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:

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

Problem 2.9 Solution

This problem is complexed, but hints have already been given in the description. Let's begin by performing integral of (2.272) over μ_{M-1}. (Note :

by integral over μ_{M-1}, we actually obtain Dirichlet distribution with $M-1$ variables.)

$$\begin{aligned} p_{M-1}(\boldsymbol{\mu}, \boldsymbol{m}, ..., \mu_{M-2}) &= \int_0^{1-\boldsymbol{\mu}-\boldsymbol{m}-...-\mu_{M-2}} C_M \prod_{k=1}^{M-1} \mu_k^{\alpha_k - 1} (1 - \sum_{j=1}^{M-1} \mu_j)^{\alpha_M - 1} d\mu_{M-1} \\ &= C_M \prod_{k=1}^{M-2} \mu_k^{\alpha_k - 1} \int_0^{1-\boldsymbol{\mu}-\boldsymbol{m}-...-\mu_{M-2}} \mu_{M-1}^{\alpha_{M-1}-1} (1 - \sum_{j=1}^{M-1} \mu_j)^{\alpha_M - 1} d\mu_{M-1} \end{aligned}$$

We change variable by :

$$t = \frac{\mu_{M-1}}{1 - \boldsymbol{\mu} - \boldsymbol{m} - ... - \mu_{M-2}}$$

The reason we do so is that $\mu_{M-1} \in [0, 1-\boldsymbol{\mu}-\boldsymbol{m}-...-\mu_{M-2}]$, by making this changing of variable, we can see that $t \in [0,1]$. Then we can further simplify the expression.

$$\begin{aligned} p_{M-1} &= C_M \prod_{k=1}^{M-2} \mu_k^{\alpha_k - 1} (1 - \sum_{j=1}^{M-2} \mu_j)^{\alpha_{M-1} + \alpha_M - 1} \int_0^1 \frac{\mu_{M-1}^{\alpha_{M-1}-1} (1 - \sum_{j=1}^{M-1} \mu_j)^{\alpha_M - 1}}{(1 - \boldsymbol{\mu} - \boldsymbol{m} - ... - \mu_{M-2})^{\alpha_{M-1} + \alpha_M - 2}} dt \\ &= C_M \prod_{k=1}^{M-2} \mu_k^{\alpha_k - 1} (1 - \sum_{j=1}^{M-2} \mu_j)^{\alpha_{M-1} + \alpha_M - 1} \int_0^1 t^{\alpha_{M-1}-1} (1-t)^{\alpha_M - 1} dt \\ &= C_M \prod_{k=1}^{M-2} \mu_k^{\alpha_k - 1} (1 - \sum_{j=1}^{M-2} \mu_j)^{\alpha_{M-1} + \alpha_M - 1} \frac{\Gamma(\alpha_{M-1} - 1)\Gamma(\alpha_M)}{\Gamma(\alpha_{M-1} + \alpha_M)} \end{aligned}$$

Comparing the expression above with a normalized Dirichlet Distribution with $M-1$ variables, and supposing that (2.272) holds for $M-1$, we can obtain that:

$$C_M \frac{\Gamma(\alpha_{M-1})\Gamma(\alpha_M)}{\Gamma(\alpha_{M-1} + \alpha_M)} = \frac{\Gamma(\alpha_1 + \alpha_2 + ... + \alpha_M)}{\Gamma(\alpha_1)\Gamma(\alpha_2)...\Gamma(\alpha_{M-1} + \alpha_M)}$$

Therefore, we obtain

$$C_M = \frac{\Gamma(\alpha_1 + \alpha_2 + ... + \alpha_M)}{\Gamma(\alpha_1)\Gamma(\alpha_2)...\Gamma(\alpha_{M-1})\Gamma(\alpha_M)}$$

as required.

Problem 2.10 Solution

Based on definition of *Expectation* and (2.38), we can write:

$$\begin{aligned}
\mathbb{E}[\mu_j] &= \int \mu_j Dir(\boldsymbol{\mu}|\boldsymbol{\alpha}) d\boldsymbol{\mu} \\
&= \int \mu_j \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_2)...\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1} d\boldsymbol{\mu} \\
&= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_2)...\Gamma(\alpha_K)} \int \mu_j \prod_{k=1}^{K} \mu_k^{\alpha_k - 1} d\boldsymbol{\mu} \\
&= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_2)...\Gamma(\alpha_K)} \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)...\Gamma(\alpha_{j-1})\Gamma(\alpha_j+1)\Gamma(\alpha_{j+1})...\Gamma(\alpha_K)}{\Gamma(\alpha_0+1)} \\
&= \frac{\Gamma(\alpha_0)\Gamma(\alpha_j+1)}{\Gamma(\alpha_j)\Gamma(\alpha_0+1)} = \frac{\alpha_j}{\alpha_0}
\end{aligned}$$

It is quite the same for variance, let's begin by calculating $\mathbb{E}[\mu_j^2]$.

$$\begin{aligned}
\mathbb{E}[\mu_j^2] &= \int \mu_j^2 Dir(\boldsymbol{\mu}|\boldsymbol{\alpha}) d\boldsymbol{\mu} \\
&= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_2)...\Gamma(\alpha_K)} \int \mu_j^2 \prod_{k=1}^{K} \mu_k^{\alpha_k - 1} d\boldsymbol{\mu} \\
&= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_2)...\Gamma(\alpha_K)} \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)...\Gamma(\alpha_{j-1})\Gamma(\alpha_j+2)\Gamma(\alpha_{j+1})...\Gamma(\alpha_K)}{\Gamma(\alpha_0+2)} \\
&= \frac{\Gamma(\alpha_0)\Gamma(\alpha_j+2)}{\Gamma(\alpha_j)\Gamma(\alpha_0+2)} = \frac{\alpha_j(\alpha_j+1)}{\alpha_0(\alpha_0+1)}
\end{aligned}$$

Hence, we obtain :

$$var[\mu_j] = \mathbb{E}[\mu_j^2] - \mathbb{E}[\mu_j]^2 = \frac{\alpha_j(\alpha_j+1)}{\alpha_0(\alpha_0+1)} - (\frac{\alpha_j}{\alpha_0})^2 = \frac{\alpha_j(\alpha_0-\alpha_j)}{\alpha_0^2(\alpha_0+1)}$$

It is the same for covariance.

$$\begin{aligned}
cov[\mu_j\mu_l] &= \int (\mu_j - \mathbb{E}[\mu_j])(\mu_l - \mathbb{E}[\mu_l]) Dir(\boldsymbol{\mu}|\boldsymbol{\alpha}) d\boldsymbol{\mu} \\
&= \int (\mu_j\mu_l - \mathbb{E}[\mu_j]\mu_l - \mathbb{E}[\mu_l]\mu_j + \mathbb{E}[\mu_j]\mathbb{E}[\mu_l]) Dir(\boldsymbol{\mu}|\boldsymbol{\alpha}) d\boldsymbol{\mu} \\
&= \frac{\Gamma(\alpha_0)\Gamma(\alpha_j+1)\Gamma(\alpha_l+1)}{\Gamma(\alpha_j)\Gamma(\alpha_l)\Gamma(\alpha_0+2)} - 2\mathbb{E}[\mu_j]\mathbb{E}[\mu_l] + \mathbb{E}[\mu_j]\mathbb{E}[\mu_l] \\
&= \frac{\alpha_j\alpha_l}{\alpha_0(\alpha_0+1)} - \mathbb{E}[\mu_j]\mathbb{E}[\mu_l] \\
&= \frac{\alpha_j\alpha_l}{\alpha_0(\alpha_0+1)} - \frac{\alpha_j\alpha_l}{\alpha_0^2} \\
&= -\frac{\alpha_j\alpha_l}{\alpha_0^2(\alpha_0+1)} \quad (j \neq l)
\end{aligned}$$

Note : when $j = l$, $cov[\mu_j \mu_l]$ will actually reduce to $var[\mu_j]$, however we cannot simply replace l with j in the expression of $cov[\mu_j \mu_l]$ to get the right result and that is because $\int \mu_j \mu_l Dir(\boldsymbol{\mu}|\boldsymbol{\alpha}) d\boldsymbol{\alpha}$ will reduce to $\int \mu_j^2 Dir(\boldsymbol{\mu}|\boldsymbol{\alpha}) d\boldsymbol{\alpha}$ in this case.

Problem 2.11 Solution

Based on definition of *Expectation* and (2.38), we first denote :

$$\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_2)...\Gamma(\alpha_K)} = K(\boldsymbol{\alpha})$$

Then we can write :

$$\begin{aligned}
\frac{\partial Dir(\boldsymbol{\mu}|\boldsymbol{\alpha})}{\partial \alpha_j} &= \partial(K(\boldsymbol{\alpha}) \prod_{i=1}^{K} \mu_i^{\alpha_i - 1}) / \partial \alpha_j \\
&= \frac{\partial K(\boldsymbol{\alpha})}{\partial \alpha_j} \prod_{i=1}^{K} \mu_i^{\alpha_i - 1} + K(\boldsymbol{\alpha}) \frac{\partial \prod_{i=1}^{K} \mu_i^{\alpha_i - 1}}{\partial \alpha_j} \\
&= \frac{\partial K(\boldsymbol{\alpha})}{\partial \alpha_j} \prod_{i=1}^{K} \mu_i^{\alpha_i - 1} + ln\mu_j \cdot Dir(\boldsymbol{\mu}|\boldsymbol{\alpha})
\end{aligned}$$

Then let us perform integral to both sides:

$$\int \frac{\partial Dir(\boldsymbol{\mu}|\boldsymbol{\alpha})}{\partial \alpha_j} d\boldsymbol{\mu} = \int \frac{\partial K(\boldsymbol{\alpha})}{\partial \alpha_j} \prod_{i=1}^{K} \mu_i^{\alpha_i - 1} d\boldsymbol{\mu} + \int ln\mu_j \cdot Dir(\boldsymbol{\mu}|\boldsymbol{\alpha}) d\boldsymbol{\mu}$$

The left side can be further simplified as :

$$\text{left side} = \frac{\partial \int Dir(\boldsymbol{\mu}|\boldsymbol{\alpha}) d\boldsymbol{\mu}}{\partial \alpha_j} = \frac{\partial 1}{\partial \alpha_j} = 0$$

The right side can be further simplified as :

$$\begin{aligned}
\text{right side} &= \frac{\partial K(\boldsymbol{\alpha})}{\partial \alpha_j} \int \prod_{i=1}^{K} \mu_i^{\alpha_i - 1} d\boldsymbol{\mu} + \mathbb{E}[ln\mu_j] \\
&= \frac{\partial K(\boldsymbol{\alpha})}{\partial \alpha_j} \frac{1}{K(\boldsymbol{\alpha})} + \mathbb{E}[ln\mu_j] \\
&= \frac{\partial lnK(\boldsymbol{\alpha})}{\partial \alpha_j} + \mathbb{E}[ln\mu_j]
\end{aligned}$$

Therefore, we obtain :

$$\begin{aligned}
\mathbb{E}[ln\mu_j] &= -\frac{\partial lnK(\boldsymbol{\alpha})}{\partial \alpha_j} \\
&= -\frac{\partial\left\{ln\Gamma(\alpha_0) - \sum_{i=1}^{K} ln\Gamma(\alpha_i)\right\}}{\partial \alpha_j} \\
&= \frac{\partial ln\Gamma(\alpha_j)}{\partial \alpha_j} - \frac{\partial ln\Gamma(\alpha_0)}{\partial \alpha_j} \\
&= \frac{\partial ln\Gamma(\alpha_j)}{\partial \alpha_j} - \frac{\partial ln\Gamma(\alpha_0)}{\partial \alpha_0}\frac{\partial \alpha_0}{\partial \alpha_j} \\
&= \frac{\partial ln\Gamma(\alpha_j)}{\partial \alpha_j} - \frac{\partial ln\Gamma(\alpha_0)}{\partial \alpha_0} \\
&= \psi(\alpha_j) - \psi(\alpha_0)
\end{aligned}$$

Therefore, the problem has been solved.

Problem 2.12 Solution

Since we have :

$$\int_a^b \frac{1}{b-a}\,dx = 1$$

It is straightforward that it is normalized. Then we calculate its mean :

$$\mathbb{E}[x] = \int_a^b x\frac{1}{b-a}\,dx = \frac{x^2}{2(b-a)}\Big|_a^b = \frac{a+b}{2}$$

Then we calculate its variance.

$$var[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \int_a^b \frac{x^2}{b-a}\,dx - (\frac{a+b}{2})^2 = \frac{x^3}{3(b-a)}\Big|_a^b - (\frac{a+b}{2})^2$$

Hence we obtain:

$$var[x] = \frac{(b-a)^2}{12}$$

Problem 2.13 Solution

This problem is an extension of Prob.1.30. We can follow the same procedure to solve it. Let's begin by calculating $ln\,\frac{p(\boldsymbol{x})}{q((\boldsymbol{x})}$:

$$ln(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}) = \frac{1}{2}ln(\frac{|\boldsymbol{L}|}{|\boldsymbol{\Sigma}|}) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})^T\boldsymbol{L}^{-1}(\boldsymbol{x}-\boldsymbol{m}) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$$

If $x \sim p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\Sigma})$, we then take advantage of the following properties.

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

$$\mathbb{E}[\boldsymbol{x}] = \int \boldsymbol{x}\, p(\boldsymbol{x})\, d\boldsymbol{x} = \mu$$

$$\mathbb{E}[(\boldsymbol{x}-\boldsymbol{a})^T \boldsymbol{A}(\boldsymbol{x}-\boldsymbol{a})] = \mathrm{tr}(\boldsymbol{A}\boldsymbol{\Sigma}) + (\boldsymbol{\mu}-\boldsymbol{a})^T \boldsymbol{A}(\boldsymbol{\mu}-\boldsymbol{a})$$

We obtain :

$$\begin{aligned}
KL &= \int \Big\{ \frac{1}{2} ln \frac{|\boldsymbol{L}|}{|\boldsymbol{\Sigma}|} - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})^T \boldsymbol{L}^{-1}(\boldsymbol{x}-\boldsymbol{m}) \Big\} p(\boldsymbol{x})\, d\boldsymbol{x} \\
&= \frac{1}{2} \ln \frac{|\boldsymbol{L}|}{|\boldsymbol{\Sigma}|} - \frac{1}{2} E[(\boldsymbol{x}-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})^T] + \frac{1}{2} E[(\boldsymbol{x}-\boldsymbol{m})^T \boldsymbol{L}^{-1}(\boldsymbol{x}-\boldsymbol{m})] \\
&= \frac{1}{2} \ln \frac{|\boldsymbol{L}|}{|\boldsymbol{\Sigma}|} - \frac{1}{2} \mathrm{tr}\,\{\boldsymbol{I}_D\} + \frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{m})^T \boldsymbol{L}^{-1}(\boldsymbol{\mu}-\boldsymbol{m}) + \frac{1}{2} \mathrm{tr}\{\boldsymbol{L}^{-1}\boldsymbol{\Sigma}\} \\
&= \frac{1}{2}[ln \frac{|\boldsymbol{L}|}{|\boldsymbol{\Sigma}|} - D + \mathrm{tr}\{\boldsymbol{L}^{-1}\boldsymbol{\Sigma}\} + (\boldsymbol{m}-\boldsymbol{\mu})^T \boldsymbol{L}^{-1}(\boldsymbol{m}-\boldsymbol{\mu})]
\end{aligned}$$

Problem 2.14 Solution

The hint given in the problem is straightforward, however it is a little bit difficult to calculate, and here we will use a more simple method to solve this problem, taking advantage of the property of *Kullback—Leibler Distance*. Let $g(\boldsymbol{x})$ be a Gaussian PDF with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$, and $f(\boldsymbol{x})$ an arbitrary PDF with the same mean and variance.

$$0 \le KL(f||g) = -\int f(\boldsymbol{x}) ln \left\{ \frac{g(\boldsymbol{x})}{f(\boldsymbol{x})} \right\} d\boldsymbol{x} = -H(f) - \int f(\boldsymbol{x}) ln g(\boldsymbol{x})\, d\boldsymbol{x} \qquad (*)$$

Let's calculate the second term of the equation above.

$$\begin{aligned}
\int f(\boldsymbol{x}) ln g(\boldsymbol{x})\, d\boldsymbol{x} &= \int f(\boldsymbol{x}) ln \left\{ \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} exp\big[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \big] \right\} d\boldsymbol{x} \\
&= \int f(\boldsymbol{x}) ln \left\{ \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \right\} d\boldsymbol{x} + \int f(\boldsymbol{x}) \big[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \big] d\boldsymbol{x} \\
&= ln \left\{ \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \right\} - \frac{1}{2} \mathbb{E}\big[(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \big] \\
&= ln \left\{ \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \right\} - \frac{1}{2} \mathrm{tr}\{I_D\} \\
&= -\left\{ \frac{1}{2} ln|\boldsymbol{\Sigma}| + \frac{D}{2}(1 + ln(2\pi)) \right\} \\
&= -H(g)
\end{aligned}$$

We take advantage of two properties of PDF $f(\boldsymbol{x})$, with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$, as listed below. What's more, we also use the result of Prob.2.15, which we will proof later.

$$\int f(\boldsymbol{x})\, d\boldsymbol{x} = 1$$

$$\mathbb{E}[(\boldsymbol{x}-\boldsymbol{a})^T \boldsymbol{A}(\boldsymbol{x}-\boldsymbol{a})] = \mathrm{tr}(\boldsymbol{A}\boldsymbol{\Sigma}) + (\boldsymbol{\mu}-\boldsymbol{a})^T \boldsymbol{A}(\boldsymbol{\mu}-\boldsymbol{a})$$

Now we can further simplify $(*)$ to obtain:

$$H(g) \geq H(f)$$

In other words, we have proved that an arbitrary PDF $f(\boldsymbol{x})$ with the same mean and variance as a Gaussian PDF $g(\boldsymbol{x})$, its entropy cannot be greater than that of Gaussian PDF.

