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 :

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

$$\lambda_2 = 0$$

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

Note that there is a typo in the official solution manual about λ_3. Moreover, in the following parts, we will substitute $p(x)$ back into the three constraints and analytically prove that $p(x)$ is Gaussian. You can skip the following part. (The writer would especially thank Dr.Spyridon Chavlis from IMBB,FORTH for this analysis)

We already know:

$$p(x) = exp(-1 + \lambda_1 + \lambda_2 x + \lambda_3 (x-\mu)^2)$$

Where the exponent is equal to:

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

Completing the square, we can obtain that:

$$ax^2 + bx + c = a(x-d)^2 + f, d = -\frac{b}{2a}, f = c - \frac{b^2}{4a}$$

Using this quadratic form, the constraints can be written as

1. $\int_{-\infty}^{\infty} p(x)dx = \int_{-\infty}^{\infty} e^{[a(x-d)^2+f]}dx = 1$
2. $\int_{-\infty}^{\infty} xp(x)dx = \int_{-\infty}^{\infty} xe^{[a(x-d)^2+f]}dx = \mu$
3. $\int_{-\infty}^{\infty} (x-\mu)^2 p(x)dx = \int_{-\infty}^{\infty} (x-\mu)^2 e^{[a(x-d)^2+f]}dx = \sigma^2$

The first constraint can be written as:

$$I_1 = \int_{-\infty}^{\infty} e^{[a(x-d)^2+f]}dx = e^f \int_{-\infty}^{\infty} e^{a(x-d)^2}dx$$

Let $u = x - d$, which gives $du = dx$, and thus:

$$I_1 = e^f \int_{-\infty}^{\infty} e^{au^2}du$$

Let $-w^2 = au^2 \Rightarrow w = \sqrt{-a}u \Rightarrow dw = \sqrt{-a}du$, and thus:

$$I_1 = \frac{e^f}{\sqrt{-a}} \int_{-\infty}^{\infty} e^{-w^2}dw$$

As e^{-x^2} is an even function the integral is written as:

$$I_1 = \frac{2e^f}{\sqrt{-a}} \int_{0}^{\infty} e^{-w^2}dw$$

Let $w^2 = t \Rightarrow w = \sqrt{t} \Rightarrow dw = \frac{1}{2\sqrt{t}}dt$, and thus:

$$I_1 = \frac{2e^f}{\sqrt{-a}} \int_{0}^{\infty} t^{-\frac{1}{2}} e^{-t}dt = \frac{2e^f}{\sqrt{-a}} \int_{0}^{\infty} \frac{1}{2} t^{\frac{1}{2}-1} e^{-t}dt = \frac{e^f}{\sqrt{-a}} \Gamma(\frac{1}{2}) = e^f \sqrt{\frac{\pi}{-a}}$$

Here the Gamma function is used. Gamma function is defined as

$$\Gamma(z) = \int_{0}^{\infty} t^{z-1} e^{-t}dt$$

where for non-negative integer values of n, we have:

$$\Gamma(\frac{1}{2}+n)=\frac{(2n)!}{4^n n!}\sqrt{\pi}$$

Thus, the first constraint can be rewritten as:

$$e^f\sqrt{\frac{\pi}{-a}}=1 \qquad (*)$$

The second constraint can be written as:

$$I_2=\int_{-\infty}^{\infty}xe^{[a(x-d)^2+f]}dx=e^f\int_{-\infty}^{\infty}xe^{a(x-d)^2}dx$$

Let $u=x-d\Rightarrow x=u+d\Rightarrow du=dx$, and thus:

$$I_2=e^f\int_{-\infty}^{\infty}(u+d)e^{au^2}du$$

Using integral additivity, we have:

$$I_2=e^f\int_{-\infty}^{\infty}ue^{au^2}du+e^f\int_{-\infty}^{\infty}de^{au^2}du$$

We first deal with the first term on the right hand side. Here we denote it as I_{21}:

$$I_{21}=e^f\int_{-\infty}^{\infty}ue^{au^2}du=e^f\left(\int_{-\infty}^{0}ue^{au^2}du+\int_{0}^{\infty}ue^{au^2}du\right)$$

Swapping the integration limits, we obtain:

$$\begin{aligned}I_{21} &= e^f\left(-\int_{0}^{-\infty}ue^{au^2}du+\int_{0}^{\infty}ue^{au^2}du\right)\\ &= e^f\left(\int_{0}^{-\infty}(-u)e^{a(-u)^2}du+\int_{0}^{\infty}ue^{au^2}du\right)\\ &= e^f\left(-\int_{0}^{\infty}(-u)e^{a(-u)^2}(-du)+\int_{0}^{\infty}ue^{au^2}du\right)=0\end{aligned}$$

Then we deal with the second term I_{22}:

$$I_{22}=e^f\int_{-\infty}^{\infty}de^{au^2}du$$

Let $-w^2=au^2\Rightarrow w=\sqrt{-a}u\Rightarrow dw=\sqrt{-a}du$, and thus:

$$I_{22}=\frac{e^f d}{\sqrt{-a}}\int_{-\infty}^{\infty}e^{-w^2}dw$$

As e^{-x^2} is an even function the integral is written as:

$$I_{22}=\frac{2e^f d}{\sqrt{-a}}\int_{0}^{\infty}e^{-w^2}dw$$

Let $w^2 = t \Rightarrow w = \sqrt{t} \Rightarrow dw = \frac{1}{2\sqrt{t}} dt$, and thus:

$$I_{22} = \frac{2e^f d}{\sqrt{-a}} \int_0^\infty t^{-\frac{1}{2}} e^{-t} dt = \frac{2e^f d}{\sqrt{-a}} \int_0^\infty \frac{1}{2} t^{\frac{1}{2}-1} e^{-t} dt = \frac{e^f d}{\sqrt{-a}} \Gamma(\frac{1}{2}) = e^f d \sqrt{\frac{\pi}{-a}}$$

Thus, the second constraint can be rewritten

$$e^f d \sqrt{\frac{\pi}{-a}} = \mu \qquad (**)$$

Combining $(*)$ and $(**)$, we can obtain that $d = \mu$. Recall that:

$$d = -\frac{b}{2a} = -\frac{\lambda_2 - 2\lambda_3 \mu}{2\lambda_3} = \mu \Rightarrow \lambda_2 - 2\lambda_3 \mu = -2\lambda_3 \mu \Rightarrow \lambda_2 = 0$$

So far, we have:

$$b = -2\lambda_3 \mu$$

And

$$f = c - \frac{b^2}{4a} = \lambda_3 \mu^2 + \lambda_1 - 1 - \frac{4\lambda_3^2 \mu^2}{4\lambda_3} = \lambda_1 - 1$$

Finally, we deal with the third also the last constraint. Substituting $\lambda_2 = 0$ into the last constraint we have:

$$I_3 = \int_{-\infty}^{\infty} (x-\mu)^2 e^{[\lambda_3 (x-\mu)^2 + \lambda_1 - 1]} dx = e^{\lambda_1 - 1} \int_{-\infty}^{\infty} (x-\mu)^2 e^{\lambda_3 (x-\mu)^2} dx$$

Let $u = x - \mu \Rightarrow du = dx$, and thus:

$$I_3 = e^{\lambda_1 - 1} \int_{-\infty}^{\infty} u^2 e^{\lambda_3 u^2} dx$$

Let $-w^2 = \lambda_3 u^2 \Rightarrow w = \sqrt{-\lambda_3} u \Rightarrow dw = \sqrt{-\lambda_3} du$, and thus:

$$I_3 = e^{\lambda_1 - 1} \int_{-\infty}^{\infty} -\frac{1}{\lambda_3} w^2 e^{-w^2} \frac{dw}{\sqrt{-\lambda_3}} = \frac{e^{\lambda_1 - 1}}{-\lambda_3^{\frac{3}{2}}} \int_{-\infty}^{\infty} w^2 e^{-w^2} dw$$