Problem 2.15 Solution

We have already used the result of this problem to solve Prob.2.14, and now we will prove it. Suppose $x \sim p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\Sigma})$:

$$\begin{aligned}
H[\boldsymbol{x}] &= -\int p(\boldsymbol{x}) ln\, p(\boldsymbol{x})\, d\boldsymbol{x} \\
&= -\int p(\boldsymbol{x}) ln \left\{ \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} exp\big[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu}) \big] \right\} d\boldsymbol{x} \\
&= -\int p(\boldsymbol{x}) ln \left\{ \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \right\} d\boldsymbol{x} - \int f(\boldsymbol{x}) \big[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu}) \big] d\boldsymbol{x} \\
&= -ln \left\{ \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \right\} + \frac{1}{2} \mathbb{E}\big[(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu}) \big] \\
&= -ln \left\{ \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \right\} + \frac{1}{2} \mathrm{tr}\{I_D\} \\
&= \frac{1}{2} ln|\boldsymbol{\Sigma}| + \frac{D}{2}(1 + ln(2\pi))
\end{aligned}$$

Where we have taken advantage of :

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

$$\mathbb{E}[(\boldsymbol{x}-\boldsymbol{a})^T \boldsymbol{A} (\boldsymbol{x}-\boldsymbol{a})] = \mathrm{tr}(\boldsymbol{A}\boldsymbol{\Sigma}) + (\boldsymbol{\mu}-\boldsymbol{a})^T \boldsymbol{A} (\boldsymbol{\mu}-\boldsymbol{a})$$

Note : Actually in Prob.2.14, we have already solved this problem, you can intuitively view it by replacing the integrand $f(\boldsymbol{x}) ln\, g(\boldsymbol{x})$ with $g(\boldsymbol{x}) ln\, g(\boldsymbol{x})$, and the same procedure in Prob.2.14 still holds to calculate $\int g(\boldsymbol{x}) ln\, g(\boldsymbol{x})\, d\boldsymbol{x}$.

Problem 2.16 Solution

Let us consider a more general conclusion about the *Probability Density Function* (PDF) of the summation of two independent random variables. We denote two random variables X and Y. Their summation $Z = X + Y$, is still a random variable. We also denote $f(\cdot)$ as PDF, and $F(\cdot)$ as *Cumulative Distribution Function* (CDF). We can obtain :

$$F_Z(z) = P(Z < z) = \iint_{x+y \leq z} f_{X,Y}(x,y)\, dx dy$$

Where z represents an arbitrary real number. We rewrite the *double integral* into *iterated integral* :

$$F_Z(z) = \int_{-\infty}^{+\infty} \left[\int_{-\infty}^{z-y} f_{X,Y}(x,y)dx \right] dy$$

We fix z and y, and then make a change of variable $x = u - y$ to the integral.

$$F_Z(z) = \int_{-\infty}^{+\infty} \left[\int_{-\infty}^{z-y} f_{X,Y}(x,y)dx \right] dy = \int_{-\infty}^{+\infty} \left[\int_{-\infty}^{z} f_{X,Y}(u-x,y)du \right] dy$$

Note: $f_{X,Y}(\cdot)$ is the joint PDF of X and Y, and then we rearrange the order, we will obtain :

$$F_Z(z) = \int_{-\infty}^{z} \left[\int_{-\infty}^{+\infty} f_{X,Y}(u-y,y)dy \right] du$$

Compare the equation above with th definition of CDF :

$$F_Z(z) = \int_{-\infty}^{z} f_Z(u)du$$

We can obtain :

$$f_Z(u) = \int_{-\infty}^{+\infty} f_{X,Y}(u-y,y)dy$$

And if X and Y are independent, which means $f_{X,Y}(x,y) = f_X(x)f_Y(y)$, we can simplify $f_Z(z)$:

$$f_Z(u) = \int_{-\infty}^{+\infty} f_X(u-y)f_Y(y)dy \quad \text{i.e.} \quad f_Z = f_X * f_Y$$

Until now we have proved that the PDF of the summation of two independent random variable is the convolution of the PDF of them. Hence it is straightforward to see that in this problem, where random variable x is the summation of random variable x_1 and x_2, the PDF of x should be the convolution of the PDF of x_1 and x_2. To find the entropy of x, we will use a simple method, taking advantage of (2.113)-(2.117). With the knowledge :

$$p(x_2) = \mathcal{N}(\mu_2, \tau_2^{-1})$$

$$p(x|x_2) = \mathcal{N}(\mu_1 + x_2, \tau_1^{-1})$$

We make analogies : x_2 in this problem to $\boldsymbol{x}$ in (2.113), x in this problem to $\boldsymbol{y}$ in (2.114). Hence by using (2.115), we can obtain $p(x)$ is still a normal distribution, and since the entropy of a Gaussian is fully decided by its variance, there is no need to calculate the mean. Still by using (2.115), the variance of x is $\tau_1^{-1} + \tau_2^{-1}$, which finally gives its entropy :

$$H[x] = \frac{1}{2}\left[1 + ln2\pi(\tau_1^{-1} + \tau_2^{-1})\right]$$

Problem 2.17 Solution

This is an extension of Prob.1.14. The same procedure can be used here. We suppose an arbitrary precision matrix Λ can be written as $\mathbf{\Lambda}^S + \mathbf{\Lambda}^A$, where they satisfy :

$$\Lambda_{ij}^S = \frac{\Lambda_{ij} + \Lambda_{ji}}{2} \quad , \quad \Lambda_{ij}^A = \frac{\Lambda_{ij} - \Lambda_{ji}}{2}$$

Hence it is straightforward that $\Lambda_{ij}^S = \Lambda_{ji}^S$, and $\Lambda_{ij}^A = -\Lambda_{ji}^A$. If we expand the quadratic form of exponent, we will obtain :

$$(\boldsymbol{x} - \boldsymbol{\mu})^T \mathbf{\Lambda} (\boldsymbol{x} - \boldsymbol{\mu}) = \sum_{i=1}^{D} \sum_{j=1}^{D} (x_i - \mu_i) \Lambda_{ij} (x_j - \mu_j) \qquad (*)$$

It is straightforward then :

$$\begin{aligned} (*) &= \sum_{i=1}^{D} \sum_{j=1}^{D} (x_i - \mu_i) \Lambda_{ij}^S (x_j - \mu_j) + \sum_{i=1}^{D} \sum_{j=1}^{D} (x_i - \mu_i) \Lambda_{ij}^A (x_j - \mu_j) \\ &= \sum_{i=1}^{D} \sum_{j=1}^{D} (x_i - \mu_i) \Lambda_{ij}^S (x_j - \mu_j) \end{aligned}$$

Therefore, we can assume precision matrix is symmetric, and so is covariance matrix.

Problem 2.18 Solution

We will just follow the hint given in the problem. Firstly, we take complex conjugate on both sides of (2.45) :

$$\overline{\mathbf{\Sigma} \boldsymbol{u_i}} = \overline{\lambda_i \boldsymbol{u_i}} \quad => \quad \mathbf{\Sigma} \overline{\boldsymbol{u_i}} = \overline{\lambda_i} \overline{\boldsymbol{u_i}}$$

Where we have taken advantage of the fact that $\mathbf{\Sigma}$ is a real matrix, i.e., $\overline{\mathbf{\Sigma}} = \mathbf{\Sigma}$. Then using that $\mathbf{\Sigma}$ is a symmetric, i.e., $\mathbf{\Sigma}^T = \mathbf{\Sigma}$:

$$\overline{\boldsymbol{u_i}}^T \mathbf{\Sigma} \boldsymbol{u_i} = \overline{\boldsymbol{u_i}}^T (\mathbf{\Sigma} \boldsymbol{u_i}) = \overline{\boldsymbol{u_i}}^T (\lambda_i \boldsymbol{u_i}) = \lambda_i \overline{\boldsymbol{u_i}}^T \boldsymbol{u_i}$$

$$\overline{\boldsymbol{u_i}}^T \mathbf{\Sigma} \boldsymbol{u_i} = (\mathbf{\Sigma} \overline{\boldsymbol{u_i}})^T \boldsymbol{u_i} = (\overline{\lambda_i} \overline{\boldsymbol{u_i}})^T \boldsymbol{u_i} = \overline{\lambda_i}^T \overline{\boldsymbol{u_i}}^T \boldsymbol{u_i}$$

Since $\boldsymbol{u_i} \neq 0$, we have $\overline{\boldsymbol{u_i}}^T \boldsymbol{u_i} \neq 0$. Thus $\lambda_i^T = \overline{\lambda_i}^T$, which means λ_i is real. Next we will proof that two eigenvectors corresponding to different eigenvalues are orthogonal.

$$\lambda_i < \boldsymbol{u_i}, \boldsymbol{u_j} > = < \lambda_i \boldsymbol{u_i}, \boldsymbol{u_j} > = < \mathbf{\Sigma} \boldsymbol{u_i}, \boldsymbol{u_j} > = < \boldsymbol{u_i}, \mathbf{\Sigma}^T \boldsymbol{u_j} > = \lambda_j < \boldsymbol{u_i}, \boldsymbol{u_j} >$$

Where we have taken advantage of $\mathbf{\Sigma}^T = \mathbf{\Sigma}$ and for arbitrary real matrix $\boldsymbol{A}$ and vector $\boldsymbol{x}$, $\boldsymbol{y}$, we have :

$$< \boldsymbol{A x}, \boldsymbol{y} > = < \boldsymbol{x}, \boldsymbol{A^T y} >$$

Provided $\lambda_i \neq \lambda_j$, we have $< \boldsymbol{u_i}, \boldsymbol{u_j} >= 0$, i.e., $\boldsymbol{u_i}$ and $\boldsymbol{u_j}$ are orthogonal. And then if we perform normalization on every eigenvector to force its *Euclidean norm* to equal to 1, (2.46) is straightforward. By performing normalization, I mean multiplying the eigenvector by a real number a to let its *Euclidean norm* (length) to equal to 1, meanwhile we should also divide its corresponding eigenvalue by a.

Problem 2.19 Solution

For every $N \times N$ real symmetric matrix, the eigenvalues are real and the eigenvectors can be chosen such that they are orthogonal to each other. Thus a real symmetric matrix $\boldsymbol{\Sigma}$ can be decomposed as $\boldsymbol{\Sigma} = \boldsymbol{U\Lambda U^T}$,where $\boldsymbol{U}$ is an orthogonal matrix, and $\boldsymbol{\Lambda}$ is a diagonal matrix whose entries are the eigenvalues of A. Hence for an arbitrary vector $\boldsymbol{x}$, we have:

$$\boldsymbol{\Sigma x} = \boldsymbol{U\Lambda U^T x} = \boldsymbol{U\Lambda} \begin{bmatrix} \boldsymbol{u_1^T x} \\ \vdots \\ \boldsymbol{u_D^T x} \end{bmatrix} = \boldsymbol{U} \begin{bmatrix} \lambda_1 \boldsymbol{u_1^T x} \\ \vdots \\ \lambda_D \boldsymbol{u_D^T x} \end{bmatrix} = (\sum_{k=1}^{D} \lambda_k \boldsymbol{u_k u_k^T}) \boldsymbol{x}$$

And since $\boldsymbol{\Sigma}^{-1} = \boldsymbol{U\Lambda^{-1}U^T}$, the same procedure can be used to prove (2.49).

Problem 2.20 Solution

Since $\boldsymbol{u_1}, \boldsymbol{u_2}, ..., \boldsymbol{u_D}$ can constitute a basis for $\mathbb{R}^D$, we can make projection for $\boldsymbol{a}$:

$$\boldsymbol{a} = a_1 \boldsymbol{u_1} + a_2 \boldsymbol{u_2} + ... + a_D \boldsymbol{u_D}$$

We substitute the expression above into $\boldsymbol{a^T \Sigma a}$, taking advantage of the property: $\boldsymbol{u_i u_j} = 1$ only if $i = j$, otherwise 0, we will obtain :

$$\begin{aligned} \boldsymbol{a^T \Sigma a} &= (a_1 \boldsymbol{u_1} + a_2 \boldsymbol{u_2} + ... + a_D \boldsymbol{u_D})^T \boldsymbol{\Sigma} (a_1 \boldsymbol{u_1} + a_2 \boldsymbol{u_2} + ... + a_D \boldsymbol{u_D}) \\ &= (a_1 \boldsymbol{u_1}^T + a_2 \boldsymbol{u_2}^T + ... + a_D \boldsymbol{u_D}^T) \boldsymbol{\Sigma} (a_1 \boldsymbol{u_1} + a_2 \boldsymbol{u_2} + ... + a_D \boldsymbol{u_D}) \\ &= (a_1 \boldsymbol{u_1}^T + a_2 \boldsymbol{u_2}^T + ... + a_D \boldsymbol{u_D}^T)(a_1 \lambda_1 \boldsymbol{u_1} + a_2 \lambda_2 \boldsymbol{u_2} + ... + a_D \lambda_D \boldsymbol{u_D}) \\ &= \lambda_1 {a_1}^2 + \lambda_2 {a_2}^2 + ... + \lambda_D {a_D}^2 \end{aligned}$$

Since $\boldsymbol{a}$ is real,the expression above will be strictly positive for any non-zero $\boldsymbol{a}$, if all eigenvalues are strictly positive. It is also clear that if an eigenvalue, λ_i, is zero or negative, there will exist a vector $\boldsymbol{a}$ (e.g. $\boldsymbol{a} = \boldsymbol{u_i}$), for which this expression will be no greater than 0. Thus, that a real symmetric matrix has eigenvectors which are all strictly positive is a sufficient and necessary condition for the matrix to be positive definite.

Problem 2.21 Solution

It is straightforward. For a symmetric matrix $\boldsymbol{\Lambda}$ of size $D \times D$, when the lower triangular part is decided, the whole matrix will be decided due to

symmetry. Hence the number of independent parameters is $D + (D-1) + ... + 1$, which equals to $D(D+1)/2$.

Problem 2.22 Solution

Suppose $\boldsymbol{A}$ is a symmetric matrix, and we need to prove that $\boldsymbol{A}^{-1}$ is also symmetric, i.e., $\boldsymbol{A}^{-1} = (\boldsymbol{A}^{-1})^T$. Since identity matrix $\boldsymbol{I}$ is also symmetric, we have :

$$\boldsymbol{A}\boldsymbol{A}^{-1} = (\boldsymbol{A}\boldsymbol{A}^{-1})^T$$

And since $\boldsymbol{A}\boldsymbol{B}^T = \boldsymbol{B}^T\boldsymbol{A}^T$ holds for arbitrary matrix $\boldsymbol{A}$ and $\boldsymbol{B}$, we will obtain :

$$\boldsymbol{A}\boldsymbol{A}^{-1} = (\boldsymbol{A}^{-1})^T\boldsymbol{A}^T$$

Since $\boldsymbol{A} = \boldsymbol{A}^T$, we substitute the right side:

$$\boldsymbol{A}\boldsymbol{A}^{-1} = (\boldsymbol{A}^{-1})^T\boldsymbol{A}$$

And note that $\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$, we rearrange the order of the left side :

$$\boldsymbol{A}^{-1}\boldsymbol{A} = (\boldsymbol{A}^{-1})^T\boldsymbol{A}$$

Finally, by multiplying $\boldsymbol{A}^{-1}$ to both sides, we can obtain:

$$\boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{A}^{-1} = (\boldsymbol{A}^{-1})^T\boldsymbol{A}\boldsymbol{A}^{-1}$$

Using $\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{I}$, we will get what we are asked :

$$\boldsymbol{A}^{-1} = (\boldsymbol{A}^{-1})^T$$

Problem 2.23 Solution

Let's reformulate the problem. What the problem wants us to prove is that if $(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = r^2$, where r^2 is a constant, we will have the volume of the hyperellipsoid decided by the equation above will equal to $V_D|\boldsymbol{\Sigma}|^{1/2}r^D$. Note that the center of this hyperellipsoid locates at $\boldsymbol{\mu}$, and a translation operation won't change its volume, thus we only need to prove that the volume of a hyperellipsoid decided by $\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} = r^2$, whose center locates at $\mathbf{0}$ equals to $V_D|\boldsymbol{\Sigma}|^{1/2}r^D$.

This problem can be viewed as two parts. Firstly, let's discuss about V_D, the volume of a unit sphere in dimension D. The expression of V_D has already be given in the solution procedure of Prob.1.18, i.e., (1.144) :

$$V_D = \frac{S_D}{D} = \frac{2\pi^{D/2}}{\Gamma(\frac{D}{2}+1)}$$

And also in the procedure, we show that a D dimensional sphere with radius r, i.e., $\boldsymbol{x}^T\boldsymbol{x} = r^2$, has volume $V(r) = V_D r^D$. We move a step forward: we

perform a linear transform using matrix $\boldsymbol{\Sigma}^{1/2}$, i.e., $\boldsymbol{y}^T\boldsymbol{y} = r^2$, where $\boldsymbol{y} = \boldsymbol{\Sigma^{1/2}x}$. After the linear transformation, we actually get a hyperellipsoid whose center locates at $\mathbf{0}$, and its volume is given by multiplying $V(r)$ with the determinant of the transformation matrix, which gives $|\boldsymbol{\Sigma}|^{1/2}V_D r^D$, just as required.