Because it is an even function, we can further obtain:

$$I_3 = 2 \frac{e^{\lambda_1 - 1}}{-\lambda_3^{\frac{3}{2}}} \int_0^{\infty} w^2 e^{-w^2} dw$$

Let $w^2 = t \Rightarrow w = \sqrt{t} \Rightarrow dw = \frac{1}{2\sqrt{t}} dt$, and thus:

$$\begin{aligned} I_3 &= 2 \frac{e^{\lambda_1 - 1}}{-\lambda_3^{\frac{3}{2}}} \int_0^{\infty} t e^{-t} \frac{1}{2\sqrt{(t)}} dt = \frac{e^{\lambda_1 - 1}}{-\lambda_3^{\frac{3}{2}}} \int_0^{\infty} t^{1-\frac{1}{2}} e^{-t} dt \\ &= \frac{e^{\lambda_1 - 1}}{-\lambda_3^{\frac{3}{2}}} \int_0^{\infty} t^{\frac{3}{2}-1} e^{-t} dt \\ &= \frac{e^{\lambda_1 - 1}}{-\lambda_3^{\frac{3}{2}}} \Gamma(\frac{3}{2}) = \frac{e^{\lambda_1 - 1}}{-\lambda_3^{\frac{3}{2}}} \frac{\pi}{2} \end{aligned}$$

Thus, the third constraint can be rewritten

$$\frac{e^{\lambda_1-1}}{-\lambda_3^{\frac{3}{2}}}\frac{\sqrt{\pi}}{2}=\sigma^2 \qquad (***)$$

Rewriting $(*)$ with $f=\lambda_1-1, d=\mu$ and $a=\lambda_3$, we obtain the following equation

$$e^{\lambda_1-1}\sqrt{\frac{\pi}{-\lambda_3}}=1 \qquad (****)$$

Substituting the equation above back into $(***)$, we obtain

$$\sqrt{\frac{-\lambda_3}{\pi}}\frac{1}{-\lambda_3^{\frac{3}{2}}}\frac{\sqrt{\pi}}{2}=\sigma^2 \Leftrightarrow -\frac{1}{\lambda_3}=2\sigma^2 \Leftrightarrow \lambda_3=-\frac{1}{2\sigma^2}$$

Substituting λ_3 back into $(****)$, we obtain:

$$e^{\lambda_1-1}\sqrt{\frac{\pi}{-\lambda_3}}=1 \Leftrightarrow e^{\lambda_1-1}\sqrt{\frac{\pi}{\frac{1}{2\sigma^2}}}=1 \Leftrightarrow e^{\lambda_1-1}=\frac{1}{\sqrt{2\pi\sigma^2}} \Leftrightarrow \lambda_1-1=\ln(\frac{1}{\sqrt{2\pi\sigma^2}})$$

Thus, we obtain:

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

So far, we have obtainde λ_i, where $i=1,2,3$. We substitute them back into $p(x)$, yielding:

$$\begin{aligned} p(x) &= \exp\left(-1+1-\frac{1}{2}\ln(2\pi\sigma^2)-\frac{1}{2\sigma^2}(x-\mu)^2\right) \\ &= \exp\left(-\frac{1}{2}\ln(2\pi\sigma^2)\right)\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \\ &= \exp\left(\ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)\right)\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \end{aligned}$$

Thus,

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

Just as required.

Problem 1.35 Solution

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

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

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) :

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

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 :

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

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 $\boldsymbol{y} = \boldsymbol{\Sigma}^{\mathbf{1/2}}\boldsymbol{x}$. 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 $\boldsymbol{\Sigma} = \boldsymbol{S}$, 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 :

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

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 :

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

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

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

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

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

Problem 3.9 Solution

We know that the prior $p(\boldsymbol{w})$ can be written as:

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

And the likelihood function $p(t_{N+1}|\boldsymbol{x_{N+1}}, \boldsymbol{w})$ can be written as:

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

According to the fact that $y(\boldsymbol{x_{N+1}}, \boldsymbol{w}) = \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x_{N+1}}) = \boldsymbol{\phi}(\boldsymbol{x_{N+1}})^T \boldsymbol{w}$, the likelihood can be further written as:

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

Then we take advantage of (2.113), (2.114) and (2.116), which gives:

$$p(\boldsymbol{w}|\boldsymbol{x_{N+1}}, t_{N+1}) = \mathcal{N}(\boldsymbol{\Sigma}\{\boldsymbol{\phi}(\boldsymbol{x_{N+1}})\beta t_{N+1} + \boldsymbol{S_N}^{-1} \boldsymbol{m_N}\}, \boldsymbol{\Sigma})$$

Where $\boldsymbol{\Sigma} = (\boldsymbol{S_N}^{-1} + \boldsymbol{\phi}(\boldsymbol{x_{N+1}})\beta\boldsymbol{\phi}(\boldsymbol{x_{N+1}})^T)^{-1}$, and we can see that the result is exactly the same as the one we obtained in the previous problem.

Problem 3.10 Solution

We have already known:

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

And

$$p(\boldsymbol{w}|\mathbf{t},\alpha,\beta) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m_N},\boldsymbol{S_N})$$

Where $\boldsymbol{m_N}$, $\boldsymbol{S_N}$ are given by (3.53) and (3.54). As what we do in previous problem, we can rewrite $p(t|\boldsymbol{w},\beta)$ as:

$$p(t|\boldsymbol{w},\beta) = \mathcal{N}(t|\boldsymbol{\phi}(\boldsymbol{x})^T\boldsymbol{w},\beta^{-1})$$

And then we take advantage of (2.113), (2.114) and (2.115), we can obtain:

$$p(t|\mathbf{t},\alpha,\beta) = \mathcal{N}(\boldsymbol{\phi}(\boldsymbol{x})^T\boldsymbol{m_N},\beta^{-1}+\boldsymbol{\phi}(\boldsymbol{x})^T\boldsymbol{S_N}\boldsymbol{\phi}(\boldsymbol{x}))$$

Which is exactly the same as (3.58), if we notice that

$$\boldsymbol{\phi}(\boldsymbol{x})^T\boldsymbol{m_N} = \boldsymbol{m_N}^T\boldsymbol{\phi}(\boldsymbol{x})$$

Problem 3.11 Solution

We need to use the result obtained in Prob.3.8. In Prob.3.8, we have derived a formula for $\boldsymbol{S_{N+1}^{-1}}$:

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

And then using (3.110), we can obtain :

$$\begin{aligned}
\boldsymbol{S_{N+1}} &= \left[\boldsymbol{S_N}^{-1} + \beta\,\boldsymbol{\phi}(\boldsymbol{x_{N+1}})\,\boldsymbol{\phi}(\boldsymbol{x_{N+1}})^T\right]^{-1} \\
&= \left[\boldsymbol{S_N}^{-1} + \sqrt{\beta}\boldsymbol{\phi}(\boldsymbol{x_{N+1}})\,\sqrt{\beta}\boldsymbol{\phi}(\boldsymbol{x_{N+1}})^T\right]^{-1} \\
&= \boldsymbol{S_N} - \frac{\boldsymbol{S_N}(\sqrt{\beta}\boldsymbol{\phi}(\boldsymbol{x_{N+1}}))(\sqrt{\beta}\boldsymbol{\phi}(\boldsymbol{x_{N+1}})^T)\boldsymbol{S_N}}{1+(\sqrt{\beta}\boldsymbol{\phi}(\boldsymbol{x_{N+1}})^T)\boldsymbol{S_N}(\sqrt{\beta}\boldsymbol{\phi}(\boldsymbol{x_{N+1}}))} \\
&= \boldsymbol{S_N} - \frac{\beta\boldsymbol{S_N}\boldsymbol{\phi}(\boldsymbol{x_{N+1}})\boldsymbol{\phi}(\boldsymbol{x_{N+1}})^T\boldsymbol{S_N}}{1+\beta\boldsymbol{\phi}(\boldsymbol{x_{N+1}})^T\boldsymbol{S_N}\boldsymbol{\phi}(\boldsymbol{x_{N+1}})}
\end{aligned}$$

Now we calculate $\sigma_N^2(\boldsymbol{x}) - \sigma_{N+1}^2(\boldsymbol{x})$ according to (3.59).

$$\begin{aligned}
\sigma_N^2(\boldsymbol{x}) - \sigma_{N+1}^2(\boldsymbol{x}) &= \boldsymbol{\phi}(\boldsymbol{x})^T(\boldsymbol{S_N}-\boldsymbol{S_{N+1}})\boldsymbol{\phi}(\boldsymbol{x}) \\
&= \boldsymbol{\phi}(\boldsymbol{x})^T\frac{\beta\boldsymbol{S_N}\boldsymbol{\phi}(\boldsymbol{x_{N+1}})\boldsymbol{\phi}(\boldsymbol{x_{N+1}})^T\boldsymbol{S_N}}{1+\beta\boldsymbol{\phi}(\boldsymbol{x_{N+1}})^T\boldsymbol{S_N}\boldsymbol{\phi}(\boldsymbol{x_{N+1}})}\boldsymbol{\phi}(\boldsymbol{x}) \\
&= \frac{\boldsymbol{\phi}(\boldsymbol{x})^T\boldsymbol{S_N}\boldsymbol{\phi}(\boldsymbol{x_{N+1}})\boldsymbol{\phi}(\boldsymbol{x_{N+1}})^T\boldsymbol{S_N}\boldsymbol{\phi}(\boldsymbol{x})}{1/\beta+\boldsymbol{\phi}(\boldsymbol{x_{N+1}})^T\boldsymbol{S_N}\boldsymbol{\phi}(\boldsymbol{x_{N+1}})} \\
&= \frac{\left[\boldsymbol{\phi}(\boldsymbol{x})^T\boldsymbol{S_N}\boldsymbol{\phi}(\boldsymbol{x_{N+1}})\right]^2}{1/\beta+\boldsymbol{\phi}(\boldsymbol{x_{N+1}})^T\boldsymbol{S_N}\boldsymbol{\phi}(\boldsymbol{x_{N+1}})} \quad (*)
\end{aligned}$$

And since $\boldsymbol{S_N}$ is positive definite, $(*)$ is larger than 0. Therefore, we have proved that $\sigma_N^2(\boldsymbol{x}) - \sigma_{N+1}^2(\boldsymbol{x}) \geq 0$

Problem 3.12 Solution

Let's begin by writing down the prior PDF $p(\boldsymbol{w},\beta)$:

$$\begin{aligned} p(\boldsymbol{w},\beta) &= \mathcal{N}(\boldsymbol{w}|\boldsymbol{m_0},\beta^{-1}\boldsymbol{S_0})\,\mathrm{Gam}(\beta|a_0,b_0) \quad (*) \\ &\propto (\frac{\beta}{|\boldsymbol{S_0}|})^2 exp(-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{m_0})^T\beta\boldsymbol{S_0^{-1}}(\boldsymbol{w}-\boldsymbol{m_0}))\, b_0^{a_0}\beta^{a_0-1}exp(-b_0\beta) \end{aligned}$$

And then we write down the likelihood function $p(\mathbf{t}|\mathbf{X},\boldsymbol{w},\beta)$:

$$\begin{aligned} p(\mathbf{t}|\mathbf{X},\boldsymbol{w},\beta) &= \prod_{n=1}^{N}\mathcal{N}(t_n|\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x_n}),\beta^{-1}) \\ &\propto \prod_{n=1}^{N}\beta^{1/2}exp\big[-\frac{\beta}{2}(t_n-\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x_n}))^2\big] \quad (**) \end{aligned}$$

According to Bayesian Inference, we have $p(\boldsymbol{w},\beta|\mathbf{t})\propto p(\mathbf{t}|\mathbf{X},\boldsymbol{w},\beta)\times p(\boldsymbol{w},\beta)$. We first focus on the quadratic term with regard to $\boldsymbol{w}$ in the exponent.

$$\begin{aligned} \text{quadratic term} &= -\frac{\beta}{2}\boldsymbol{w}^T{\boldsymbol{S_0}}^{-1}\boldsymbol{w}+\sum_{n=1}^{N}-\frac{\beta}{2}\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x_n})\boldsymbol{\phi}(\boldsymbol{x_n})^T\boldsymbol{w} \\ &= -\frac{\beta}{2}\boldsymbol{w}^T\big[{\boldsymbol{S_0}}^{-1}+\sum_{n=1}^{N}\boldsymbol{\phi}(\boldsymbol{x_n})\boldsymbol{\phi}(\boldsymbol{x_n})^T\big]\boldsymbol{w} \end{aligned}$$

Where the first term is generated by $(*)$, and the second by $(**)$. By now, we know that:

$${\boldsymbol{S_N}}^{-1} = {\boldsymbol{S_0}}^{-1}+\sum_{n=1}^{N}\boldsymbol{\phi}(\boldsymbol{x_n})\boldsymbol{\phi}(\boldsymbol{x_n})^T$$

We then focus on the linear term with regard to $\boldsymbol{w}$ in the exponent.

$$\begin{aligned} \text{linear term} &= \beta{\boldsymbol{m_0}}^T{\boldsymbol{S_0}}^{-1}\boldsymbol{w}+\sum_{n=1}^{N}\beta t_n\boldsymbol{\phi}(\boldsymbol{x_n})^T\boldsymbol{w} \\ &= \beta\big[{\boldsymbol{m_0}}^T{\boldsymbol{S_0}}^{-1}+\sum_{n=1}^{N}t_n\boldsymbol{\phi}(\boldsymbol{x_n})^T\big]\boldsymbol{w} \end{aligned}$$

Again, the first term is generated by $(*)$, and the second by $(**)$. We can also obtain:

$${\boldsymbol{m_N}}^T{\boldsymbol{S_N}}^{-1} = {\boldsymbol{m_0}}^T{\boldsymbol{S_0}}^{-1}+\sum_{n=1}^{N}t_n\boldsymbol{\phi}(\boldsymbol{x_n})^T$$

Which gives:

$$\boldsymbol{m_N} = \boldsymbol{S_N}\big[{\boldsymbol{S_0}}^{-1}\boldsymbol{m_0}+\sum_{n=1}^{N}t_n\boldsymbol{\phi}(\boldsymbol{x_n})\big]$$

Then we focus on the constant term with regard to $\boldsymbol{w}$ in the exponent.

$$\begin{aligned} \text{constant term} &= (-\frac{\beta}{2}{\boldsymbol{m_0}}^T{\boldsymbol{S_0}}^{-1}\boldsymbol{m_0}-b_0\beta)-\frac{\beta}{2}\sum_{n=1}^{N}t_n^2 \\ &= -\beta\big[\frac{1}{2}{\boldsymbol{m_0}}^T{\boldsymbol{S_0}}^{-1}\boldsymbol{m_0}+b_0+\frac{1}{2}\sum_{n=1}^{N}t_n^2\big] \end{aligned}$$

Therefore, we can obtain:

$$\frac{1}{2}\boldsymbol{m_N}^T\boldsymbol{S_N}^{-1}\boldsymbol{m_N} + b_N = \frac{1}{2}\boldsymbol{m_0}^T\boldsymbol{S_0}^{-1}\boldsymbol{m_0} + b_0 + \frac{1}{2}\sum_{n=1}^{N} t_n^2$$