Problem 2.24 Solution

We just following the hint, and firstly let's calculate :

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix} \times \begin{bmatrix} \boldsymbol{M} & -\boldsymbol{MBD^{-1}} \\ -\boldsymbol{D^{-1}CM} & \boldsymbol{D^{-1}} + \boldsymbol{D^{-1}CMBD^{-1}} \end{bmatrix}$$

The result can also be partitioned into four blocks. The block located at left top equals to :

$$\boldsymbol{AM} - \boldsymbol{BD^{-1}CM} = (\boldsymbol{A} - \boldsymbol{BD^{-1}C})(\boldsymbol{A} - \boldsymbol{BD^{-1}C})^{-1} = \boldsymbol{I}$$

Where we have taken advantage of (2.77). And the right top equals to :

$$-\boldsymbol{AMBD^{-1}} + \boldsymbol{BD^{-1}} + \boldsymbol{BD^{-1}CMBD^{-1}} = (\boldsymbol{I} - \boldsymbol{AM} + \boldsymbol{BD^{-1}CM})\boldsymbol{BD^{-1}} = \mathbf{0}$$

Where we have used the result of the left top block. And the left bottom equals to :

$$\boldsymbol{CM} - \boldsymbol{DD^{-1}CM} = \mathbf{0}$$

And the right bottom equals to :

$$-\boldsymbol{CMBD^{-1}} + \boldsymbol{DD^{-1}} + \boldsymbol{DD^{-1}CMDD^{-1}} = \boldsymbol{I}$$

we have proved what we are asked. Note: if you want to be more precise, you should also multiply the block matrix on the right side of (2.76) and then prove that it will equal to a identity matrix. However, the procedure above can be also used there, so we omit the proof and what's more, if two arbitrary square matrix $\boldsymbol{X}$ and $\boldsymbol{Y}$ satisfied $\boldsymbol{XY} = \boldsymbol{I}$, it can be shown that $\boldsymbol{YX} = \boldsymbol{I}$ also holds.

Problem 2.25 Solution

We will take advantage of the result of (2.94)-(2.98). Let's first begin by grouping $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$ together, and then we rewrite what has been given as :

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_{a,b} \\ \boldsymbol{x}_c \end{pmatrix} \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_{a,b} \\ \boldsymbol{\mu}_c \end{pmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{(a,b)(a,b)} & \boldsymbol{\Sigma}_{(a,b)c} \\ \boldsymbol{\Sigma}_{(a,b)c} & \boldsymbol{\Sigma}_{cc} \end{bmatrix}$$

Then we take advantage of (2.98), we can obtain :

$$p(\boldsymbol{x}_{a,b}) = \mathcal{N}(\boldsymbol{x}_{a,b}|\boldsymbol{\mu}_{a,b}, \boldsymbol{\Sigma}_{(a,b)(a,b)})$$

Where we have defined:

$$\boldsymbol{\mu}_{a,b} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \quad \boldsymbol{\Sigma}_{(a,b)(a,b)} = \begin{bmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{bmatrix}$$

Since now we have obtained the joint contribution of $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$, we will take advantage of (2.96) (2.97) to obtain conditional distribution, which gives:

$$p(\boldsymbol{x}_a|\boldsymbol{x}_b) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Lambda}_{aa}^{-1})$$

Where we have defined

$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b - \boldsymbol{\mu}_b)$$

And the expression of $\boldsymbol{\Lambda}_{aa}^{-1}$ and $\boldsymbol{\Lambda}_{ab}$ can be given by using (2.76) and (2.77) once we notice that the following relation exits:

$$\begin{bmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{bmatrix}^{-1}$$

Problem 2.26 Solution

This problem is quite straightforward, if we just follow the hint.

$$\begin{aligned}
&(\boldsymbol{A}+\boldsymbol{BCD})\left(\boldsymbol{A}^{-1}-\boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{C}^{-1}+\boldsymbol{DA}^{-1}\boldsymbol{B})^{-1}\boldsymbol{DA}^{-1}\right) \\
&= \boldsymbol{AA}^{-1}-\boldsymbol{AA}^{-1}\boldsymbol{B}(\boldsymbol{C}^{-1}+\boldsymbol{DA}^{-1}\boldsymbol{B})^{-1}\boldsymbol{DA}^{-1}+\boldsymbol{BCDA}^{-1}-\boldsymbol{BCDA}^{-1}\boldsymbol{B}(\boldsymbol{C}^{-1}+\boldsymbol{DA}^{-1}\boldsymbol{B})^{-1}\boldsymbol{DA}^{-1} \\
&= \boldsymbol{I}-\boldsymbol{B}(\boldsymbol{C}^{-1}+\boldsymbol{DA}^{-1}\boldsymbol{B})^{-1}\boldsymbol{DA}^{-1}+\boldsymbol{BCDA}^{-1}+\boldsymbol{B}(\boldsymbol{C}^{-1}+\boldsymbol{DA}^{-1}\boldsymbol{B})^{-1}\boldsymbol{DA}^{-1}-\boldsymbol{BCDA}^{-1} \\
&= \boldsymbol{I}
\end{aligned}$$

Where we have taken advantage of

$$\begin{aligned}
&-\boldsymbol{BCDA}^{-1}\boldsymbol{B}(\boldsymbol{C}^{-1}+\boldsymbol{DA}^{-1}\boldsymbol{B})^{-1}\boldsymbol{DA}^{-1} \\
&= -\boldsymbol{BC}(-\boldsymbol{C}^{-1}+\boldsymbol{C}^{-1}+\boldsymbol{DA}^{-1}\boldsymbol{B})(\boldsymbol{C}^{-1}+\boldsymbol{DA}^{-1}\boldsymbol{B})^{-1}\boldsymbol{DA}^{-1} \\
&= (-\boldsymbol{BC})(-\boldsymbol{C}^{-1})(\boldsymbol{C}^{-1}+\boldsymbol{DA}^{-1}\boldsymbol{B})^{-1}\boldsymbol{DA}^{-1}+(-\boldsymbol{BC})(\boldsymbol{C}^{-1}+\boldsymbol{DA}^{-1}\boldsymbol{B})(\boldsymbol{C}^{-1}+\boldsymbol{DA}^{-1}\boldsymbol{B})^{-1}\boldsymbol{DA}^{-1} \\
&= \boldsymbol{B}(\boldsymbol{C}^{-1}+\boldsymbol{DA}^{-1}\boldsymbol{B})^{-1}\boldsymbol{DA}^{-1}-\boldsymbol{BCDA}^{-1}
\end{aligned}$$

Here we will also directly calculate the inverse matrix instead to give another solution. Let's first begin by introducing two useful formulas.

$$\begin{aligned}
(\boldsymbol{I}+\boldsymbol{P})^{-1} &= (\boldsymbol{I}+\boldsymbol{P})^{-1}(\boldsymbol{I}+\boldsymbol{P}-\boldsymbol{P}) \\
&= \boldsymbol{I}-(\boldsymbol{I}+\boldsymbol{P})^{-1}\boldsymbol{P}
\end{aligned}$$

And since

$$\boldsymbol{P}+\boldsymbol{PQP} = \boldsymbol{P}(\boldsymbol{I}+\boldsymbol{QP}) = (\boldsymbol{I}+\boldsymbol{PQ})\boldsymbol{P}$$

The second formula is :

$$(\boldsymbol{I}+\boldsymbol{PQ})^{-1}\boldsymbol{P} = \boldsymbol{P}(\boldsymbol{I}+\boldsymbol{QP})^{-1}$$

And now let's directly calculate $(\boldsymbol{A}+\boldsymbol{BCD})^{-1}$:

$$\begin{aligned}(\boldsymbol{A}+\boldsymbol{BCD})^{-1} &= [\boldsymbol{A}(\boldsymbol{I}+\boldsymbol{A^{-1}BCD})]^{-1} \\ &= (\boldsymbol{I}+\boldsymbol{A^{-1}BCD})^{-1}\boldsymbol{A}^{-1} \\ &= \left[\boldsymbol{I}-(\boldsymbol{I}+\boldsymbol{A}^{-1}\boldsymbol{BCD})^{-1}\boldsymbol{A}^{-1}\boldsymbol{BCD}\right]\boldsymbol{A}^{-1} \\ &= \boldsymbol{A}^{-1}-(\boldsymbol{I}+\boldsymbol{A}^{-1}\boldsymbol{BCD})^{-1}\boldsymbol{A}^{-1}\boldsymbol{BCDA}^{-1}\end{aligned}$$

Where we have assumed that $\boldsymbol{A}$ is invertible and also used the first formula we introduced. Then we also assume that $\boldsymbol{C}$ is invertible and recursively use the second formula :

$$\begin{aligned}(\boldsymbol{A}+\boldsymbol{BCD})^{-1} &= \boldsymbol{A}^{-1}-(\boldsymbol{I}+\boldsymbol{A}^{-1}\boldsymbol{BCD})^{-1}\boldsymbol{A}^{-1}\boldsymbol{BCDA}^{-1} \\ &= \boldsymbol{A}^{-1}-\boldsymbol{A}^{-1}(\boldsymbol{I}+\boldsymbol{BCDA}^{-1})^{-1}\boldsymbol{BCDA}^{-1} \\ &= \boldsymbol{A}^{-1}-\boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{I}+\boldsymbol{CDA}^{-1}\boldsymbol{B})^{-1}\boldsymbol{CDA}^{-1} \\ &= \boldsymbol{A}^{-1}-\boldsymbol{A}^{-1}\boldsymbol{B}\left[\boldsymbol{C}(\boldsymbol{C}^{-1}+\boldsymbol{DA}^{-1}\boldsymbol{B})\right]^{-1}\boldsymbol{CDA}^{-1} \\ &= \boldsymbol{A}^{-1}-\boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{C}^{-1}+\boldsymbol{DA}^{-1}\boldsymbol{B})^{-1}\boldsymbol{C}^{-1}\boldsymbol{CDA}^{-1} \\ &= \boldsymbol{A}^{-1}-\boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{C}^{-1}+\boldsymbol{DA}^{-1}\boldsymbol{B})^{-1}\boldsymbol{DA}^{-1}\end{aligned}$$

Just as required.

Problem 2.27 Solution

The same procedure used in Prob.1.10 can be used here similarly.

$$\begin{aligned}\mathbb{E}[\boldsymbol{x}+\boldsymbol{z}] &= \int\int(\boldsymbol{x}+\boldsymbol{z})p(\boldsymbol{x},\boldsymbol{z})d\boldsymbol{x}d\boldsymbol{z} \\ &= \int\int(\boldsymbol{x}+\boldsymbol{z})p(\boldsymbol{x})p(\boldsymbol{z})d\boldsymbol{x}d\boldsymbol{z} \\ &= \int\int\boldsymbol{x}p(\boldsymbol{x})p(\boldsymbol{z})d\boldsymbol{x}d\boldsymbol{z}+\int\int\boldsymbol{z}p(\boldsymbol{x})p(\boldsymbol{z})d\boldsymbol{x}d\boldsymbol{z} \\ &= \int(\int p(\boldsymbol{z})d\boldsymbol{z})\boldsymbol{x}p(\boldsymbol{x})d\boldsymbol{x}+\int(\int p(\boldsymbol{x})d\boldsymbol{x})\boldsymbol{z}p(\boldsymbol{z})d\boldsymbol{z} \\ &= \int\boldsymbol{x}p(\boldsymbol{x})d\boldsymbol{x}+\int\boldsymbol{z}p(\boldsymbol{z})d\boldsymbol{z} \\ &= \mathbb{E}[\boldsymbol{x}]+\mathbb{E}[\boldsymbol{z}]\end{aligned}$$

And for covariance matrix, we will use matrix integral :

$$cov[\boldsymbol{x}+\boldsymbol{z}] = \int\int(\boldsymbol{x}+\boldsymbol{z}-\mathbb{E}[\boldsymbol{x}+\boldsymbol{z}])(\boldsymbol{x}+\boldsymbol{z}-\mathbb{E}[\boldsymbol{x}+\boldsymbol{z}])^{T}p(\boldsymbol{x},\boldsymbol{z})d\boldsymbol{x}d\boldsymbol{z}$$

Also the same procedure can be used here. We omit the proof for simplicity.

Problem 2.28 Solution

It is quite straightforward when we compare the problem with (2.94)-(2.98). We treat $\boldsymbol{x}$ in (2.94) as $\boldsymbol{z}$ in this problem, $\boldsymbol{x}_a$ in (2.94) as $\boldsymbol{x}$ in this problem, $\boldsymbol{x}_b$ in (2.94) as $\boldsymbol{y}$ in this problem. In other words, we rewrite the problem in the form of (2.94)-(2.98), which gives :

$$\boldsymbol{z} = \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix} \quad \mathbb{E}(\boldsymbol{z}) = \begin{pmatrix} \boldsymbol{\mu} \\ \boldsymbol{A\mu} + \boldsymbol{b} \end{pmatrix} \quad cov(\boldsymbol{z}) = \begin{bmatrix} \boldsymbol{\Lambda}^{-1} & \boldsymbol{\Lambda}^{-1}\boldsymbol{A}^T \\ \boldsymbol{A\Lambda}^{-1} & \boldsymbol{L}^{-1} + \boldsymbol{A\Lambda}^{-1}\boldsymbol{A}^T \end{bmatrix}$$

By using (2.98), we can obtain:

$$p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$$

And by using (2.96) and (2.97), we can obtain :

$$p(\boldsymbol{y}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{\mu}_{\boldsymbol{y}|\boldsymbol{x}}, \boldsymbol{\Lambda}_{\boldsymbol{yy}}^{-1})$$

Where $\boldsymbol{\Lambda}_{\boldsymbol{yy}}$ can be obtained by the right bottom part of (2.104),which gives $\boldsymbol{\Lambda}_{\boldsymbol{yy}} = \boldsymbol{L}^{-1}$, and you can also calculate it using (2.105) combined with (2.78) and (2.79). Finally the conditional mean is given by (2.97) :

$$\boldsymbol{\mu}_{\boldsymbol{y}|\boldsymbol{x}} = \boldsymbol{A\mu} + \boldsymbol{L} - \boldsymbol{L}^{-1}(-\boldsymbol{LA})(\boldsymbol{x} - \boldsymbol{\mu}) = \boldsymbol{Ax} + \boldsymbol{L}$$

Problem 2.29 Solution

It is straightforward. Firstly, we calculate the left top block :

$$\text{left top} = \left[(\boldsymbol{\Lambda} + \boldsymbol{A}^T\boldsymbol{LA}) - (-\boldsymbol{A}^T\boldsymbol{L})(\boldsymbol{L}^{-1})(-\boldsymbol{LA})\right]^{-1} = \boldsymbol{\Lambda}^{-1}$$

And then the right top block :

$$\text{right top} = -\boldsymbol{\Lambda}^{-1}(-\boldsymbol{A}^T\boldsymbol{L})\boldsymbol{L}^{-1} = \boldsymbol{\Lambda}^{-1}\boldsymbol{A}^T$$

And then the left bottom block :

$$\text{left bottom} = -\boldsymbol{L}^{-1}(-\boldsymbol{LA})\boldsymbol{\Lambda}^{-1} = \boldsymbol{A\Lambda}^{-1}$$

Finally the right bottom block :

$$\text{right bottom} = \boldsymbol{L}^{-1} + \boldsymbol{L}^{-1}(-\boldsymbol{LA})\boldsymbol{\Lambda}^{-1}(-\boldsymbol{A}^T\boldsymbol{L})\boldsymbol{L}^{-1} = \boldsymbol{L}^{-1} + \boldsymbol{A\Lambda}^{-1}\boldsymbol{A}^T$$

Problem 2.30 Solution

It is straightforward by multiplying (2.105) and (2.107), which gives :

$$\begin{pmatrix} \boldsymbol{\Lambda}^{-1} & \boldsymbol{\Lambda}^{-1}\boldsymbol{A}^T \\ \boldsymbol{A\Lambda}^{-1} & \boldsymbol{L}^{-1} + \boldsymbol{A\Lambda}^{-1}\boldsymbol{A}^T \end{pmatrix} \begin{pmatrix} \boldsymbol{\Lambda\mu} - \boldsymbol{A}^T\boldsymbol{Lb} \\ \boldsymbol{Lb} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu} \\ \boldsymbol{A\mu} + \boldsymbol{b} \end{pmatrix}$$

Just as required in the problem.

Problem 2.31 Solution

According to the problem, we can write two expressions :

$$p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu_x}, \boldsymbol{\Sigma_x}), \quad p(\boldsymbol{y}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{\mu_z} + \boldsymbol{x}, \boldsymbol{\Sigma_z})$$

By comparing the expression above and (2.113)-(2.117), we can write the expression of $p(\boldsymbol{y})$:

$$p(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{\mu_x} + \boldsymbol{\mu_z}, \boldsymbol{\Sigma_x} + \boldsymbol{\Sigma_z})$$

Problem 2.32 Solution

Let's make this problem more clear. The deduction in the main text, i.e., (2.101-2.110), firstly denote a new random variable $\boldsymbol{z}$ corresponding to the joint distribution, and then by completing square according to $\boldsymbol{z}$,i.e.,(2.103), obtain the precision matrix $\boldsymbol{R}$ by comparing (2.103) with the PDF of a multivariate Gaussian Distribution, and then it takes the inverse of precision matrix to obtain covariance matrix, and finally it obtains the linear term i.e., (2.106) to calculate the mean.

In this problem, we are asked to solve the problem from another perspective: we need to write the joint distribution $p(\boldsymbol{x}, \boldsymbol{y})$ and then perform integration over $\boldsymbol{x}$ to obtain marginal distribution $p(\boldsymbol{y})$. Let's begin by write the quadratic form in the exponential of $p(\boldsymbol{x}, \boldsymbol{y})$:

$$-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Lambda}(\boldsymbol{x} - \boldsymbol{\mu}) - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})^T \boldsymbol{L}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})$$

We extract those terms involving $\boldsymbol{x}$:

$$\begin{aligned} &= -\frac{1}{2}\boldsymbol{x}^T(\boldsymbol{\Lambda} + \boldsymbol{A}^T\boldsymbol{L}\boldsymbol{A})\boldsymbol{x} + \boldsymbol{x}^T[\boldsymbol{\Lambda}\boldsymbol{\mu} + \boldsymbol{A}^T\boldsymbol{L}(\boldsymbol{y} - \boldsymbol{b})] + const \\ &= -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{m})^T(\boldsymbol{\Lambda} + \boldsymbol{A}^T\boldsymbol{L}\boldsymbol{A})(\boldsymbol{x} - \boldsymbol{m}) + \frac{1}{2}\boldsymbol{m}^T(\boldsymbol{\Lambda} + \boldsymbol{A}^T\boldsymbol{L}\boldsymbol{A})\boldsymbol{m} + const \end{aligned}$$

Where we have defined :

$$\boldsymbol{m} = (\boldsymbol{\Lambda} + \boldsymbol{A}^T\boldsymbol{L}\boldsymbol{A})^{-1}[\boldsymbol{\Lambda}\boldsymbol{\mu} + \boldsymbol{A}^T\boldsymbol{L}(\boldsymbol{y} - \boldsymbol{b})]$$

Now if we perform integration over $\boldsymbol{x}$, we will see that the first term vanish to a constant, and we extract the terms including $\boldsymbol{y}$ from the remaining parts, we can obtain :

$$\begin{aligned} = \; & -\frac{1}{2}\boldsymbol{y}^T\left[\boldsymbol{L} - \boldsymbol{L}\boldsymbol{A}(\boldsymbol{\Lambda} + \boldsymbol{A}^T\boldsymbol{L}\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{L}\right]\boldsymbol{y} \\ & + \boldsymbol{y}^T\Big\{\left[\boldsymbol{L} - \boldsymbol{L}\boldsymbol{A}(\boldsymbol{\Lambda} + \boldsymbol{A}^T\boldsymbol{L}\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{L}\right]\boldsymbol{b} \\ & + \boldsymbol{L}\boldsymbol{A}(\boldsymbol{\Lambda} + \boldsymbol{A}^T\boldsymbol{L}\boldsymbol{A})^{-1}\boldsymbol{\Lambda}\boldsymbol{\mu}\Big\} \end{aligned}$$

We firstly view the quadratic term to obtain the precision matrix, and then we take advantage of (2.289), we will obtain (2.110). Finally, using the

linear term combined with the already known covariance matrix, we can obtain (2.109).

Problem 2.33 Solution

According to Bayesian Formula, we can write $p(\boldsymbol{x}|\boldsymbol{y}) = \frac{p(\boldsymbol{x},\boldsymbol{y})}{p(\boldsymbol{y})}$, where we have already known the joint distribution $p(\boldsymbol{x},\boldsymbol{y})$ in (2.105) and (2.108), and the marginal distribution $p(\boldsymbol{y})$ in Prob.2.32., we can follow the same procedure in Prob.2.32., i.e. firstly obtain the covariance matrix from the quadratic term and then obtain the mean from the linear term. The details are omitted here.

Problem 2.34 Solution

Let's follow the hint by firstly calculating the derivative of (2.118) with respect to $\boldsymbol{\Sigma}$ and let it equal to 0 :

$$-\frac{N}{2}\frac{\partial}{\partial \boldsymbol{\Sigma}} ln|\boldsymbol{\Sigma}| - \frac{1}{2}\frac{\partial}{\partial \boldsymbol{\Sigma}} \sum_{n=1}^{N} (\boldsymbol{x_n} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x_n} - \boldsymbol{\mu}) = 0$$

By using (C.28), the first term can be reduced to :

$$-\frac{N}{2}\frac{\partial}{\partial \boldsymbol{\Sigma}} ln|\boldsymbol{\Sigma}| = -\frac{N}{2}(\boldsymbol{\Sigma}^{-\mathbf{1}})^T = -\frac{N}{2}\boldsymbol{\Sigma}^{-\mathbf{1}}$$

Provided with the result that the optimal covariance matrix is the sample covariance, we denote sample matrix $\boldsymbol{S}$ as :

$$\boldsymbol{S} = \frac{1}{N}\sum_{n=1}^{N} (\boldsymbol{x_n} - \boldsymbol{\mu})(\boldsymbol{x_n} - \boldsymbol{\mu})^T$$

We rewrite the second term :

$$\begin{aligned} \text{second term} &= -\frac{1}{2}\frac{\partial}{\partial \boldsymbol{\Sigma}} \sum_{n=1}^{N} (\boldsymbol{x_n} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x_n} - \boldsymbol{\mu}) \\ &= -\frac{N}{2}\frac{\partial}{\partial \boldsymbol{\Sigma}} Tr[\boldsymbol{\Sigma}^{-1}\boldsymbol{S}] \\ &= \frac{N}{2}\boldsymbol{\Sigma}^{-1}\boldsymbol{S}\boldsymbol{\Sigma}^{-1} \end{aligned}$$

Where we have taken advantage of the following property, combined with the fact that $\boldsymbol{S}$ and $\boldsymbol{\Sigma}$ is symmetric. (Note : this property can be found in *The Matrix Cookbook*.)

$$\frac{\partial}{\partial \boldsymbol{X}} Tr(\boldsymbol{A}\boldsymbol{X}^{-\mathbf{1}}\boldsymbol{B}) = -(\boldsymbol{X}^{-\mathbf{1}}\boldsymbol{B}\boldsymbol{A}\boldsymbol{X}^{-\mathbf{1}})^T = -(\boldsymbol{X}^{-1})^T\boldsymbol{A}^T\boldsymbol{B}^T(\boldsymbol{X}^{-1})^T$$

Thus we obtain :

$$-\frac{N}{2}\boldsymbol{\Sigma}^{-\mathbf{1}} + \frac{N}{2}\boldsymbol{\Sigma}^{-1}\boldsymbol{S}\boldsymbol{\Sigma}^{-1} = 0$$

Obviously, we obtain $\boldsymbol{\Sigma}^{-1} = \boldsymbol{S}$, just as required.

Problem 2.35 Solution

The proof of (2.62) is quite clear in the main text, i.e., from page 82 to page 83 and hence we won't repeat it here. Let's prove (2.124). We first begin by proving (2.123) :

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

Where we have taken advantage of the fact that $\boldsymbol{x_n}$ is independently and identically distributed (i.i.d).

Then we use the expression in (2.122) :

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

By using (2.291), the first term will equal to :

$$\text{first term} = \frac{1}{N}\cdot N(\boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\Sigma}) = \boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\Sigma}$$

The second term will equal to :

$$\begin{aligned}
\text{second term} &= -2\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[\boldsymbol{\mu_{ML}}\boldsymbol{x_n}^T] \\
&= -2\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[\frac{1}{N}(\sum_{m=1}^{N}\boldsymbol{x_m})\boldsymbol{x_n}^T] \\
&= -2\frac{1}{N^2}\sum_{n=1}^{N}\sum_{m=1}^{N}\mathbb{E}[\boldsymbol{x_m}\boldsymbol{x_n}^T] \\
&= -2\frac{1}{N^2}\sum_{n=1}^{N}\sum_{m=1}^{N}(\boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{I}_{nm}\boldsymbol{\Sigma}) \\
&= -2\frac{1}{N^2}(N^2\boldsymbol{\mu}\boldsymbol{\mu}^T + N\boldsymbol{\Sigma}) \\
&= -2(\boldsymbol{\mu}\boldsymbol{\mu}^T + \frac{1}{N}\boldsymbol{\Sigma})
\end{aligned}$$

Similarly, the third term will equal to :

$$\begin{aligned}
\text{third term} &= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[\boldsymbol{\mu_{ML}}\boldsymbol{\mu_{ML}^{T}}] \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[(\frac{1}{N}\sum_{j=1}^{N}\boldsymbol{x_j})\cdot(\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x_i})] \\
&= \frac{1}{N^3}\sum_{n=1}^{N}\mathbb{E}[(\sum_{j=1}^{N}\boldsymbol{x_j})\cdot(\sum_{i=1}^{N}\boldsymbol{x_i})] \\
&= \frac{1}{N^3}\sum_{n=1}^{N}(N^2\boldsymbol{\mu}\boldsymbol{\mu}^T + N\boldsymbol{\Sigma}) \\
&= \boldsymbol{\mu}\boldsymbol{\mu}^T + \frac{1}{N}\boldsymbol{\Sigma}
\end{aligned}$$

Finally, we combine those three terms, which gives:

$$\mathbb{E}[\boldsymbol{\Sigma_{ML}}] = \frac{N-1}{N}\boldsymbol{\Sigma}$$

Note: the same procedure from (2.59) to (2.62) can be carried out to prove (2.291) and the only difference is that we need to introduce index m and n to represent the samples. (2.291) is quite straightforward if we see it in this way: If $m = n$, which means $\boldsymbol{x_n}$ and $\boldsymbol{x_m}$ are actually the same sample, (2.291) will reduce to (2.262) (i.e. the correlation between different dimensions exists) and if $m \neq n$, which means $\boldsymbol{x_n}$ and $\boldsymbol{x_m}$ are different samples, also i.i.d, then no correlation should exist, we can guess $\mathbb{E}[\boldsymbol{x_n}\boldsymbol{x_m}^T] = \boldsymbol{\mu}\boldsymbol{\mu}^T$ in this case.

Problem 2.36 Solution

Let's follow the hint. However, firstly we will find the sequential expression based on definition, which will make the latter process on finding coefficient a_{N-1} more easily. Suppose we have N observations in total, and then we can write:

$$\begin{aligned}
\sigma_{ML}^{2(N)} &= \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{ML}^{(N)})^2 \\
&= \frac{1}{N}\left[\sum_{n=1}^{N-1}(x_n - \mu_{ML}^{(N)})^2 + (x_N - \mu_{ML}^{(N)})^2\right] \\
&= \frac{N-1}{N}\frac{1}{N-1}\sum_{n=1}^{N-1}(x_n - \mu_{ML}^{(N)})^2 + \frac{1}{N}(x_N - \mu_{ML}^{(N)})^2 \\
&= \frac{N-1}{N}\sigma_{ML}^{2(N-1)} + \frac{1}{N}(x_N - \mu_{ML}^{(N)})^2 \\
&= \sigma_{ML}^{2(N-1)} + \frac{1}{N}\left[(x_N - \mu_{ML}^{(N)})^2 - \sigma_{ML}^{2(N-1)}\right]
\end{aligned}$$

And then let us write the expression for σ_{ML}.

$$\frac{\partial}{\partial \sigma^2}\left\{\frac{1}{N}\sum_{n=1}^{N} lnp(x_n|\mu,\sigma)\right\}\Big|_{\sigma_{ML}} = 0$$

By exchanging the summation and the derivative, and letting $N \to +\infty$, we can obtain :

$$\lim_{N\to+\infty}\frac{1}{N}\sum_{n=1}^{N}\frac{\partial}{\partial\sigma^2}lnp(x_n|\mu,\sigma) = \mathbb{E}_x\left[\frac{\partial}{\partial\sigma^2}lnp(x_n|\mu,\sigma)\right]$$

Comparing it with (2.127), we can obtain the sequential formula to estimate σ_{ML} :

$$\begin{aligned}\sigma_{ML}^{2(N)} &= \sigma_{ML}^{2(N-1)} + a_{N-1}\frac{\partial}{\partial\sigma_{ML}^{2(N-1)}}lnp(x_N|\mu_{ML}^{(N)},\sigma_{ML}^{(N-1)}) \quad (*)\\ &= \sigma_{ML}^{2(N-1)} + a_{N-1}\left[-\frac{1}{2\sigma_{ML}^{2(N-1)}} + \frac{(x_N-\mu_{ML}^{(N)})^2}{2\sigma_{ML}^{4(N-1)}}\right]\end{aligned}$$

Where we use $\sigma_{ML}^{2(N)}$ to represent the Nth estimation of σ_{ML}^2, i.e., the estimation of σ_{ML}^2 after the Nth observation. What's more, if we choose :

$$a_{N-1} = \frac{2\sigma_{ML}^{4(N-1)}}{N}$$

Then we will obtain :

$$\sigma_{ML}^{2(N)} = \sigma_{ML}^{2(N-1)} + \frac{1}{N}\left[-\sigma_{ML}^{2(N-1)} + (x_N-\mu_{ML}^{(N)})^2\right]$$

We can see that the results are the same. An important thing should be noticed : In maximum likelihood, when estimating variance $\sigma_{ML}^{2(N)}$, we will first estimate mean $\mu_{ML}^{(N)}$, and then we we will calculate variance $\sigma_{ML}^{2(N)}$.

In other words, they are decoupled. It is the same in sequential method. For instance, if we want to estimate both mean and variance sequentially, after observing the Nth sample (i.e., x_N), firstly we can use $\mu_{ML}^{(N-1)}$ together with (2.126) to estimate $\mu_{ML}^{(N)}$ and then use the conclusion in this problem to obtain $\sigma_{ML}^{(N)}$. That is why in $(*)$ we write $lnp(x_N|\mu_{ML}^{(N)},\sigma_{ML}^{(N-1)})$ instead of $lnp(x_N|\mu_{ML}^{(N-1)},\sigma_{ML}^{(N-1)})$.

Problem 2.37 Solution (Wait for revising)

We follow the same procedure in Prob.2.36 to solve this problem. Firstly,

we can obtain the sequential formula based on definition.

$$\begin{aligned}
\boldsymbol{\Sigma}_{\boldsymbol{ML}}^{(\boldsymbol{N})} &= \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x_n}-\boldsymbol{\mu}_{\boldsymbol{ML}}^{(\boldsymbol{N})})(\boldsymbol{x_n}-\boldsymbol{\mu}_{\boldsymbol{ML}}^{(\boldsymbol{N})})^T \\
&= \frac{1}{N}\left[\sum_{n=1}^{N-1}(\boldsymbol{x_n}-\boldsymbol{\mu}_{\boldsymbol{ML}}^{(\boldsymbol{N})})(\boldsymbol{x_n}-\boldsymbol{\mu}_{\boldsymbol{ML}}^{(\boldsymbol{N})})^T+(\boldsymbol{x_N}-\boldsymbol{\mu}_{\boldsymbol{ML}}^{(\boldsymbol{N})})(\boldsymbol{x_N}-\boldsymbol{\mu}_{\boldsymbol{ML}}^{(\boldsymbol{N})})^T\right] \\
&= \frac{N-1}{N}\boldsymbol{\Sigma}_{\boldsymbol{ML}}^{(\boldsymbol{N-1})}+\frac{1}{N}(\boldsymbol{x_N}-\boldsymbol{\mu}_{\boldsymbol{ML}}^{(\boldsymbol{N})})(\boldsymbol{x_N}-\boldsymbol{\mu}_{\boldsymbol{ML}}^{(\boldsymbol{N})})^T \\
&= \boldsymbol{\Sigma}_{\boldsymbol{ML}}^{(\boldsymbol{N-1})}+\frac{1}{N}\left[(\boldsymbol{x_N}-\boldsymbol{\mu}_{\boldsymbol{ML}}^{(\boldsymbol{N})})(\boldsymbol{x_N}-\boldsymbol{\mu}_{\boldsymbol{ML}}^{(\boldsymbol{N})})^T-\boldsymbol{\Sigma}_{\boldsymbol{ML}}^{(\boldsymbol{N-1})}\right]
\end{aligned}$$

If we use *Robbins-Monro sequential estimation formula*, i.e., (2.135), we can obtain :

$$\begin{aligned}
\boldsymbol{\Sigma}_{\boldsymbol{ML}}^{(\boldsymbol{N})} &= \boldsymbol{\Sigma}_{\boldsymbol{ML}}^{(\boldsymbol{N-1})}+\boldsymbol{a}_{N-1}\frac{\partial}{\partial\boldsymbol{\Sigma}_{\boldsymbol{ML}}^{(\boldsymbol{N-1})}}lnp(\boldsymbol{x_N}|\boldsymbol{\mu}_{\boldsymbol{ML}}^{(\boldsymbol{N})},\boldsymbol{\Sigma}_{\boldsymbol{ML}}^{(\boldsymbol{N-1})}) \\
&= \boldsymbol{\Sigma}_{\boldsymbol{ML}}^{(\boldsymbol{N-1})}+\boldsymbol{a}_{N-1}\frac{\partial}{\partial\boldsymbol{\Sigma}_{\boldsymbol{ML}}^{(\boldsymbol{N-1})}}lnp(\boldsymbol{x_N}|\boldsymbol{\mu}_{\boldsymbol{ML}}^{(\boldsymbol{N})},\boldsymbol{\Sigma}_{\boldsymbol{ML}}^{(\boldsymbol{N-1})}) \\
&= \boldsymbol{\Sigma}_{\boldsymbol{ML}}^{(\boldsymbol{N-1})}+\boldsymbol{a}_{N-1}\left[-\frac{1}{2}[\boldsymbol{\Sigma}_{\boldsymbol{ML}}^{(\boldsymbol{N-1})}]^{-1}+\frac{1}{2}[\boldsymbol{\Sigma}_{\boldsymbol{ML}}^{(\boldsymbol{N-1})}]^{-1}(\boldsymbol{x_n}-\boldsymbol{\mu}_{\boldsymbol{ML}}^{(\boldsymbol{N-1})})(\boldsymbol{x_n}-\boldsymbol{\mu}_{\boldsymbol{ML}}^{(\boldsymbol{N-1})})^T[\boldsymbol{\Sigma}_{\boldsymbol{ML}}^{(\boldsymbol{N-1})}]^{-1}\right]
\end{aligned}$$

Where we have taken advantage of the procedure we carried out in Prob.2.34 to calculate the derivative, and if we choose :

$$\boldsymbol{a}_{N-1}=\frac{2}{N}\boldsymbol{\Sigma}_{\boldsymbol{ML}}^{\boldsymbol{2(N-1)}}$$

We can see that the equation above will be identical with our previous conclusion based on definition.