Which gives :

$$b_N = \frac{1}{2}\boldsymbol{m_0}^T\boldsymbol{S_0}^{-1}\boldsymbol{m_0} + b_0 + \frac{1}{2}\sum_{n=1}^{N} t_n^2 - \frac{1}{2}\boldsymbol{m_N}^T\boldsymbol{S_N}^{-1}\boldsymbol{m_N}$$

Finally, we focus on the exponential term whose base is β.

$$\text{exponent term} = (2 + a_0 - 1) + \frac{N}{2}$$

Which gives:

$$2 + a_N - 1 = (2 + a_0 - 1) + \frac{N}{2}$$

Hence,

$$a_N = a_0 + \frac{N}{2}$$

Problem 3.13 Solution(Waiting for update)

Similar to (3.57), we write down the expression of the predictive distribution $p(t|\mathbf{X},\mathbf{t})$:

$$p(t|\mathbf{X},\mathbf{t}) = \int\int p(t|\boldsymbol{w},\beta)\, p(\boldsymbol{w},\beta|\mathbf{X},\mathbf{t})\, d\boldsymbol{w}\, d\beta \qquad (*)$$

We know that:

$$p(t|\boldsymbol{w},\beta) = \mathcal{N}(t|y(\boldsymbol{x},\boldsymbol{w}),\beta^{-1}) = \mathcal{N}(t|\boldsymbol{\phi}(\boldsymbol{x})^T\boldsymbol{w},\beta^{-1})$$

And that:

$$p(\boldsymbol{w},\beta|\mathbf{X},\mathbf{t}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m_N},\beta^{-1}\boldsymbol{S_N})\,\mathrm{Gam}(\beta|a_N,b_N)$$

We go back to $(*)$, and we first deal with the integral with regard to $\boldsymbol{w}$:

$$\begin{aligned} p(t|\mathbf{X},\mathbf{t}) &= \int\Big[\int \mathcal{N}(t|\boldsymbol{\phi}(\boldsymbol{x})^T\boldsymbol{w},\beta^{-1})\,\mathcal{N}(\boldsymbol{w}|\boldsymbol{m_N},\beta^{-1}\boldsymbol{S_N})\, d\boldsymbol{w}\Big]\mathrm{Gam}(\beta|a_N,b_N)\, d\beta \\ &= \int \mathcal{N}(t|\boldsymbol{\phi}(\boldsymbol{x})^T\boldsymbol{m_N},\beta^{-1} + \boldsymbol{\phi}(\boldsymbol{x})^T\beta^{-1}\boldsymbol{S_N}\boldsymbol{\phi}(\boldsymbol{x}))\,\mathrm{Gam}(\beta|a_N,b_N)\, d\beta \\ &= \int \mathcal{N}\big[t|\boldsymbol{\phi}(\boldsymbol{x})^T\boldsymbol{m_N},\beta^{-1}(1 + \boldsymbol{\phi}(\boldsymbol{x})^T\boldsymbol{S_N}\boldsymbol{\phi}(\boldsymbol{x}))\big]\,\mathrm{Gam}(\beta|a_N,b_N)\, d\beta \end{aligned}$$

Where we have used (2.113), (2.114) and (2.115). Then, we compare the expression above with (2.160), we can see that $p(t|\mathbf{X},\mathbf{t}) = \mathrm{St}(t|\mu,\lambda,\upsilon)$, where we have defined:

$$\mu = \boldsymbol{\phi}(\boldsymbol{x})^T\boldsymbol{m_N}, \quad \lambda = \big[1 + \boldsymbol{\phi}(\boldsymbol{x})^T\boldsymbol{S_N}\boldsymbol{\phi}(\boldsymbol{x})\big]^{-1}, \quad \upsilon = 2a_N$$

Problem 3.14 Solution(Wait for updating)

Firstly, according to (3.16), if we use the new orthonormal basis set specified in the problem to construct $\mathbf{\Phi}$, we can obtain an important property: $\mathbf{\Phi}^T\mathbf{\Phi} = \mathbf{I}$. Hence, if $\alpha = 0$, together with (3.54), we know that $\mathbf{S_N} = 1/\beta$. Finally, according to (3.62), we can obtain:

$$k(\boldsymbol{x},\boldsymbol{x}') = \beta\boldsymbol{\psi}(\boldsymbol{x})^T\mathbf{S_N}\boldsymbol{\psi}(\boldsymbol{x}') = \boldsymbol{\psi}(\boldsymbol{x})^T\boldsymbol{\psi}(\boldsymbol{x}')$$

Problem 3.15 Solution

It is quite obvious if we substitute (3.92) and (3.95) into (3.82), which gives,

$$E(\boldsymbol{m_N}) = \frac{\beta}{2}||\mathbf{t}-\Phi\boldsymbol{m_N}||^2 + \frac{\alpha}{2}\boldsymbol{m_N}^T\boldsymbol{m_N} = \frac{N-\gamma}{2} + \frac{\gamma}{2} = \frac{N}{2}$$

Problem 3.16 Solution(Waiting for update)

We know that

$$p(\mathbf{t}|\boldsymbol{w},\beta) = \prod_{n=1}^{N}\mathcal{N}(\boldsymbol{\phi}(\boldsymbol{x_n})^T\boldsymbol{w},\beta^{-1}) \propto \mathcal{N}(\mathbf{\Phi}\boldsymbol{w},\beta^{-1}\mathbf{I})$$

And

$$p(\boldsymbol{w}|\alpha) = \mathcal{N}(\mathbf{0},\alpha^{-1}\mathbf{I})$$

Comparing them with (2.113), (2.114) and (2.115), we can obtain:

$$p(\mathbf{t}|\alpha,\beta) = \mathcal{N}(\mathbf{0},\beta^{-1}\mathbf{I}+\alpha^{-1}\mathbf{\Phi}\mathbf{\Phi}^T)$$

Problem 3.17 Solution

We know that:

$$\begin{aligned}
p(\mathbf{t}|\boldsymbol{w},\beta) &= \prod_{n=1}^{N}\mathcal{N}(\boldsymbol{\phi}(\boldsymbol{x_n})^T\boldsymbol{w},\beta^{-1}) \\
&= \prod_{n=1}^{N}\frac{1}{(2\pi\beta^{-1})^{1/2}}exp\{-\frac{1}{2\beta^{-1}}(t_n-\boldsymbol{\phi}(\boldsymbol{x_n})^T\boldsymbol{w})^2\} \\
&= (\frac{\beta}{2\pi})^{N/2}exp\{\sum_{n=1}^{N}-\frac{\beta}{2}(t_n-\boldsymbol{\phi}(\boldsymbol{x_n})^T\boldsymbol{w})^2\} \\
&= (\frac{\beta}{2\pi})^{N/2}exp\{-\frac{\beta}{2}||\mathbf{t}-\Phi\boldsymbol{w}||^2\}
\end{aligned}$$

And that:

$$\begin{aligned}
p(\boldsymbol{w}|\alpha) &= \mathcal{N}(\mathbf{0},\alpha^{-1}\mathbf{I}) \\
&= \frac{\alpha^{M/2}}{(2\pi)^{M/2}}exp\{-\frac{\alpha}{2}||\boldsymbol{w}||^2\}
\end{aligned}$$

If we substitute the expressions above into (3.77), we can obtain (3.78) just as required.