Problem 2.38 Solution

It is straightforward. Based on (2.137), (2.138) and (2.139), we focus on the exponential term of the posterior distribution $p(\mu|\boldsymbol{X})$, which gives :

$$-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2=-\frac{1}{2\sigma_N^2}(\mu-\mu_N)^2$$

We rewrite the left side regarding to μ.

$$\text{quadratic term}=-(\frac{N}{2\sigma^2}+\frac{1}{2\sigma_0^2})\,\mu^2$$

$$\text{linear term}=(\frac{\sum_{n=1}^{N}x_n}{\sigma^2}+\frac{\mu_0}{\sigma_0^2})\,\mu$$

We also rewrite the right side regarding to μ, and hence we will obtain :

$$-(\frac{N}{2\sigma^2}+\frac{1}{2\sigma_0^2})\,\mu^2=-\frac{1}{2\sigma_N^2}\mu^2\,,\;(\frac{\sum_{n=1}^N x_n}{\sigma^2}+\frac{\mu_0}{\sigma_0^2})\,\mu=\frac{\mu_N}{\sigma_N^2}\mu$$

Then we will obtain :

$$\frac{1}{\sigma_N^2}=\frac{1}{\sigma_0^2}+\frac{N}{\sigma^2}$$

And with the prior knowledge that $\sum_{n=1}^N x_n = N\cdot\mu_{ML}$, we can write :

$$\begin{aligned}
\mu_N &= \sigma_N^2\cdot(\frac{\sum_{n=1}^N x_n}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}) \\
&= (\frac{1}{\sigma_0^2}+\frac{N}{\sigma^2})^{-1}\cdot(\frac{N\mu_{ML}}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}) \\
&= \frac{\sigma_0^2\sigma^2}{\sigma^2+N\sigma_0^2}\cdot\frac{N\mu_{ML}\sigma_0^2+\mu_0\sigma^2}{\sigma\sigma_0^2} \\
&= \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0+\frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\mu_{ML}
\end{aligned}$$

Problem 2.39 Solution

Let's follow the hint.

$$\frac{1}{\sigma_N^2}=\frac{1}{\sigma_0^2}+\frac{N}{\sigma^2}=\frac{1}{\sigma_0^2}+\frac{N-1}{\sigma^2}+\frac{1}{\sigma^2}=\frac{1}{\sigma_{N-1}^2}+\frac{1}{\sigma^2}$$

However, it is complicated to derive a sequential formula for μ_N directly. Based on (2.142), we see that the denominator in (2.141) can be eliminated if we multiply $1/\sigma_N^2$ on both side of (2.141). Therefore we will derive a sequential formula for μ_N/σ_N^2 instead.

$$\begin{aligned}
\frac{\mu_N}{\sigma_N^2} &= \frac{\sigma^2+N\sigma_0^2}{\sigma_0^2\sigma^2}(\frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0+\frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\mu_{ML}^{(N)}) \\
&= \frac{\sigma^2+N\sigma_0^2}{\sigma_0^2\sigma^2}(\frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0+\frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\mu_{ML}^{(N)}) \\
&= \frac{\mu_0}{\sigma_0^2}+\frac{N\mu_{ML}^{(N)}}{\sigma^2}=\frac{\mu_0}{\sigma_0^2}+\frac{\sum_{n=1}^N x_n}{\sigma^2} \\
&= \frac{\mu_0}{\sigma_0^2}+\frac{\sum_{n=1}^{N-1} x_n}{\sigma^2}+\frac{x_N}{\sigma^2} \\
&= \frac{\mu_{N-1}}{\sigma_{N-1}^2}+\frac{x_N}{\sigma^2}
\end{aligned}$$

Another possible solution is also given in the problem. We solve it by completing the square.

$$-\frac{1}{2\sigma^2}(x_N-\mu)^2-\frac{1}{2\sigma_{N-1}^2}(\mu-\mu_{N-1})^2=-\frac{1}{2\sigma_N^2}(\mu-\mu_N)^2$$

By comparing the quadratic and linear term regarding to μ, we can obtain:

$$\frac{1}{\sigma_N^2}=\frac{1}{\sigma^2}+\frac{1}{\sigma_{N-1}^2}$$

And :

$$\frac{\mu_N}{\sigma_N^2}=\frac{x_N}{\sigma^2}+\frac{\mu_{N-1}}{\sigma_{N-1}^2}$$

It is the same as previous result. Note: after obtaining the Nth observation, we will firstly use the sequential formula to calculate σ_N^2, and then μ_N. This is because the sequential formula for μ_N is dependent on σ_N^2.

Problem 2.40 Solution

Based on *Bayes Theorem*, we can write :

$$p(\boldsymbol{\mu}|\boldsymbol{X})\propto p(\boldsymbol{X}|\boldsymbol{\mu})p(\boldsymbol{\mu})$$

We focus on the exponential term on the right side and then rearrange it regarding to $\boldsymbol{\mu}$.

$$\begin{aligned}\text{right} &= \left[\sum_{n=1}^{N}-\frac{1}{2}(\boldsymbol{x_n}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x_n}-\boldsymbol{\mu})\right]-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\mu_0})^T{\boldsymbol{\Sigma_0}}^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu_0})\\ &= \left[\sum_{n=1}^{N}-\frac{1}{2}(\boldsymbol{x_n}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x_n}-\boldsymbol{\mu})\right]-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\mu_0})^T{\boldsymbol{\Sigma_0}}^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu_0})\\ &= -\frac{1}{2}\boldsymbol{\mu}({\boldsymbol{\Sigma_0}}^{-1}+N\boldsymbol{\Sigma}^{-1})\boldsymbol{\mu}+\boldsymbol{\mu}^T(\boldsymbol{\Sigma_0^{-1}}\boldsymbol{\mu_0}+\boldsymbol{\Sigma^{-1}}\sum_{\boldsymbol{n=1}}^{\boldsymbol{N}}\boldsymbol{x_n})+\text{const}\end{aligned}$$

Where 'const' represents all the constant terms independent of $\boldsymbol{\mu}$. According to the quadratic term, we can obtain the posterior covariance matrix.

$$\boldsymbol{\Sigma_N^{-1}}={\boldsymbol{\Sigma_0}}^{-1}+N\boldsymbol{\Sigma}^{-1}$$

Then using the linear term, we can obtain :

$$\boldsymbol{\Sigma_N^{-1}}\boldsymbol{\mu_N}=(\boldsymbol{\Sigma_0^{-1}}\boldsymbol{\mu_0}+\boldsymbol{\Sigma^{-1}}\sum_{\boldsymbol{n=1}}^{\boldsymbol{N}}\boldsymbol{x_n})$$

Finally we obtain posterior mean :

$$\boldsymbol{\mu_N}=({\boldsymbol{\Sigma_0}}^{-1}+N\boldsymbol{\Sigma}^{-1})^{-1}(\boldsymbol{\Sigma_0^{-1}}\boldsymbol{\mu_0}+\boldsymbol{\Sigma^{-1}}\sum_{\boldsymbol{n=1}}^{\boldsymbol{N}}\boldsymbol{x_n})$$

Which can also be written as :

$$\boldsymbol{\mu_N} = (\boldsymbol{\Sigma_0}^{-1} + N\boldsymbol{\Sigma}^{-1})^{-1}(\boldsymbol{\Sigma_0}^{-1}\boldsymbol{\mu_0} + \boldsymbol{\Sigma}^{-1}N\boldsymbol{\mu_{ML}})$$

Problem 2.41 Solution

Let's compute the integral of (2.146) over λ.

$$\begin{aligned}
\int_0^{+\infty} \frac{1}{\Gamma(a)} b^a \lambda^{a-1} exp(-b\lambda)\, d\lambda &= \frac{b^a}{\Gamma(a)} \int_0^{+\infty} \lambda^{a-1} exp(-b\lambda)\, d\lambda \\
&= \frac{b^a}{\Gamma(a)} \int_0^{+\infty} (\frac{u}{b})^{a-1} exp(-u) \frac{1}{b}\, du \\
&= \frac{1}{\Gamma(a)} \int_0^{+\infty} u^{a-1} exp(-u)\, du \\
&= \frac{1}{\Gamma(a)} \cdot \Gamma(a) = 1
\end{aligned}$$

Where we first perform change of variable $b\lambda = u$, and then take advantage of the definition of gamma function:

$$\Gamma(x) = \int_0^{+\infty} u^{x-1} e^{-u}\, du$$

Problem 2.42 Solution

We first calculate its mean.

$$\begin{aligned}
\int_0^{+\infty} \lambda \frac{1}{\Gamma(a)} b^a \lambda^{a-1} exp(-b\lambda)\, d\lambda &= \frac{b^a}{\Gamma(a)} \int_0^{+\infty} \lambda^a exp(-b\lambda)\, d\lambda \\
&= \frac{b^a}{\Gamma(a)} \int_0^{+\infty} (\frac{u}{b})^a exp(-u) \frac{1}{b}\, du \\
&= \frac{1}{\Gamma(a) \cdot b} \int_0^{+\infty} u^a exp(-u)\, du \\
&= \frac{1}{\Gamma(a) \cdot b} \cdot \Gamma(a+1) = \frac{a}{b}
\end{aligned}$$

Where we have taken advantage of the property $\Gamma(a+1) = a\Gamma(a)$. Then we calculate $\mathbb{E}[\lambda^2]$.

$$\begin{aligned}
\int_0^{+\infty} \lambda^2 \frac{1}{\Gamma(a)} b^a \lambda^{a-1} exp(-b\lambda)\, d\lambda &= \frac{b^a}{\Gamma(a)} \int_0^{+\infty} \lambda^{a+1} exp(-b\lambda)\, d\lambda \\
&= \frac{b^a}{\Gamma(a)} \int_0^{+\infty} (\frac{u}{b})^{a+1} exp(-u) \frac{1}{b}\, du \\
&= \frac{1}{\Gamma(a) \cdot b^2} \int_0^{+\infty} u^{a+1} exp(-u)\, du \\
&= \frac{1}{\Gamma(a) \cdot b^2} \cdot \Gamma(a+2) = \frac{a(a+1)}{b^2}
\end{aligned}$$

Therefore, according to $var[\lambda] = \mathbb{E}[\lambda^2] - \mathbb{E}[\lambda]^2$, we can obtain :

$$var[\lambda] = \mathbb{E}[\lambda^2] - \mathbb{E}[\lambda]^2 = \frac{a(a+1)}{b^2} - (\frac{a}{b})^2 = \frac{a}{b^2}$$

For the mode of a gamma distribution, we need to find where the maximum of the PDF occurs, and hence we will calculate the derivative of the gamma distribution with respect to λ.

$$\frac{d}{d\lambda}\left[\frac{1}{\Gamma(a)} b^a \lambda^{a-1} exp(-b\lambda)\right] = [(a-1) - b\lambda]\frac{1}{\Gamma(a)} b^a \lambda^{a-2} exp(-b\lambda)$$

It is obvious that $\text{Gam}(\lambda|a,b)$ has its maximum at $\lambda = (a-1)/b$. In other words, the gamma distribution $\text{Gam}(\lambda|a,b)$ has mode $(a-1)/b$.

Problem 2.43 Solution

Let's firstly calculate the following integral.

$$\begin{aligned}
\int_{-\infty}^{+\infty} exp(-\frac{|x|^q}{2\sigma^2})\,dx &= 2\int_{-\infty}^{+\infty} exp(-\frac{x^q}{2\sigma^2})\,dx \\
&= 2\int_0^{+\infty} exp(-u)\frac{(2\sigma^2)^{\frac{1}{q}}}{q} u^{\frac{1}{q}-1}\,du \\
&= 2\frac{(2\sigma^2)^{\frac{1}{q}}}{q}\int_0^{+\infty} exp(-u)u^{\frac{1}{q}-1}\,dx \\
&= 2\frac{(2\sigma^2)^{\frac{1}{q}}}{q}\Gamma(\frac{1}{q})
\end{aligned}$$

And then it is obvious that (2.293) is normalized. Next, we consider about the log likelihood function. Since $\epsilon = t - y(\boldsymbol{x},\boldsymbol{w})$ and $\epsilon \sim p(\epsilon|\sigma^2,q)$, we can write:

$$\begin{aligned}
\ln p(\mathbf{t}|\boldsymbol{X},\boldsymbol{w},\sigma^2) &= \sum_{n=1}^{N} \ln p\left(y(\boldsymbol{x_n},\boldsymbol{w}) - t_n|\sigma^2,q\right) \\
&= -\frac{1}{2\sigma^2}\sum_{n=1}^{N}|y(\boldsymbol{x_n},\boldsymbol{w}) - t_n|^q + N\cdot\ln\left[\frac{q}{2(2\sigma^2)^{1/q}\Gamma(1/q)}\right] \\
&= -\frac{1}{2\sigma^2}\sum_{n=1}^{N}|y(\boldsymbol{x_n},\boldsymbol{w}) - t_n|^q - \frac{N}{q}\ln(2\sigma^2) + \text{const}
\end{aligned}$$

Problem 2.44 Solution

Here we use a simple method to solve this problem by taking advantage of (2.152) and (2.153). By writing the prior distribution in the form of (2.153), i.e., $p(\mu,\lambda|\beta,c,d)$, we can easily obtain the posterior distribution.

$$\begin{aligned}
p(\mu,\lambda|\boldsymbol{X}) &\propto p(\boldsymbol{X}|\mu,\lambda)\cdot p(\mu,\lambda) \\
&\propto \left[\lambda^{1/2} exp(-\frac{\lambda\mu^2}{2})\right]^{N+\beta} exp\left[(c+\sum_{n=1}^{N} x_n)\lambda\mu - (d+\sum_{n=1}^{N}\frac{x_n^2}{2})\lambda\right]
\end{aligned}$$

Therefore, we can see that the posterior distribution has parameters: $\beta' = \beta + N$, $c' = c + \sum_{n=1}^{N} x_n$, $d' = d + \sum_{n=1}^{N} \frac{x_n^2}{2}$. And since the prior distribution is actually the product of a Gaussian distribution and a Gamma distribution:

$$p(\mu, \lambda | \mu_0, \beta, a, b) = \mathcal{N}\left[\mu | \mu_0, (\beta\lambda)^{-1}\right] \mathrm{Gam}(\lambda | a, b)$$

Where $\mu_0 = c/\beta$, $a = 1 + \beta/2$, $b = d - c^2/2\beta$. Hence the posterior distribution can also be written as the product of a Gaussian distribution and a Gamma distribution.

$$p(\mu, \lambda | \boldsymbol{X}) = \mathcal{N}\left[\mu | \mu_0', (\beta'\lambda)^{-1}\right] \mathrm{Gam}(\lambda | a', b')$$

Where we have defined:

$$\mu_0' = c'/\beta' = (c + \sum_{n=1}^{N} x_n)/(N + \beta)$$

$$a' = 1 + \beta'/2 = 1 + (N + \beta)/2$$

$$b' = d' - c'^2/2\beta' = d + \sum_{n=1}^{N} \frac{x_n^2}{2} - (c + \sum_{n=1}^{N} x_n)^2/(2(\beta + N))$$

Problem 2.45 Solution

Let's begin by writing down the dependency of the prior distribution $\mathcal{W}(\boldsymbol{\Lambda} | \boldsymbol{W}, v)$ and the likelihood function $p(\boldsymbol{X} | \boldsymbol{\mu}, \boldsymbol{\Lambda})$ on $\boldsymbol{\Lambda}$.

$$p(\boldsymbol{X} | \boldsymbol{\mu}, \boldsymbol{\Lambda}) \propto |\boldsymbol{\Lambda}|^{N/2} exp\big[\sum_{n=1}^{N} -\frac{1}{2}(\boldsymbol{x_n} - \boldsymbol{\mu})^T \boldsymbol{\Lambda}(\boldsymbol{x_n} - \boldsymbol{\mu})\big]$$

And if we denote

$$\boldsymbol{S} = \frac{1}{N} \sum_{n=1}^{N} (\boldsymbol{x_n} - \boldsymbol{\mu})(\boldsymbol{x_n} - \boldsymbol{\mu})^T$$

Then we can rewrite the equation above as:

$$p(\boldsymbol{X} | \boldsymbol{\mu}, \boldsymbol{\Lambda}) \propto |\boldsymbol{\Lambda}|^{N/2} exp\big[-\frac{1}{2}\mathrm{Tr}(\boldsymbol{S}\boldsymbol{\Lambda})\big]$$