Problem 3.18 Solution

We expand (3.79) as follows:

$$\begin{aligned} E(\boldsymbol{w}) &= \frac{\beta}{2}||\mathbf{t}-\boldsymbol{\Phi}\boldsymbol{w}||^2+\frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w} \\ &= \frac{\beta}{2}(\mathbf{t}^T\mathbf{t}-2\mathbf{t}^T\boldsymbol{\Phi}\boldsymbol{w}+\boldsymbol{w}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{w})+\frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w} \\ &= \frac{1}{2}\left[\boldsymbol{w}^T(\beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}+\alpha\mathbf{I})\boldsymbol{w}-2\beta\mathbf{t}^T\boldsymbol{\Phi}\boldsymbol{w}+\beta\mathbf{t}^T\mathbf{t}\right] \end{aligned}$$

Observing the equation above, we see that $E(\boldsymbol{w})$ contains the following term :

$$\frac{1}{2}(\boldsymbol{w}-\boldsymbol{m}_{\boldsymbol{N}})^T\mathbf{A}(\boldsymbol{w}-\boldsymbol{m}_{\boldsymbol{N}}) \qquad (*)$$

Now, we need to solve $\mathbf{A}$ and $\boldsymbol{m}_{\boldsymbol{N}}$. We expand $(*)$ and obtain:

$$(*)=\frac{1}{2}(\boldsymbol{w}^T\mathbf{A}\boldsymbol{w}-2\boldsymbol{m}_{\boldsymbol{N}}{}^T\mathbf{A}\boldsymbol{w}+\boldsymbol{m}_{\boldsymbol{N}}{}^T\mathbf{A}\boldsymbol{m}_{\boldsymbol{N}})$$

We firstly compare the quadratic term, which gives:

$$\mathbf{A}=\beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}+\alpha\mathbf{I}$$

And then we compare the linear term, which gives:

$$\boldsymbol{m}_{\boldsymbol{N}}{}^T\mathbf{A}=\beta\mathbf{t}^T\boldsymbol{\Phi}$$

Noticing that $\mathbf{A}=\mathbf{A}^T$, which implies $\mathbf{A}^{-1}$ is also symmetric, we first transpose and then multiply $\mathbf{A}^{-1}$ on both sides, which gives:

$$\boldsymbol{m}_{\boldsymbol{N}}=\beta\mathbf{A}^{-1}\boldsymbol{\Phi}^T\mathbf{t}$$

Now we rewrite $E(\boldsymbol{w})$:

$$\begin{aligned} E(\boldsymbol{w}) &= \frac{1}{2}\left[\boldsymbol{w}^T(\beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}+\alpha\mathbf{I})\boldsymbol{w}-2\beta\mathbf{t}^T\boldsymbol{\Phi}\boldsymbol{w}+\beta\mathbf{t}^T\mathbf{t}\right] \\ &= \frac{1}{2}\left[(\boldsymbol{w}-\boldsymbol{m}_{\boldsymbol{N}})^T\mathbf{A}(\boldsymbol{w}-\boldsymbol{m}_{\boldsymbol{N}})+\beta\mathbf{t}^T\mathbf{t}-\boldsymbol{m}_{\boldsymbol{N}}{}^T\mathbf{A}\boldsymbol{m}_{\boldsymbol{N}}\right] \\ &= \frac{1}{2}(\boldsymbol{w}-\boldsymbol{m}_{\boldsymbol{N}})^T\mathbf{A}(\boldsymbol{w}-\boldsymbol{m}_{\boldsymbol{N}})+\frac{1}{2}(\beta\mathbf{t}^T\mathbf{t}-\boldsymbol{m}_{\boldsymbol{N}}{}^T\mathbf{A}\boldsymbol{m}_{\boldsymbol{N}}) \\ &= \frac{1}{2}(\boldsymbol{w}-\boldsymbol{m}_{\boldsymbol{N}})^T\mathbf{A}(\boldsymbol{w}-\boldsymbol{m}_{\boldsymbol{N}})+\frac{1}{2}(\beta\mathbf{t}^T\mathbf{t}-2\boldsymbol{m}_{\boldsymbol{N}}{}^T\mathbf{A}\boldsymbol{m}_{\boldsymbol{N}}+\boldsymbol{m}_{\boldsymbol{N}}{}^T\mathbf{A}\boldsymbol{m}_{\boldsymbol{N}}) \\ &= \frac{1}{2}(\boldsymbol{w}-\boldsymbol{m}_{\boldsymbol{N}})^T\mathbf{A}(\boldsymbol{w}-\boldsymbol{m}_{\boldsymbol{N}})+\frac{1}{2}(\beta\mathbf{t}^T\mathbf{t}-2\boldsymbol{m}_{\boldsymbol{N}}{}^T\mathbf{A}\boldsymbol{m}_{\boldsymbol{N}}+\boldsymbol{m}_{\boldsymbol{N}}{}^T(\beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}+\alpha\mathbf{I})\boldsymbol{m}_{\boldsymbol{N}}) \\ &= \frac{1}{2}(\boldsymbol{w}-\boldsymbol{m}_{\boldsymbol{N}})^T\mathbf{A}(\boldsymbol{w}-\boldsymbol{m}_{\boldsymbol{N}})+\frac{1}{2}\left[\beta\mathbf{t}^T\mathbf{t}-2\beta\mathbf{t}^T\boldsymbol{\Phi}\boldsymbol{m}_{\boldsymbol{N}}+\boldsymbol{m}_{\boldsymbol{N}}{}^T(\beta\boldsymbol{\Phi}^T\boldsymbol{\Phi})\boldsymbol{m}_{\boldsymbol{N}}\right]+\frac{\alpha}{2}\boldsymbol{m}_{\boldsymbol{N}}{}^T\boldsymbol{m}_{\boldsymbol{N}} \\ &= \frac{1}{2}(\boldsymbol{w}-\boldsymbol{m}_{\boldsymbol{N}})^T\mathbf{A}(\boldsymbol{w}-\boldsymbol{m}_{\boldsymbol{N}})+\frac{\beta}{2}||\mathbf{t}-\boldsymbol{\Phi}\boldsymbol{m}_{\boldsymbol{N}}||^2+\frac{\alpha}{2}\boldsymbol{m}_{\boldsymbol{N}}{}^T\boldsymbol{m}_{\boldsymbol{N}} \end{aligned}$$

Just as required.

Problem 3.19 Solution

Based on the standard form of a multivariate normal distribution, we know that

$$\int \frac{1}{(2\pi)^{M/2}} \frac{1}{|\mathbf{A}|^{1/2}} exp\{-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{m_N})^T \mathbf{A}(\boldsymbol{w}-\boldsymbol{m_N})\}\, d\boldsymbol{w} = 1$$

Hence,

$$\int exp\{-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{m_N})^T \mathbf{A}(\boldsymbol{w}-\boldsymbol{m_N})\}\, d\boldsymbol{w} = (2\pi)^{M/2}|\mathbf{A}|^{1/2}$$

And since $E(\boldsymbol{m_N})$ doesn't depend on $\boldsymbol{w}$, (3.85) is quite obvious. Then we substitute (3.85) into (3.78), which will immediately gives (3.86).

Problem 3.20 Solution

You can just follow the steps from (3.87) to (3.92), which is already very clear.

Problem 3.21 Solution

Let's first prove (3.117). According to (C.47) and (C.48), we know that if $\mathbf{A}$ is a $M \times M$ real symmetric matrix, with eigenvalues λ_i , $i = 1,2,...,M$, $|\mathbf{A}|$ and $\mathrm{Tr}(\mathbf{A})$ can be written as:

$$|\mathbf{A}| = \prod_{i=1}^{M} \lambda_i \,, \quad \mathrm{Tr}(\mathbf{A}) = \sum_{i=1}^{M} \lambda_i$$

Back to this problem, according to section 3.5.2, we know that $\mathbf{A}$ has eigenvalues $\alpha + \lambda_i$, $i = 1,2,...,M$. Hence the left side of (3.117) equals to:

$$\text{left side} = \frac{d}{d\alpha} ln\big[\prod_{i=1}^{M}(\alpha+\lambda_i)\big] = \sum_{i=1}^{M} \frac{d}{d\alpha} ln(\alpha+\lambda_i) = \sum_{i=1}^{M} \frac{1}{\alpha+\lambda_i}$$

And according to (3.81), we can obtain:

$$\mathbf{A}^{-1}\frac{d}{d\alpha}\mathbf{A} = \mathbf{A}^{-1}\mathbf{I} = \mathbf{A}^{-1}$$

For the symmetric matrix $\mathbf{A}$, its inverse $\mathbf{A}^{-1}$ has eigenvalues $1/(\alpha+\lambda_i)$, $i = 1,2,...,M$. Therefore,

$$\mathrm{Tr}(\mathbf{A}^{-1}\frac{d}{d\alpha}\mathbf{A}) = \sum_{i=1}^{M} \frac{1}{\alpha+\lambda_i}$$

Hence there are the same, and (3.92) is quite obvious.

Problem 3.22 Solution

Let's derive (3.86) with regard to β. The first term dependent on β in (3.86) is :

$$\frac{d}{d\beta}(\frac{N}{2}ln\beta) = \frac{N}{2\beta}$$

The second term is :

$$\frac{d}{d\beta}E(\boldsymbol{m_N}) = \frac{1}{2}||\mathbf{t}-\boldsymbol{\Phi m_N}||^2 + \frac{\beta}{2}\frac{d}{d\beta}||\mathbf{t}-\boldsymbol{\Phi m_N}||^2 + \frac{d}{d\beta}\frac{\alpha}{2}\boldsymbol{m_N}^T\boldsymbol{m_N}$$

The last two terms in the equation above can be further written as:

$$\begin{aligned}
\frac{\beta}{2}\frac{d}{d\beta}||\mathbf{t}-\boldsymbol{\Phi m_N}||^2 + \frac{d}{d\beta}\frac{\alpha}{2}\boldsymbol{m_N}^T\boldsymbol{m_N} &= \{\frac{\beta}{2}\frac{d}{d\boldsymbol{m_N}}||\mathbf{t}-\boldsymbol{\Phi m_N}||^2 + \frac{d}{d\boldsymbol{m_N}}\frac{\alpha}{2}\boldsymbol{m_N}^T\boldsymbol{m_N}\}\cdot\frac{d\boldsymbol{m_N}}{d\beta} \\
&= \{\frac{\beta}{2}[-2\boldsymbol{\Phi}^T(\mathbf{t}-\boldsymbol{\Phi m_N})] + \frac{\alpha}{2}2\boldsymbol{m_N}\}\cdot\frac{d\boldsymbol{m_N}}{d\beta} \\
&= \{-\beta\boldsymbol{\Phi}^T(\mathbf{t}-\boldsymbol{\Phi m_N}) + \alpha\boldsymbol{m_N}\}\cdot\frac{d\boldsymbol{m_N}}{d\beta} \\
&= \{-\beta\boldsymbol{\Phi}^T\mathbf{t} + (\alpha\mathbf{I}+\beta\boldsymbol{\Phi}^T\boldsymbol{\Phi})\boldsymbol{m_N}\}\cdot\frac{d\boldsymbol{m_N}}{d\beta} \\
&= \{-\beta\boldsymbol{\Phi}^T\mathbf{t} + \mathbf{A}\boldsymbol{m_N}\}\cdot\frac{d\boldsymbol{m_N}}{d\beta} \\
&= 0
\end{aligned}$$

Where we have taken advantage of (3.83) and (3.84). Hence

$$\frac{d}{d\beta}E(\boldsymbol{m_N}) = \frac{1}{2}||\mathbf{t}-\boldsymbol{\Phi m_N}||^2 = \frac{1}{2}\sum_{n=1}^{N}(t_n - \boldsymbol{m_N}^T\boldsymbol{\phi}(\boldsymbol{x_n}))^2$$

The last term dependent on β in (3.86) is:

$$\frac{d}{d\beta}(\frac{1}{2}ln|\mathbf{A}|) = \frac{\gamma}{2\beta}$$

Therefore, if we combine all those expressions together, we will obtain (3.94). And then if we rearrange it, we will obtain (3.95).

Problem 3.23 Solution

First, according to (3.10), we know that $p(\mathbf{t}|\mathbf{X},\boldsymbol{w},\beta)$ can be further written as $p(\mathbf{t}|\mathbf{X},\boldsymbol{w},\beta) = \mathcal{N}(\mathbf{t}|\boldsymbol{\Phi w},\beta^{-1}\mathbf{I})$, and given that $p(\boldsymbol{w}|\beta) = \mathcal{N}(\boldsymbol{m_0},\beta^{-1}\mathbf{S_0})$ and

$p(\beta) = \text{Gam}(\beta|a_0, b_0)$. Therefore, we just follow the hint in the problem.

$$\begin{aligned}
p(\mathbf{t}) &= \int\int p(\mathbf{t}|\mathbf{X}, \boldsymbol{w}, \beta)\, p(\boldsymbol{w}|\beta)\, d\boldsymbol{w}\, p(\beta)\, d\beta \\
&= \int\int (\frac{\beta}{2\pi})^{N/2} exp\{-\frac{\beta}{2}(\mathbf{t}-\boldsymbol{\Phi w})^T(\mathbf{t}-\boldsymbol{\Phi w})\} \cdot \\
&\quad (\frac{\beta}{2\pi})^{M/2} |\mathbf{S_0}|^{-1/2} exp\{-\frac{\beta}{2}(\boldsymbol{w}-\boldsymbol{m_0})^T \mathbf{S_0}^{-1}(\boldsymbol{w}-\boldsymbol{m_0})\}\, d\boldsymbol{w} \\
&\quad \Gamma(a_0)^{-1} b_0^{a_0} \beta^{a_0-1} exp(-b_0\beta)\, d\beta \\
&= \frac{b_0^{a_0}}{(2\pi)^{(M+N)/2}|\mathbf{S_0}|^{1/2}} \int\int exp\{-\frac{\beta}{2}(\mathbf{t}-\boldsymbol{\Phi w})^T(\mathbf{t}-\boldsymbol{\Phi w})\} \\
&\quad exp\{-\frac{\beta}{2}(\boldsymbol{w}-\boldsymbol{m_0})^T \mathbf{S_0}^{-1}(\boldsymbol{w}-\boldsymbol{m_0})\}\, d\boldsymbol{w} \\
&\quad \beta^{a_0-1+N/2+M/2} exp(-b_0\beta)\, d\beta \\
&= \frac{b_0^{a_0}}{(2\pi)^{(M+N)/2}|\mathbf{S_0}|^{1/2}} \int\int exp\{-\frac{\beta}{2}(\boldsymbol{w}-\boldsymbol{m_N})^T \mathbf{S_N}^{-1}(\boldsymbol{w}-\boldsymbol{m_N})\}\, d\boldsymbol{w} \\
&\quad exp\{-\frac{\beta}{2}(\mathbf{t}^T\mathbf{t} + \boldsymbol{m_0}^T \mathbf{S_0}^{-1}\boldsymbol{m_0} - \boldsymbol{m_N}^T \mathbf{S_N}^{-1}\boldsymbol{m_N})\} \\
&\quad \beta^{a_N-1+M/2} exp(-b_0\beta)\, d\beta
\end{aligned}$$

Where we have defined

$$\boldsymbol{m_N} = \mathbf{S_N}(\mathbf{S_0}^{-1}\boldsymbol{m_0} + \Phi^T\mathbf{t})$$

$$\boldsymbol{S_N}^{-1} = \mathbf{S_0}^{-1} + \Phi^T\Phi$$

$$a_N = a_0 + \frac{N}{2}$$

$$b_N = b_0 + \frac{1}{2}(\boldsymbol{m_0}^T \mathbf{S_0}^{-1}\boldsymbol{m_0} - \boldsymbol{m_N}^T \mathbf{S_N}^{-1}\boldsymbol{m_N} + \sum_{n=1}^{N} t_n^2)$$