Just as what we have done in Prob.2.34, and comparing this problem with Prob.2.34, one important thing should be noticed: since $\boldsymbol{S}$ and $\boldsymbol{\Lambda}$ are both symmetric, we have: $\mathrm{Tr}(\boldsymbol{S}\boldsymbol{\Lambda}) = \mathrm{Tr}\big((\boldsymbol{S}\boldsymbol{\Lambda})^T\big) = \mathrm{Tr}(\boldsymbol{\Lambda}^{\boldsymbol{T}}\boldsymbol{S}^{\boldsymbol{T}}) = \mathrm{Tr}(\boldsymbol{\Lambda}\boldsymbol{S})$. And we can also write down the prior distribution as:

$$\mathcal{W}(\boldsymbol{\Lambda} | \boldsymbol{W}, v) \propto |\boldsymbol{\Lambda}|^{(v-D-1)/2} exp\big[-\frac{1}{2}\mathrm{Tr}(\boldsymbol{W}^{-1}\boldsymbol{\Lambda})\big]$$

Therefore, the posterior distribution can be obtained:

$$\begin{aligned} p(\boldsymbol{\Lambda} | \boldsymbol{X}, \boldsymbol{W}, v) &\propto p(\boldsymbol{X} | \boldsymbol{\mu}, \boldsymbol{\Lambda}) \cdot \mathcal{W}(\boldsymbol{\Lambda} | \boldsymbol{W}, v) \\ &\propto |\boldsymbol{\Lambda}|^{(N+v-D-1)/2} exp\big\{-\frac{1}{2}\mathrm{Tr}\big[(\boldsymbol{W}^{-1} + \boldsymbol{S})\boldsymbol{\Lambda}\big]\big\} \end{aligned}$$

Therefore, $p(\boldsymbol{\Lambda}|\boldsymbol{X},\boldsymbol{W},v)$ is also a *Wishart* distribution, with parameters:

$$v_N = N + v$$

$$\boldsymbol{W}_N = (\boldsymbol{W}^{-1} + \boldsymbol{S})^{-1}$$

Problem 2.46 Solution

It is quite straightforward.

$$\begin{aligned} p(x|\mu,a,b) &= \int_0^\infty \mathcal{N}(x|\mu,\tau^{-1})\mathrm{Gam}(\tau|a,b)d\tau \\ &= \int_0^\infty \frac{b^a exp(-b\tau)\tau^{a-1}}{\Gamma(a)}(\frac{\tau}{2\pi})^{1/2} exp\left\{-\frac{\tau}{2}(x-\mu)^2\right\} d\tau \\ &= \frac{b^a}{\Gamma(a)}(\frac{1}{2\pi})^{1/2}\int_0^\infty \tau^{a-1/2} exp\left\{-b\tau - \frac{\tau}{2}(x-\mu)^2\right\} d\tau \end{aligned}$$

And if we make change of variable: $z = \tau[b + (x-\mu)^2/2]$, the integral above can be written as:

$$\begin{aligned} p(x|\mu,a,b) &= \frac{b^a}{\Gamma(a)}(\frac{1}{2\pi})^{1/2}\int_0^\infty \tau^{a-1/2} exp\left\{-b\tau - \frac{\tau}{2}(x-\mu)^2\right\} d\tau \\ &= \frac{b^a}{\Gamma(a)}(\frac{1}{2\pi})^{1/2}\int_0^\infty \left[\frac{z}{b+(x-\mu)^2/2}\right]^{a-1/2} exp\{-z\}\frac{1}{b+(x-\mu)^2/2}dz \\ &= \frac{b^a}{\Gamma(a)}(\frac{1}{2\pi})^{1/2}\left[\frac{1}{b+(x-\mu)^2/2}\right]^{a+1/2}\int_0^\infty z^{a-1/2} exp\{-z\}dz \\ &= \frac{b^a}{\Gamma(a)}(\frac{1}{2\pi})^{1/2}\left[b+\frac{(x-\mu)^2}{2}\right]^{-a-1/2}\Gamma(a+1/2) \end{aligned}$$

And if we substitute $a = v/2$ and $b = v/2\lambda$, we will obtain (2.159).

Problem 2.47 Solution

We focus on the dependency of (2.159) on x.

$$\begin{aligned} \mathrm{St}(x|\mu,\lambda,v) &\propto \left[1+\frac{\lambda(x-\mu)^2}{v}\right]^{-v/2-1/2} \\ &\propto exp\left[\frac{-v-1}{2}ln(1+\frac{\lambda(x-\mu)^2}{v})\right] \\ &\propto exp\left[\frac{-v-1}{2}(\frac{\lambda(x-\mu)^2}{v}+O(v^{-2}))\right] \\ &\approx exp\left[-\frac{\lambda(x-\mu)^2}{2}\right] \quad (v\to\infty) \end{aligned}$$

Where we have used *Taylor Expansion*: $ln(1+\epsilon) = \epsilon + O(\epsilon^2)$. We see that this, up to an overall constant, is a Gaussian distribution with mean μ and precision λ.

Problem 2.48 Solution

The same steps in Prob.2.46 can be used here.

$$\begin{aligned}
\mathrm{St}(\boldsymbol{x}\,|\,\boldsymbol{\mu},\boldsymbol{\Lambda},\upsilon) &= \int_0^{+\infty} \mathcal{N}(\boldsymbol{x}\,|\,\boldsymbol{\mu},(\eta\boldsymbol{\Lambda})^{-1})\cdot \mathrm{Gam}(\eta\,|\,\frac{\upsilon}{2},\frac{\upsilon}{2})\; d\eta \\
&= \int_0^{+\infty} \frac{1}{(2\pi)^{D/2}}|\eta\boldsymbol{\Lambda}|^{1/2} exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T(\eta\boldsymbol{\Lambda})(\boldsymbol{x}-\boldsymbol{\mu})-\frac{\upsilon\eta}{2}\right\}\frac{1}{\Gamma(\upsilon/2)}(\frac{\upsilon}{2})^{\upsilon/2}\eta^{\upsilon/2-1}\,d\eta \\
&= \frac{(\upsilon/2)^{\upsilon/2}\,|\boldsymbol{\Lambda}|^{1/2}}{(2\pi)^{D/2}\,\Gamma(\upsilon/2)}\int_0^{+\infty} exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T(\eta\boldsymbol{\Lambda})(\boldsymbol{x}-\boldsymbol{\mu})-\frac{\upsilon\eta}{2}\right\}\eta^{D/2+\upsilon/2-1}\,d\eta
\end{aligned}$$

Where we have taken advantage of the property: $|\eta\boldsymbol{\Lambda}| = \eta^D|\boldsymbol{\Lambda}|$, and if we denote:

$$\boldsymbol{\Delta}^2 = (\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Lambda}(\boldsymbol{x}-\boldsymbol{\mu}) \quad \text{and} \quad z = \frac{\eta}{2}(\boldsymbol{\Delta}^2+\upsilon)$$

The expression above can be reduced to :

$$\begin{aligned}
\mathrm{St}(\boldsymbol{x}\,|\,\boldsymbol{\mu},\boldsymbol{\Lambda},\upsilon) &= \frac{(\upsilon/2)^{\upsilon/2}\,|\boldsymbol{\Lambda}|^{1/2}}{(2\pi)^{D/2}\,\Gamma(\upsilon/2)}\int_0^{+\infty} exp(-z)(\frac{2z}{\boldsymbol{\Delta}^2+\upsilon})^{D/2+\upsilon/2-1}\cdot\frac{2}{\boldsymbol{\Delta}^2+\upsilon}\,d\eta \\
&= \frac{(\upsilon/2)^{\upsilon/2}\,|\boldsymbol{\Lambda}|^{1/2}}{(2\pi)^{D/2}\,\Gamma(\upsilon/2)}(\frac{2}{\boldsymbol{\Delta}^2+\upsilon})^{D/2+\upsilon/2}\int_0^{+\infty} exp(-z)\cdot z^{D/2+\upsilon/2-1}\,d\eta \\
&= \frac{(\upsilon/2)^{\upsilon/2}\,|\boldsymbol{\Lambda}|^{1/2}}{(2\pi)^{D/2}\,\Gamma(\upsilon/2)}(\frac{2}{\boldsymbol{\Delta}^2+\upsilon})^{D/2+\upsilon/2}\Gamma(D/2+\upsilon/2)
\end{aligned}$$

And if we rearrange the expression above, we will obtain (2.162) just as required.

Problem 2.49 Solution

Firstly, we notice that if and only if $\boldsymbol{x}=\boldsymbol{\mu}$, $\boldsymbol{\Delta}^2$ equals to 0, so that $\mathrm{St}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Lambda},\upsilon)$ achieves its maximum. In other words, the mode of $\mathrm{St}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Lambda},\upsilon)$ is $\boldsymbol{\mu}$. Then we consider about its mean $\mathbb{E}[\boldsymbol{x}]$.

$$\begin{aligned}
\mathbb{E}[\boldsymbol{x}] &= \int_{\boldsymbol{x}\in\mathbb{R}^D} \mathrm{St}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Lambda},\upsilon)\cdot\boldsymbol{x}\,d\boldsymbol{x} \\
&= \int_{\boldsymbol{x}\in\mathbb{R}^D}\left[\int_0^{+\infty}\mathcal{N}(\boldsymbol{x}\,|\,\boldsymbol{\mu},(\eta\boldsymbol{\Lambda})^{-1})\cdot\mathrm{Gam}(\eta\,|\,\frac{\upsilon}{2},\frac{\upsilon}{2})\;d\eta\,\boldsymbol{x}\right]d\boldsymbol{x} \\
&= \int_{\boldsymbol{x}\in\mathbb{R}^D}\int_0^{+\infty}\boldsymbol{x}\mathcal{N}(\boldsymbol{x}\,|\,\boldsymbol{\mu},(\eta\boldsymbol{\Lambda})^{-1})\cdot\mathrm{Gam}(\eta\,|\,\frac{\upsilon}{2},\frac{\upsilon}{2})\;d\eta\,d\boldsymbol{x} \\
&= \int_0^{+\infty}\left[\int_{\boldsymbol{x}\in\mathbb{R}^D}\boldsymbol{x}\mathcal{N}(\boldsymbol{x}\,|\,\boldsymbol{\mu},(\eta\boldsymbol{\Lambda})^{-1})\,d\boldsymbol{x}\cdot\mathrm{Gam}(\eta\,|\,\frac{\upsilon}{2},\frac{\upsilon}{2})\right]\;d\eta \\
&= \int_0^{+\infty}\left[\boldsymbol{\mu}\cdot\mathrm{Gam}(\eta\,|\,\frac{\upsilon}{2},\frac{\upsilon}{2})\right]\;d\eta \\
&= \boldsymbol{\mu}\int_0^{+\infty}\mathrm{Gam}(\eta\,|\,\frac{\upsilon}{2},\frac{\upsilon}{2})\;d\eta = \boldsymbol{\mu}
\end{aligned}$$

Where we have taken the following property:

$$\int_{\boldsymbol{x}\in\mathbb{R}^D}\boldsymbol{x}\mathcal{N}(\boldsymbol{x}\,|\,\boldsymbol{\mu},(\eta\boldsymbol{\Lambda})^{-1})\,d\boldsymbol{x} = \mathbb{E}[\boldsymbol{x}] = \boldsymbol{\mu}$$

Then we calculate $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T]$. The steps above can also be used here.

$$\begin{aligned}
\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T] &= \int_{\boldsymbol{x}\in\mathbb{R}^D} \mathrm{St}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Lambda},v)\cdot\boldsymbol{x}\boldsymbol{x}^T\,d\boldsymbol{x} \\
&= \int_{\boldsymbol{x}\in\mathbb{R}^D}\left[\int_0^{+\infty}\mathcal{N}(\boldsymbol{x}\,|\,\boldsymbol{\mu},(\eta\boldsymbol{\Lambda})^{-1})\cdot\mathrm{Gam}(\eta\,|\,\frac{v}{2},\frac{v}{2})\;d\eta\,\boldsymbol{x}\boldsymbol{x}^T\right]d\boldsymbol{x} \\
&= \int_{\boldsymbol{x}\in\mathbb{R}^D}\int_0^{+\infty}\boldsymbol{x}\boldsymbol{x}^T\mathcal{N}(\boldsymbol{x}\,|\,\boldsymbol{\mu},(\eta\boldsymbol{\Lambda})^{-1})\cdot\mathrm{Gam}(\eta\,|\,\frac{v}{2},\frac{v}{2})\;d\eta\,d\boldsymbol{x} \\
&= \int_0^{+\infty}\left[\int_{\boldsymbol{x}\in\mathbb{R}^D}\boldsymbol{x}\boldsymbol{x}^T\mathcal{N}(\boldsymbol{x}\,|\,\boldsymbol{\mu},(\eta\boldsymbol{\Lambda})^{-1})\,d\boldsymbol{x}\cdot\mathrm{Gam}(\eta\,|\,\frac{v}{2},\frac{v}{2})\right]\;d\eta \\
&= \int_0^{+\infty}\left[\mathbb{E}[\boldsymbol{\mu}\boldsymbol{\mu}^T]\cdot\mathrm{Gam}(\eta\,|\,\frac{v}{2},\frac{v}{2})\right]\;d\eta \\
&= \int_0^{+\infty}\left[\boldsymbol{\mu}\boldsymbol{\mu}^T+(\eta\boldsymbol{\Lambda})^{-1}\right]\mathrm{Gam}(\eta\,|\,\frac{v}{2},\frac{v}{2})\;d\eta \\
&= \boldsymbol{\mu}\boldsymbol{\mu}^T+\int_0^{+\infty}(\eta\boldsymbol{\Lambda})^{-1}\cdot\mathrm{Gam}(\eta\,|\,\frac{v}{2},\frac{v}{2})\;d\eta \\
&= \boldsymbol{\mu}\boldsymbol{\mu}^T+\int_0^{+\infty}(\eta\boldsymbol{\Lambda})^{-1}\cdot\frac{1}{\Gamma(v/2)}(\frac{v}{2})^{v/2}\eta^{v/2-1}exp(-\frac{v}{2}\eta)\,d\eta \\
&= \boldsymbol{\mu}\boldsymbol{\mu}^T+\boldsymbol{\Lambda}^{-1}\frac{1}{\Gamma(v/2)}(\frac{v}{2})^{v/2}\int_0^{+\infty}\eta^{v/2-2}exp(-\frac{v}{2}\eta)\,d\eta
\end{aligned}$$

If we denote: $z=\frac{v\eta}{2}$, the equation above can be reduced to :

$$\begin{aligned}
\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T] &= \boldsymbol{\mu}\boldsymbol{\mu}^T+\boldsymbol{\Lambda}^{-1}\frac{1}{\Gamma(v/2)}(\frac{v}{2})^{v/2}\int_0^{+\infty}(\frac{2z}{v})^{v/2-2}exp(-z)\frac{2}{v}\,dz \\
&= \boldsymbol{\mu}\boldsymbol{\mu}^T+\boldsymbol{\Lambda}^{-1}\frac{1}{\Gamma(v/2)}\cdot\frac{v}{2}\int_0^{+\infty}z^{v/2-2}exp(-z)\,dz \\
&= \boldsymbol{\mu}\boldsymbol{\mu}^T+\boldsymbol{\Lambda}^{-1}\frac{\Gamma(v/2-1)}{\Gamma(v/2)}\cdot\frac{v}{2} \\
&= \boldsymbol{\mu}\boldsymbol{\mu}^T+\boldsymbol{\Lambda}^{-1}\frac{1}{v/2-1}\frac{v}{2} \\
&= \boldsymbol{\mu}\boldsymbol{\mu}^T+\frac{v}{v-2}\boldsymbol{\Lambda}^{-1}
\end{aligned}$$

Where we have taken advantage of the property: $\Gamma(x+1)=x\Gamma(x)$, and since we have $cov[\boldsymbol{x}]=\mathbb{E}\left[(\boldsymbol{x}-\mathbb{E}[\boldsymbol{x}])(\boldsymbol{x}-\mathbb{E}[\boldsymbol{x}])^T\right]$, together with $\mathbb{E}[\boldsymbol{x}]=\boldsymbol{\mu}$, we can obtain:

$$cov[\boldsymbol{x}]=\frac{v}{v-2}\boldsymbol{\Lambda}^{-1}$$

Problem 2.50 Solution

The same steps in Prob.2.47 can be used here.

$$\begin{aligned}
\mathrm{St}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Lambda},v) &\propto \left[1+\frac{\boldsymbol{\Delta}^2}{v}\right]^{-D/2-v/2} \\
&\propto exp\left[(-D/2-v/2)\cdot ln(1+\frac{\boldsymbol{\Delta}^2}{v})\right] \\
&\propto exp\left[-\frac{D+v}{2}\cdot(\frac{\boldsymbol{\Delta}^2}{v}+O(v^{-2}))\right] \\
&\approx exp(-\frac{\boldsymbol{\Delta}^2}{2}) \quad (v\to\infty)
\end{aligned}$$

Where we have used *Taylor Expansion*: $ln(1+\epsilon)=\epsilon+O(\epsilon^2)$. And since $\boldsymbol{\Delta}^2=(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Lambda}(\boldsymbol{x}-\boldsymbol{\mu})$, we see that this, up to an overall constant, is a Gaussian distribution with mean $\boldsymbol{\mu}$ and precision $\boldsymbol{\Lambda}$.

Problem 2.51 Solution

We first prove (2.177). Since we have $exp(iA)\cdot exp(-iA)=1$, and $exp(iA)=cosA+isinA$. We can obtain:

$$(cosA+isinA)\cdot(cosA-isinA)=1$$

Which gives $cos^2A+sin^2A=1$. And then we prove (2.178) using the hint.

$$\begin{aligned}
cos(A-B) &= \Re[exp(i(A-B))] \\
&= \Re[exp(iA)/exp(iB)] \\
&= \Re[\frac{cosA+isinA}{cosB+isinB}] \\
&= \Re[\frac{(cosA+isinA)(cosB-isinB)}{(cosB+isinB)(cosB-isinB)}] \\
&= \Re[(cosA+isinA)(cosB-isinB)] \\
&= cosAcosB+sinAsinB
\end{aligned}$$

It is quite similar for (2.183).

$$\begin{aligned}
sin(A-B) &= \Im[exp(i(A-B))] \\
&= \Im[(cosA+isinA)(cosB-isinB)] \\
&= sinAcosB-cosAsinB
\end{aligned}$$

Problem 2.52 Solution

Let's follow the hint. We first derive an approximation for $exp[mcos(\theta-$

$\theta_0)]$.

$$\begin{aligned} exp\{mcos(\theta-\theta_0)\} &= exp\left\{m\left[1-\frac{(\theta-\theta_0)^2}{2}+O((\theta-\theta_0)^4)\right]\right\} \\ &= exp\left\{m-m\frac{(\theta-\theta_0)^2}{2}-mO((\theta-\theta_0)^4)\right\} \\ &= exp(m)\cdot exp\left\{-m\frac{(\theta-\theta_0)^2}{2}\right\}\cdot exp\left\{-mO((\theta-\theta_0)^4)\right\} \end{aligned}$$

It is same for $exp(mcos\theta)$:

$$exp\{mcos\theta\} = exp(m)\cdot exp(-m\frac{\theta^2}{2})\cdot exp\left\{-mO(\theta^4)\right\}$$

Now we rearrange (2.179):

$$\begin{aligned} p(\theta|\theta_0,m) &= \frac{1}{2\pi I_0(m)}exp\{mcos(\theta-\theta_0)\} \\ &= \frac{1}{\int_0^{2\pi}exp\{mcos\theta\}\,d\theta}exp\{mcos(\theta-\theta_0)\} \\ &= \frac{exp(m)\cdot exp\left\{-m\frac{(\theta-\theta_0)^2}{2}\right\}\cdot exp\left\{-mO((\theta-\theta_0)^4)\right\}}{\int_0^{2\pi}exp(m)\cdot exp(-m\frac{\theta^2}{2})\cdot exp\left\{-mO(\theta^4)\right\}\,d\theta} \\ &= \frac{1}{\int_0^{2\pi}exp(-m\frac{\theta^2}{2})d\theta}exp\left\{-m\frac{(\theta-\theta_0)^2}{2}\right\} \end{aligned}$$

Where we have taken advantage of the following fact:

$$exp\left\{-mO((\theta-\theta_0)^4)\right\}\approx exp\left\{-mO(\theta^4)\right\} \quad (\text{when } m\to\infty)$$

Therefore, it is straightforward that when $m\to\infty$, (2.179) reduces to a Gaussian Distribution with mean θ_0 and precision m.

Problem 2.53 Solution

Let's rearrange (2.182) according to (2.183).

$$\begin{aligned} \sum_{n=1}^{N}sin(\theta-\theta_0) &= \sum_{n=1}^{N}(sin\theta_n cos\theta_0 - cos\theta_n sin\theta_0) \\ &= cos\theta_0\sum_{n=1}^{N}sin\theta_n - sin\theta_0\sum_{n=1}^{N}cos\theta_n \end{aligned}$$

Where we have used (2.183), and then together with (2.182), we can obtain :

$$cos\theta_0\sum_{n=1}^{N}sin\theta_n - sin\theta_0\sum_{n=1}^{N}cos\theta_n = 0$$

Which gives:

$$\theta_0^{ML} = tan^{-1}\left\{\frac{\sum_n sin\theta_n}{\sum_n cos\theta_n}\right\}$$

Problem 2.54 Solution

We calculate the first and second derivative of (2.179) with respect to θ.

$$p(\theta|\theta_0,m)' = \frac{1}{2\pi I_0(m)}[-msin(\theta-\theta_0)]\,exp\{mcos(\theta-\theta_0)\}$$

$$p(\theta|\theta_0,m)'' = \frac{1}{2\pi I_0(m)}\left[-mcos(\theta-\theta_0)+(-msin(\theta-\theta_0))^2\right]exp\{mcos(\theta-\theta_0)\}$$

If we let $p(\theta|\theta_0,m)'$ equals to 0, we will obtain its root:

$$\theta = \theta_0 + k\pi \quad (k \in Z)$$

When $k \equiv 0\,(mod 2)$, i.e. $\theta \equiv \theta_0\,(mod 2\pi)$, we have:

$$p(\theta|\theta_0,m)'' = \frac{-m\,exp(m)}{2\pi I_0(m)} < 0$$

Therefore, when $\theta = \theta_0$, (2.179) obtains its maximum. And when $k \equiv 1\,(mod 2)$, i.e. $\theta \equiv \theta_0 + \pi\,(mod 2\pi)$, we have:

$$p(\theta|\theta_0,m)'' = \frac{m\,exp(-m)}{2\pi I_0(m)} > 0$$

Therefore, when $\theta = \theta_0 + \pi\,(mod 2\pi)$, (2.179) obtains its minimum.

Problem 2.55 Solution

According to (2.185), we have :

$$A(m_{ML}) = \frac{1}{N}\sum_{n=1}^{N} cos(\theta_n - \theta_0^{ML})$$

By using (2.178), we can write :

$$\begin{aligned} A(m_{ML}) &= \frac{1}{N}\sum_{n=1}^{N} cos(\theta_n - \theta_0^{ML}) \\ &= \frac{1}{N}\sum_{n=1}^{N}\left(cos\theta_n cos\theta_0^{ML} + sin\theta_n sin\theta_0^{ML}\right) \\ &= \left(\frac{1}{N}\sum_{n=1}^{N} cos\theta_N\right)cos\theta_0^{ML} + \left(\frac{1}{N}\sum_{n=1}^{N} sin\theta_N\right)sin\theta_0^{ML} \end{aligned}$$

By using (2.168), we can further derive:

$$\begin{aligned} A(m_{ML}) &= \left(\frac{1}{N}\sum_{n=1}^{N} cos\theta_N\right)cos\theta_0^{ML} + \left(\frac{1}{N}\sum_{n=1}^{N} sin\theta_N\right)sin\theta_0^{ML} \\ &= \bar{r}cos\bar{\theta}\cdot cos\theta_0^{ML} + \bar{r}sin\bar{\theta}\cdot sin\theta_0^{ML} \\ &= \bar{r}cos(\bar{\theta} - \theta_0^{ML}) \end{aligned}$$

And then by using (2.169) and (2.184), it is obvious that $\bar{\theta} = \theta_0^{ML}$, and hence $A(m_{ML}) = \bar{r}$.

Problem 2.56 Solution

Recall that the distributions belonging to the exponential family have the form:

$$p(\boldsymbol{x}|\boldsymbol{\eta}) = h(\boldsymbol{x})g(\boldsymbol{\eta})exp(\boldsymbol{\eta}^T\boldsymbol{u}(\boldsymbol{x}))$$

And according to (2.13), the beta distribution can be written as:

$$\begin{aligned}\text{Beta}(x|a,b) &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1} \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}exp\left[(a-1)lnx + (b-1)ln(1-x)\right] \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{exp\left[alnx + bln(1-x)\right]}{x(1-x)}\end{aligned}$$

Comparing it with the standard form of exponential family, we can obtain:

$$\begin{cases}\boldsymbol{\eta} = [a,\, b]^T \\ \boldsymbol{u}(x) = [lnx,\, ln(1-x)]^T \\ g(\boldsymbol{\eta}) = \Gamma(\eta_1 + \eta_2)\big/\left[\Gamma(\eta_1)\Gamma(\eta_2)\right] \\ h(x) = 1\big/(x(1-x))\end{cases}$$

Where η_1 means the first element of $\boldsymbol{\eta}$, i.e. $\eta_1 = a-1$, and η_2 means the second element of $\boldsymbol{\eta}$, i.e. $\eta_2 = b-1$. According to (2.146), Gamma distribution can be written as:

$$\text{Gam}(x|a,b) = \frac{1}{\Gamma(a)}b^a x^{a-1}exp(-bx)$$

Comparing it with the standard form of exponential family, we can obtain:

$$\begin{cases}\boldsymbol{\eta} = [a,\, b]^T \\ \boldsymbol{u}(x) = [0,\, -x] \\ g(\boldsymbol{\eta}) = \eta_2^{\eta_1}\Big/\Gamma(\eta_1) \\ h(x) = x^{\eta_1 - 1}\end{cases}$$

According to (2.179), the von Mises distribution can be written as:

$$\begin{aligned}p(x|\theta_0,m) &= \frac{1}{2\pi I_0(m)}exp(mcos(x-\theta_0)) \\ &= \frac{1}{2\pi I_0(m)}exp\left[m(cosxcos\theta_0 + sinxsin\theta_0)\right]\end{aligned}$$

Comparing it with the standard form of exponential family, we can obtain:

$$\begin{cases} \boldsymbol{\eta} = [mcos\theta_0, msin\theta_0]^T \\ \boldsymbol{u}(x) = [cosx, sinx] \\ g(\boldsymbol{\eta}) = 1\Big/2\pi I_0(\sqrt{\eta_1^2 + \eta_2^2}) \\ h(x) = 1 \end{cases}$$

Note : a given distribution can be written into the exponential family in several ways with different natural parameters.

Problem 2.57 Solution

Recall that the distributions belonging to the exponential family have the form:

$$p(\boldsymbol{x}|\boldsymbol{\eta}) = h(\boldsymbol{x})g(\boldsymbol{\eta})exp(\boldsymbol{\eta}^T\boldsymbol{u}(\boldsymbol{x}))$$

And the multivariate Gaussian Distribution has the form:

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\Sigma|^{1}/2}exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}$$

We expand the exponential term with respect to $\boldsymbol{\mu}$.

$$\begin{aligned} \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) &= \frac{1}{(2\pi)^{D/2}}\frac{1}{|\Sigma|^{1/2}}exp\left\{-\frac{1}{2}(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - 2\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \boldsymbol{\mu}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\right\} \\ &= \frac{1}{(2\pi)^{D/2}}\frac{1}{|\Sigma|^{1/2}}exp\left\{-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\right\}exp\left\{-\frac{1}{2}\boldsymbol{\mu}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\right\} \end{aligned}$$

Comparing it with the standard form of exponential family, we can obtain:

$$\begin{cases} \boldsymbol{\eta} = [\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, -\frac{1}{2}vec(\boldsymbol{\Sigma}^{-1})\,]^T \\ \boldsymbol{u}(\boldsymbol{x}) = [\boldsymbol{x}, vec(\boldsymbol{x}\boldsymbol{x}^T)\,] \\ g(\boldsymbol{\eta}) = exp(\frac{1}{4}\boldsymbol{\eta_1}^T\boldsymbol{\eta_2}^{-1}\boldsymbol{\eta_1}) + |-2\boldsymbol{\eta_2}|^{1/2} \\ h(x) = (2\pi)^{-D/2} \end{cases}$$

Where we have used $\boldsymbol{\eta}_1$ to denote the first element of $\boldsymbol{\eta}$, and $\boldsymbol{\eta}_2$ to denote the second element of $\boldsymbol{\eta}$. And we also take advantage of the vectorizing operator, i.e.$vec(\cdot)$. The vectorization of a matrix is a linear transformation which converts the matrix into a column vector. This can be viewed in an example :

$$\boldsymbol{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \quad => \quad vec(\boldsymbol{A}) = [a, c, b, d]^T$$

Note: By introducing vectorizing operator, we actually have $vec(\boldsymbol{\Sigma^{-1}}) \cdot vec(\boldsymbol{x}\boldsymbol{x}^T) = \boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}$

Problem 2.58 Solution (Wait for updating)

Based on (2.226), we rewrite the expression for $\nabla g(\boldsymbol{\eta})$.

$$\nabla g(\boldsymbol{\eta}) = -g(\boldsymbol{\eta})\mathbb{E}[\boldsymbol{u}(\boldsymbol{x})]$$

And then we calculate the derivative of both sides of the equation above with respect to $\boldsymbol{\eta}$.

$$\nabla\nabla g(\boldsymbol{\eta}) = -\left[\nabla g(\boldsymbol{\eta})\mathbb{E}[\boldsymbol{u}(\boldsymbol{x})^T] + g(\boldsymbol{\eta})\nabla\mathbb{E}[\boldsymbol{u}(\boldsymbol{x})^T]\right]$$

If we multiply both sides by $-\frac{1}{g(\boldsymbol{\eta})}$, we can obtain :

$$-\nabla\nabla lng(\boldsymbol{\eta}) = \nabla lng(\boldsymbol{\eta})\mathbb{E}[\boldsymbol{u}(\boldsymbol{x})^T] + \nabla\mathbb{E}[\boldsymbol{u}(\boldsymbol{x})^T]$$

According to (2.225), we calculate $\nabla\mathbb{E}[\boldsymbol{u}(\boldsymbol{x})^T]$.

$$\nabla\mathbb{E}[\boldsymbol{u}(\boldsymbol{x})^T] = \nabla g(\boldsymbol{\eta})\int h(\boldsymbol{x})exp\left\{\boldsymbol{\eta}^T\boldsymbol{u}(\boldsymbol{x})\right\}\boldsymbol{u}(\boldsymbol{x})^T d\boldsymbol{x} + g(\boldsymbol{\eta})\int h(\boldsymbol{x})exp\left\{\boldsymbol{\eta}^T\boldsymbol{u}(\boldsymbol{x})\right\}\boldsymbol{u}(\boldsymbol{x})\boldsymbol{u}(\boldsymbol{x})^T d\boldsymbol{x}$$

$$=> \nabla\mathbb{E}[\boldsymbol{u}(\boldsymbol{x})^T] = \nabla lng(\boldsymbol{\eta})\mathbb{E}[\boldsymbol{u}(\boldsymbol{x})^T] + \mathbb{E}[\boldsymbol{u}(\boldsymbol{x})\boldsymbol{u}(\boldsymbol{x})^T]$$

Therefore, we obtain :

$$-\nabla\nabla lng(\boldsymbol{\eta}) = 2\nabla lng(\boldsymbol{\eta})\mathbb{E}[\boldsymbol{u}(\boldsymbol{x})^T] + \mathbb{E}[\boldsymbol{u}(\boldsymbol{x})\boldsymbol{u}(\boldsymbol{x})^T] = -2\mathbb{E}[\boldsymbol{u}(\boldsymbol{x})]\mathbb{E}[\boldsymbol{u}(\boldsymbol{x})^T] + \mathbb{E}[\boldsymbol{u}(\boldsymbol{x})\boldsymbol{u}(\boldsymbol{x})^T]$$

Problem 2.59 Solution

It is straightforward.

$$\begin{aligned}\int p(x|\sigma)dx &= \int \frac{1}{\sigma}f(\frac{x}{\sigma})dx \\ &= \int \frac{1}{\sigma}f(u)\sigma\, du \\ &= \int f(u)du = 1\end{aligned}$$

Where we have denoted $u = x/\sigma$.

Problem 2.60 Solution

Firstly, we write down the log likelihood function.

$$\sum_{n=1}^{N} lnp(\boldsymbol{x_n}) = \sum_{i=1}^{M} n_i\, ln(h_i)$$

Some details should be explained here. If $\boldsymbol{x_n}$ falls into region Δ_i, then $p(\boldsymbol{x_n})$ will equal to h_i, and since we have already been given that among all the N observations, there are n_i samples fall into region Δ_i, we can easily write down the likelihood function just as the equation above, and note we use

M to denote the number of different regions. Therefore, an implicit equation should hold:

$$\sum_{i=1}^{M} n_i = N$$

We now need to take account of the constraint that $p(\boldsymbol{x})$ must integrate to unity, which can be written as $\sum_{j=1}^{M} h_j \Delta_j = 1$. We introduce a Lagrange multiplier to the expression, and then we need to minimize:

$$\sum_{i=1}^{M} n_i \, ln(h_i) + \lambda(\sum_{j}^{M} h_j \Delta_j - 1)$$

We calculate its derivative with respect to h_i and let it equal to 0.

$$\frac{n_i}{h_i} + \lambda \Delta_i = 0$$

Multiplying both sides by h_i, performing summation over i and then using the constraint, we can obtain:

$$N + \lambda = 0$$

In other words, $\lambda = -N$. Then we substitute the result into the likelihood function, which gives:

$$h_i = \frac{n_i}{N} \frac{1}{\Delta_i}$$

Problem 2.61 Solution

It is straightforward. In *K nearest neighbours* (*KNN*), when we want to estimate probability density at a point $\boldsymbol{x_i}$, we will consider a small sphere centered on $\boldsymbol{x_i}$ and then allow the radius to grow until it contains K data points, and then $p(\boldsymbol{x_i})$ will equal to $K/(NV_i)$, where N is total observations and V_i is the volume of the sphere centered on $\boldsymbol{x_i}$. We can assume that V_i is small enough that $p(\boldsymbol{x_i})$ is roughly constant in it. In this way, We can write down the integral:

$$\int p(\boldsymbol{x}) d\boldsymbol{x} \approx \sum_{i=1}^{N} p(\boldsymbol{x_i}) \cdot V_i = \sum_{i=1}^{N} \frac{K}{NV_i} \cdot V_i = K \neq 1$$

We also see that if we use "*1NN*" ($K = 1$), the probability density will be well normalized. Note that if and only if the volume of all the spheres are small enough and N is large enough, the equation above will hold. Fortunately, these two conditions can be satisfied in *KNN*.