Which are exactly the same as those in Prob.3.12, and then we evaluate the integral, taking advantage of the normalized property of multivariate Gaussian Distribution and Gamma Distribution.

$$\begin{aligned}
p(\mathbf{t}) &= \frac{b_0^{a_0}}{(2\pi)^{(M+N)/2}|\mathbf{S_0}|^{1/2}} (\frac{2\pi}{\beta})^{M/2} |\mathbf{S_N}|^{1/2} \int \beta^{a_N-1+M/2} exp(-b_N\beta)\, d\beta \\
&= \frac{b_0^{a_0}}{(2\pi)^{(M+N)/2}|\mathbf{S_0}|^{1/2}} (2\pi)^{M/2} |\mathbf{S_N}|^{1/2} \int \beta^{a_N-1} exp(-b_N\beta)\, d\beta \\
&= \frac{1}{(2\pi)^{N/2}} \frac{|\mathbf{S_N}|^{1/2}}{|\mathbf{S_0}|^{1/2}} \frac{b_0^{a_0}}{b_N^{a_N}} \frac{\Gamma(a_N)}{\Gamma(b_N)}
\end{aligned}$$

Just as required.

Problem 3.24 Solution

Let's just follow the hint and we begin by writing down expression for the likelihood, prior and posterior PDF. We know that $p(\mathbf{t}|\boldsymbol{w},\beta) = \mathcal{N}(\mathbf{t}|\boldsymbol{\Phi w},\beta^{-1}\mathbf{I})$. What's more, the form of the prior and posterior are quite similar:

$$p(\boldsymbol{w},\beta) = \mathcal{N}(\boldsymbol{w}|\mathbf{m_0},\beta^{-1}\mathbf{S_0})\,\mathrm{Gam}(\beta|a_0,b_0)$$

And

$$p(\boldsymbol{w},\beta|\mathbf{t}) = \mathcal{N}(\boldsymbol{w}|\mathbf{m_N},\beta^{-1}\mathbf{S_N})\,\mathrm{Gam}(\beta|a_N,b_N)$$

Where the relationships among those parameters are shown in Prob.3.12, Prob.3.23. Now according to (3.119), we can write:

$$\begin{aligned}
p(\mathbf{t}) &= \mathcal{N}(\mathbf{t}|\boldsymbol{\Phi w},\beta^{-1}\mathbf{I})\,\frac{\mathcal{N}(\boldsymbol{w}|\mathbf{m_0},\beta^{-1}\mathbf{S_0})\,\mathrm{Gam}(\beta|a_0,b_0)}{\mathcal{N}(\boldsymbol{w}|\mathbf{m_N},\beta^{-1}\mathbf{S_N})\,\mathrm{Gam}(\beta|a_N,b_N)} \\
&= \mathcal{N}(\mathbf{t}|\boldsymbol{\Phi w},\beta^{-1}\mathbf{I})\,\frac{\mathcal{N}(\boldsymbol{w}|\mathbf{m_0},\beta^{-1}\mathbf{S_0})}{\mathcal{N}(\boldsymbol{w}|\mathbf{m_N},\beta^{-1}\mathbf{S_N})}\,\frac{b_0^{a_0}\beta^{a_0-1}exp(-b_0\beta)/\Gamma(a_0)}{b_N^{a_N}\beta^{a_N-1}exp(-b_N\beta)/\Gamma(a_N)} \\
&= \mathcal{N}(\mathbf{t}|\boldsymbol{\Phi w},\beta^{-1}\mathbf{I})\,\frac{\mathcal{N}(\boldsymbol{w}|\mathbf{m_0},\beta^{-1}\mathbf{S_0})}{\mathcal{N}(\boldsymbol{w}|\mathbf{m_N},\beta^{-1}\mathbf{S_N})}\,\frac{b_0^{a_0}}{b_N^{a_N}}\,\frac{\Gamma(a_N)}{\Gamma(a_0)}\beta^{a_0-a_N}exp\{-(b_0-b_N)\beta\} \\
&= \mathcal{N}(\mathbf{t}|\boldsymbol{\Phi w},\beta^{-1}\mathbf{I})\,\frac{\mathcal{N}(\boldsymbol{w}|\mathbf{m_0},\beta^{-1}\mathbf{S_0})}{\mathcal{N}(\boldsymbol{w}|\mathbf{m_N},\beta^{-1}\mathbf{S_N})}\,exp\{-(b_0-b_N)\beta\}\,\frac{b_0^{a_0}}{b_N^{a_N}}\,\frac{\Gamma(a_N)}{\Gamma(a_0)}\beta^{-N/2}
\end{aligned}$$

Where we have used $a_N = a_0 + \frac{N}{2}$. Now we deal with the terms expressed in the form of Gaussian Distribution:

$$\begin{aligned}
\text{Gaussian terms} &= \mathcal{N}(\mathbf{t}|\boldsymbol{\Phi w},\beta^{-1}\mathbf{I})\,\frac{\mathcal{N}(\boldsymbol{w}|\mathbf{m_0},\beta^{-1}\mathbf{S_0})}{\mathcal{N}(\boldsymbol{w}|\mathbf{m_N},\beta^{-1}\mathbf{S_N})} \\
&= (\frac{\beta}{2\pi})^{N/2}exp\{-\frac{\beta}{2}(\mathbf{t}-\boldsymbol{\Phi w})^T(\mathbf{t}-\boldsymbol{\Phi w})\}\cdot \\
&\quad \frac{|\beta^{-1}\mathbf{S_N}|^{1/2}}{|\beta^{-1}\mathbf{S_0}|^{1/2}}\,\frac{exp\{-\frac{\beta}{2}(\boldsymbol{w}-\mathbf{m_0})^T\mathbf{S_0}^{-1}(\boldsymbol{w}-\mathbf{m_0})\}}{exp\{-\frac{\beta}{2}(\boldsymbol{w}-\mathbf{m_N})^T\mathbf{S_N}^{-1}(\boldsymbol{w}-\mathbf{m_N})\}} \\
&= (\frac{\beta}{2\pi})^{N/2}\frac{|\mathbf{S_N}|^{1/2}}{|\mathbf{S_0}|^{1/2}}exp\{-\frac{\beta}{2}(\mathbf{t}-\boldsymbol{\Phi w})^T(\mathbf{t}-\boldsymbol{\Phi w})\}\cdot \\
&\quad \frac{exp\{-\frac{\beta}{2}(\boldsymbol{w}-\mathbf{m_0})^T\mathbf{S_0}^{-1}(\boldsymbol{w}-\mathbf{m_0})\}}{exp\{-\frac{\beta}{2}(\boldsymbol{w}-\mathbf{m_N})^T\mathbf{S_N}^{-1}(\boldsymbol{w}-\mathbf{m_N})\}}
\end{aligned}$$

We look back to the previous problem and we notice that at the last step in the deduction of $p(\mathbf{t})$, we complete the square according to $\boldsymbol{w}$. And if we carefully compare the left and right side at the last step, we can obtain :

$$\begin{aligned}
& exp\{-\frac{\beta}{2}(\mathbf{t}-\boldsymbol{\Phi w})^T(\mathbf{t}-\boldsymbol{\Phi w})\}exp\{-\frac{\beta}{2}(\boldsymbol{w}-\boldsymbol{m_0})^T\mathbf{S_0}^{-1}(\boldsymbol{w}-\boldsymbol{m_0})\} \\
= \;& exp\{-\frac{\beta}{2}(\boldsymbol{w}-\boldsymbol{m_N})^T\mathbf{S_N}^{-1}(\boldsymbol{w}-\boldsymbol{m_N})\}exp\{-(b_N-b_0)\beta\}
\end{aligned}$$

Hence, we go back to deal with the Gaussian terms:

$$\text{Gaussian terms} \quad = \quad (\frac{\beta}{2\pi})^{N/2} \frac{|\mathbf{S_N}|^{1/2}}{|\mathbf{S_0}|^{1/2}} exp\{-(b_N - b_0)\beta\}$$

If we substitute the expressions above into $p(\mathbf{t})$, we will obtain (3.118) immediately.