0.3 Probability Distribution

Problem 3.1 Solution

Based on (3.6), we can write :

$$2\sigma(2a)-1 = \frac{2}{1+exp(-2a)} - 1 = \frac{1-exp(-2a)}{1+exp(-2a)} = \frac{exp(a)-exp(-a)}{exp(a)+exp(-a)}$$

Which is exactly $tanh(a)$. Then we will find the relation between μ_i, w_i in (3.101) and (3.102). Let's start from (3.101).

$$\begin{aligned} y(x,\boldsymbol{w}) &= w_0 + \sum_{j=1}^{M} w_j \sigma(\frac{x-\mu_j}{s}) \\ &= w_0 + \sum_{j=1}^{M} w_j \frac{tanh(\frac{x-\mu_j}{2s})+1}{2} \\ &= w_0 + \frac{1}{2}\sum_{j=1}^{M} w_j + \sum_{j=1}^{M} \frac{w_j}{2} tanh(\frac{x-\mu_j}{2s}) \end{aligned}$$

Hence the relation is given by :

$$\mu_0 = w_0 + \frac{1}{2}\sum_{j=1}^{M} w_j \quad \text{and} \quad \mu_j = \frac{w_j}{2}$$

Note: there is a typo in (3.102), the denominator should be $2s$ instead of s, or alternatively you can view it as a new s', which equals to $2s$.

Problem 3.2 Solution

We first need to show that $(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}$ is invertible. Suppose, for the sake of contradiction, $\boldsymbol{c}$ is a nonzero vector in the kernel(Null space) of $\boldsymbol{\Phi}^T\boldsymbol{\Phi}$. Then $\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{c}$ equals to $\boldsymbol{0}$ and so we have:

$$0 = \boldsymbol{c}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{c} = (\boldsymbol{\Phi}\boldsymbol{c})^T\boldsymbol{\Phi}\boldsymbol{c} = ||\boldsymbol{\Phi}\boldsymbol{c}||^2$$

The equation above shows that $\boldsymbol{\Phi}\boldsymbol{c} = \boldsymbol{0}$. However, $\boldsymbol{\Phi}\boldsymbol{c} = c_1\boldsymbol{\phi}_1 + c_2\boldsymbol{\phi}_2 + ... + c_M\boldsymbol{\phi}_M$ and $\{\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, ,..., \boldsymbol{\phi}_M\}$ is a basis for $\boldsymbol{\Phi}$, there is no linear relation between the $\boldsymbol{\phi}_i$ and therefore we cannot have $c_1\boldsymbol{\phi}_1 + c_2\boldsymbol{\phi}_2 + ... + c_M\boldsymbol{\phi}_M = 0$. This is the contradiction. Hence $\boldsymbol{\Phi}^T\boldsymbol{\Phi}$ is invertible. Then let's first prove two specific cases.

Case 1: $\boldsymbol{w}_1$ is in $\boldsymbol{\Phi}$. In this case, we have $\boldsymbol{\Phi}\boldsymbol{c} = \boldsymbol{w}_1$ for some $\boldsymbol{c}$. So we have:

$$\boldsymbol{\Phi}(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\boldsymbol{w}_1 = \boldsymbol{\Phi}(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{c} = \boldsymbol{\Phi}\boldsymbol{c} = \boldsymbol{w}_1$$

Case 2:$\boldsymbol{w}_2$ is in $\boldsymbol{\Phi}^\perp$, where $\boldsymbol{\Phi}^\perp$ is used to denote the *orthogonal complement* of $\boldsymbol{\Phi}$ and then we have $\boldsymbol{\Phi}^T\boldsymbol{w}_2 = \boldsymbol{0}$, which leads to:

$$\boldsymbol{\Phi}(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\boldsymbol{w}_2 = \boldsymbol{0}$$

Recall that any vector $\boldsymbol{x} \in R^M$ can be divided into the summation of two vectors $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$, were $\boldsymbol{w}_1 \in \boldsymbol{\Phi}$ and $\boldsymbol{w}_2 \in \boldsymbol{\Phi}^{\perp}$ separately. And so we have:

$$\boldsymbol{\Phi}(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\boldsymbol{w} = \boldsymbol{\Phi}(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T(\boldsymbol{w}_1 + \boldsymbol{w}_2) = \boldsymbol{w}_1$$

Which is exactly what orthogonal projection is supposed to do.

Problem 3.3 Solution

Let's calculate the derivative of (3.104) with respect to $\boldsymbol{w}$.

$$\nabla E_D(\boldsymbol{w}) = \sum_{n=1}^{N} r_n \left\{ t_n - \boldsymbol{w}^T\boldsymbol{\Phi}(\boldsymbol{x_n}) \right\} \boldsymbol{\Phi}(\boldsymbol{x_n})^T$$

We set the derivative equal to 0.

$$0 = \sum_{n=1}^{N} r_n t_n \boldsymbol{\Phi}(\boldsymbol{x_n})^T - \boldsymbol{w}^T \left(\sum_{n=1}^{N} r_n \boldsymbol{\Phi}(\boldsymbol{x_n})\boldsymbol{\Phi}(\boldsymbol{x_n})^T \right)$$

If we denote $\sqrt{r_n}\,\boldsymbol{\phi}(\boldsymbol{x_n}) = \boldsymbol{\phi}'(\boldsymbol{x_n})$ and $\sqrt{r_n}t_n = t'_n$, we can obtain:

$$0 = \sum_{n=1}^{N} t'_n \boldsymbol{\Phi}'(\boldsymbol{x_n})^T - \boldsymbol{w}^T \left(\sum_{n=1}^{N} \boldsymbol{\Phi}'(\boldsymbol{x_n})\boldsymbol{\Phi}'(\boldsymbol{x_n})^T \right)$$

Taking advantage of (3.11) – (3.17), we can derive a similar result, i.e. $\boldsymbol{w}_{ML} = (\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\boldsymbol{t}$. But here, we define $\boldsymbol{t}$ as:

$$\boldsymbol{t} = \left[\sqrt{r_1}t_1,\, \sqrt{r_2}t_2,\, \ldots,\, \sqrt{r_N}t_N \right]^T$$

We also define $\boldsymbol{\Phi}$ as a $N \times M$ matrix, with element $\boldsymbol{\Phi}(i,j) = \sqrt{r_i}\,\phi_j(\boldsymbol{x_i})$.

Problem 3.4 Solution

Firstly, we rearrange $E_D(\boldsymbol{w})$.

$$\begin{aligned}
E_D(\boldsymbol{w}) &= \frac{1}{2}\sum_{n=1}^{N} \left\{ \left[w_0 + \sum_{i=1}^{D} w_i(x_i + \epsilon_i) \right] - t_n \right\}^2 \\
&= \frac{1}{2}\sum_{n=1}^{N} \left\{ \left(w_0 + \sum_{i=1}^{D} w_i x_i \right) - t_n + \sum_{i=1}^{D} w_i \epsilon_i \right\}^2 \\
&= \frac{1}{2}\sum_{n=1}^{N} \left\{ y(x_n, \boldsymbol{w}) - t_n + \sum_{i=1}^{D} w_i \epsilon_i \right\}^2 \\
&= \frac{1}{2}\sum_{n=1}^{N} \left\{ \left(y(x_n, \boldsymbol{w}) - t_n \right)^2 + \left(\sum_{i=1}^{D} w_i \epsilon_i \right)^2 + 2\left(\sum_{i=1}^{D} w_i \epsilon_i \right)\left(y(x_n, \boldsymbol{w}) - t_n \right) \right\}
\end{aligned}$$

Where we have used $y(x_n, \boldsymbol{w})$ to denote the output of the linear model when input variable is x_n, without noise added. For the second term in the equation above, we can obtain :

$$\mathbb{E}_\epsilon[(\sum_{i=1}^{D} w_i\epsilon_i)^2] = \mathbb{E}_\epsilon[\sum_{i=1}^{D}\sum_{j=1}^{D} w_i w_j \epsilon_i \epsilon_j] = \sum_{i=1}^{D}\sum_{j=1}^{D} w_i w_j \mathbb{E}_\epsilon[\epsilon_i\epsilon_j] = \sigma^2 \sum_{i=1}^{D}\sum_{j=1}^{D} w_i w_j \delta_{ij}$$

Which gives

$$\mathbb{E}_\epsilon[(\sum_{i=1}^{D} w_i\epsilon_i)^2] = \sigma^2 \sum_{i=1}^{D} w_i^2$$

For the third term, we can obtain:

$$\begin{aligned}\mathbb{E}_\epsilon[2\big(\sum_{i=1}^{D} w_i\epsilon_i\big)\big(y(x_n,\boldsymbol{w})-t_n\big)] &= 2\big(y(x_n,\boldsymbol{w})-t_n\big)\,\mathbb{E}_\epsilon[\sum_{i=1}^{D} w_i\epsilon_i] \\ &= 2\big(y(x_n,\boldsymbol{w})-t_n\big)\sum_{i=1}^{D}\mathbb{E}_\epsilon[w_i\epsilon_i] \\ &= 0\end{aligned}$$

Therefore, if we calculate the expectation of $E_D(\boldsymbol{w})$ with respect to ϵ, we can obtain:

$$\mathbb{E}_\epsilon[E_D(\boldsymbol{w})] = \frac{1}{2}\sum_{n=1}^{N}\big(y(x_n,\boldsymbol{w})-t_n\big)^2 + \frac{\sigma^2}{2}\sum_{i=1}^{D} w_i^2$$

Problem 3.5 Solution

We can firstly rewrite the constraint (3.30) as :

$$\frac{1}{2}\left(\sum_{j=1}^{M}|w_j|^q - \eta\right) \le 0$$

Where we deliberately introduce scaling factor 1/2 for convenience.Then it is straightforward to obtain the Lagrange function.

$$L(\boldsymbol{w},\lambda) = \frac{1}{2}\sum_{n=1}^{N}\Big\{t_n - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x_n})\Big\}^2 + \frac{\lambda}{2}\left(\sum_{j=1}^{M}|w_j|^q - \eta\right)$$

It is obvious that $L(\boldsymbol{w},\lambda)$ and (3.29) has the same dependence on $\boldsymbol{w}$. Meanwhile, if we denote the optimal $\boldsymbol{w}$ that can minimize $L(\boldsymbol{w},\lambda)$ as $\boldsymbol{w}^\star(\lambda)$, we can see that

$$\eta = \sum_{j=1}^{M}|w_j^\star|^q$$

Problem 3.6 Solution

Firstly, we write down the log likelihood function.

$$lnp(\boldsymbol{T}|\boldsymbol{X},\boldsymbol{W},\beta) = -\frac{N}{2}ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}\big[\boldsymbol{t_n} - \boldsymbol{W}^T\boldsymbol{\phi}(\boldsymbol{x_n})\big]^T\boldsymbol{\Sigma}^{-1}\big[\boldsymbol{t_n} - \boldsymbol{W}^T\boldsymbol{\phi}(\boldsymbol{x_n})\big]$$

Where we have already omitted the constant term. We set the derivative of the equation above with respect to $\boldsymbol{W}$ equals to zero.

$$\boldsymbol{0} = -\sum_{n=1}^{N}\boldsymbol{\Sigma}^{-1}[\boldsymbol{t_n} - \boldsymbol{W}^T\boldsymbol{\phi}(\boldsymbol{x_n})]\boldsymbol{\phi}(\boldsymbol{x_n})^T$$

Therefore, we can obtain similar result for $\boldsymbol{W}$ as (3.15). For $\boldsymbol{\Sigma}$, comparing with (2.118) – (2.124), we can easily write down a similar result :

$$\boldsymbol{\Sigma} = \frac{1}{N}\sum_{n=1}^{N}[\boldsymbol{t_n} - \boldsymbol{W}_{ML}^T\boldsymbol{\phi}(\boldsymbol{x_n})]\,[\boldsymbol{t_n} - \boldsymbol{W}_{ML}^T\boldsymbol{\phi}(\boldsymbol{x_n})]^T$$

We can see that the solutions for $\boldsymbol{W}$ and $\boldsymbol{\Sigma}$ are also decoupled.

Problem 3.7 Solution

Let's begin by writing down the prior distribution $p(\boldsymbol{w})$ and likelihood function $p(\boldsymbol{t}|\boldsymbol{X},\boldsymbol{w},\beta)$.

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_0,\boldsymbol{S}_0)\ , \quad p(\boldsymbol{t}|\boldsymbol{X},\boldsymbol{w},\beta) = \prod_{n=1}^{N}\mathcal{N}(t_n|\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x_n}),\beta^{-1})$$

Since the posterior PDF equals to the product of the prior PDF and likelihood function, up to a normalized constant. We mainly focus on the exponential term of the product.

$$\begin{aligned}
\text{exponential term} &= -\frac{\beta}{2}\sum_{n=1}^{N}\left\{t_n - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x_n})\right\}^2 - \frac{1}{2}(\boldsymbol{w}-\boldsymbol{m}_0)^T\boldsymbol{S}_0^{-1}(\boldsymbol{w}-\boldsymbol{m}_0) \\
&= -\frac{\beta}{2}\sum_{n=1}^{N}\left\{t_n^2 - 2t_n\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x_n}) + \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x_n})\boldsymbol{\phi}(\boldsymbol{x_n})^T\boldsymbol{w}\right\} - \frac{1}{2}(\boldsymbol{w}-\boldsymbol{m}_0)^T\boldsymbol{S}_0^{-1}(\boldsymbol{w}-\boldsymbol{m}_0) \\
&= -\frac{1}{2}\boldsymbol{w}^T\left[\sum_{n=1}^{N}\beta\boldsymbol{\phi}(\boldsymbol{x_n})\boldsymbol{\phi}(\boldsymbol{x_n})^T + \boldsymbol{S}_0^{-1}\right]\boldsymbol{w} \\
&\quad -\frac{1}{2}\left[-2\boldsymbol{m}_0^T\boldsymbol{S}_0^{-1} - \sum_{n=1}^{N}2\beta t_n\boldsymbol{\phi}(\boldsymbol{x_n})^T\right]\boldsymbol{w} \\
&\quad + \text{const}
\end{aligned}$$

Hence, by comparing the quadratic term with standard Gaussian Distribution, we can obtain: $\boldsymbol{S}_N^{-1} = \boldsymbol{S}_0^{-1} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}$. And then comparing the linear term, we can obtain :

$$-2\boldsymbol{m_N}^T\boldsymbol{S_N}^{-1} = -2\boldsymbol{m}_0^T\boldsymbol{S}_0^{-1} - \sum_{n=1}^{N}2\beta t_n\boldsymbol{\phi}(\boldsymbol{x_n})^T$$

If we multiply -0.5 on both sides, and then transpose both sides, we can easily see that $\boldsymbol{m_N} = \boldsymbol{S_N}(\boldsymbol{S_0}^{-1}\boldsymbol{m_0} + \beta\boldsymbol{\Phi}^T\boldsymbol{t})$

Problem 3.8 Solution

Firstly, we write down the prior :

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{m_N},\boldsymbol{S_N})$$

Where $\boldsymbol{m_N}, \boldsymbol{S_N}$ are given by (3.50) and (3.51). And if now we observe another sample $(\boldsymbol{X_{N+1}}, t_{N+1})$, we can write down the likelihood function :

$$p(t_{N+1}|\boldsymbol{x_{N+1}}, \boldsymbol{w}) = \mathcal{N}(t|y(\boldsymbol{x_{N+1}}, \boldsymbol{w}), \beta^{-1})$$

Since the posterior equals to the production of likelihood function and the prior, up to a constant, we focus on the exponential term.

$$\begin{aligned} \text{exponential term} &= (\boldsymbol{w} - \boldsymbol{m_N})^T \boldsymbol{S_N^{-1}} (\boldsymbol{w} - \boldsymbol{m_N}) + \beta (t_{N+1} - \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x_{N+1}})^2 \\ &= \boldsymbol{w}^T \left[\boldsymbol{S_N}^{-1} + \beta \boldsymbol{\phi}(\boldsymbol{x_{N+1}}) \boldsymbol{\phi}(\boldsymbol{x_{N+1}})^T \right] \boldsymbol{w} \\ &\quad -2\boldsymbol{w}^T \left[\boldsymbol{S_N^{-1}} \boldsymbol{m_N} + \beta \boldsymbol{\phi}(\boldsymbol{x_{N+1}}) \boldsymbol{t_{N+1}} \right] \\ &\quad +\text{const} \end{aligned}$$

Therefore, after observing $(\boldsymbol{X_{N+1}}, t_{N+1})$, we have $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{m_{N+1}}, \boldsymbol{S_{N+1}})$, where we have defined:

$$\boldsymbol{S_{N+1}^{-1}} = \boldsymbol{S_N^{-1}} + \beta \, \boldsymbol{\phi}(\boldsymbol{x_{N+1}}) \, \boldsymbol{\phi}(\boldsymbol{x_{N+1}})^T$$

and

$$\boldsymbol{m_{N+1}} = \boldsymbol{S_{N+1}} \left(\boldsymbol{S_N^{-1}} \boldsymbol{m_N} + \beta \boldsymbol{\phi}(\boldsymbol{x_{N+1}}) \boldsymbol{t_{N+1}} \right)$$