0.4 Linear Models Classification

Problem 4.1 Solution

If the convex hull of $\{\mathbf{x_n}\}$ and $\{\mathbf{y_n}\}$ intersects, we know that there will be a point $\mathbf{z}$ which can be written as $\mathbf{z} = \sum_n \alpha_n \mathbf{x_n}$ and also $\mathbf{z} = \sum_n \beta_n \mathbf{y_n}$. Hence we can obtain:

$$\begin{aligned} \widehat{\mathbf{w}}^T \mathbf{z} + w_0 &= \widehat{\mathbf{w}}^T (\sum_n \alpha_n \mathbf{x_n}) + w_0 \\ &= (\sum_n \alpha_n \widehat{\mathbf{w}}^T \mathbf{x_n}) + (\sum_n \alpha_n) w_0 \\ &= \sum_n \alpha_n (\widehat{\mathbf{w}}^T \mathbf{x_n} + w_0) \quad (*) \end{aligned}$$

Where we have used $\sum_n \alpha_n = 1$. And if $\{\mathbf{x_n}\}$ and $\{\mathbf{y_n}\}$ are linearly separable, we have $\widehat{\mathbf{w}}^T \mathbf{x_n} + w_0 > 0$ and $\widehat{\mathbf{w}}^T \mathbf{y_n} + w_0 < 0$, for $\forall \mathbf{x_n}, \mathbf{y_n}$. Together with $\alpha_n \geq 0$ and $(*)$, we know that $\widehat{\mathbf{w}}^T \mathbf{z} + w_0 > 0$. And if we calculate $\widehat{\mathbf{w}}^T \mathbf{z} + w_0$ from the perspective of $\{\mathbf{y_n}\}$ following the same procedure, we can obtain $\widehat{\mathbf{w}}^T \mathbf{z} + w_0 < 0$. Hence contradictory occurs. In other words, they are not linearly separable if their convex hulls intersect.

Now let's assume they are linearly separable and try to prove their convex hulls don't intersect. This is obvious. We can obtain $\widehat{\mathbf{w}}^T \mathbf{x_n} + w_0 > 0$ and $\widehat{\mathbf{w}}^T \mathbf{y_n} + w_0 < 0$, for $\forall \mathbf{x_n}, \mathbf{y_n}$, if the two sets are linearly separable. And if there is a point $\mathbf{z}$, which belongs to both convex hulls of $\{\mathbf{x_n}\}$ and $\{\mathbf{y_n}\}$, following the same procedure as above, we can see that $\widehat{\mathbf{w}}^T \mathbf{z} + w_0 > 0$ from the perspective of $\{\mathbf{x_n}\}$ and $\widehat{\mathbf{w}}^T \mathbf{z} + w_0 < 0$ from the perspective of $\{\mathbf{y_n}\}$, which directly leads to a contradictory. Therefore, the convex hulls don't intersect.

Problem 4.2 Solution

Let's make the dependency of $E_D(\widetilde{\mathbf{W}})$ on w_0 explicitly:

$$E_D(\widetilde{\mathbf{W}}) = \frac{1}{2} \text{Tr}\{(\mathbf{XW} + \mathbf{1w_0}^T - \mathbf{T})^T (\mathbf{XW} + \mathbf{1w_0}^T - \mathbf{T})\}$$

Then we calculate the derivative of $E_D(\widetilde{\mathbf{W}})$ with respect to $\mathbf{w_0}$:

$$\frac{\partial E_D(\widetilde{\mathbf{W}})}{\partial \mathbf{w_0}} = 2N\mathbf{w_0} + 2(\mathbf{XW} - \mathbf{T})^T \mathbf{1}$$

Where we have used the property:

$$\frac{\partial}{\partial \mathbf{X}} \mathrm{Tr}\left[(\mathbf{AXB}+\mathbf{C})(\mathbf{AXB}+\mathbf{C})^T\right] = 2\mathbf{A}^T(\mathbf{AXB}+\mathbf{C})\mathbf{B}^T$$

We set the derivative equals to 0, which gives:

$$\mathbf{w_0} = -\frac{1}{N}(\mathbf{XW}-\mathbf{T})^T\mathbf{1} = \bar{\mathbf{t}} - \mathbf{W}^T\bar{\mathbf{x}}$$

Where we have denoted:

$$\bar{\mathbf{t}} = \frac{1}{N}\mathbf{T}^T\mathbf{1}, \quad \text{and} \quad \bar{\mathbf{x}} = \frac{1}{N}\mathbf{X}^T\mathbf{1}$$

If we substitute the equations above into $E_D(\widetilde{\mathbf{W}})$, we can obtain:

$$E_D(\widetilde{\mathbf{W}}) = \frac{1}{2}\mathrm{Tr}\{(\mathbf{XW}+\bar{\mathbf{T}}-\bar{\mathbf{X}}\mathbf{W}-\mathbf{T})^T(\mathbf{XW}+\bar{\mathbf{T}}-\bar{\mathbf{X}}\mathbf{W}-\mathbf{T})\}$$

Where we further denote

$$\bar{\mathbf{T}} = \mathbf{1}\bar{\mathbf{t}}^T, \quad \text{and} \quad \bar{\mathbf{X}} = \mathbf{1}\bar{\mathbf{x}}^T$$

Then we set the derivative of $E_D(\widetilde{\mathbf{W}})$ with regard to $\mathbf{W}$ to 0, which gives:

$$\mathbf{W} = \widehat{\mathbf{X}}^{\dagger}\widehat{\mathbf{T}}$$

Where we have defined:

$$\widehat{\mathbf{X}} = \mathbf{X}-\bar{\mathbf{X}}, \quad \text{and} \quad \widehat{\mathbf{T}} = \mathbf{T}-\bar{\mathbf{T}}$$

Now consider the prediction for a new given $\mathbf{x}$, we have:

$$\begin{aligned}
\mathbf{y}(\mathbf{x}) &= \mathbf{W}^T\mathbf{x}+\mathbf{w_0} \\
&= \mathbf{W}^T\mathbf{x}+\bar{\mathbf{t}}-\mathbf{W}^T\bar{\mathbf{x}} \\
&= \bar{\mathbf{t}}+\mathbf{W}^T(\mathbf{x}-\bar{\mathbf{x}})
\end{aligned}$$

If we know that $\mathbf{a}^T\mathbf{t_n}+b = 0$ holds for some $\mathbf{a}$ and b, we can obtain:

$$\mathbf{a}^T\bar{\mathbf{t}} = \frac{1}{N}\mathbf{a}^T\mathbf{T}^T\mathbf{1} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{a^T t_n} = -b$$

Therefore,

$$\begin{aligned}
\mathbf{a}^T\mathbf{y}(\mathbf{x}) &= \mathbf{a}^T\left[\bar{\mathbf{t}}+\mathbf{W}^T(\mathbf{x}-\bar{\mathbf{x}})\right] \\
&= \mathbf{a}^T\bar{\mathbf{t}}+\mathbf{a}^T\mathbf{W}^T(\mathbf{x}-\bar{\mathbf{x}}) \\
&= -b+\mathbf{a}^T\widehat{\mathbf{T}}^T(\widehat{\mathbf{X}}^{\dagger})^T(\mathbf{x}-\bar{\mathbf{x}}) \\
&= -b
\end{aligned}$$

Where we have used:

$$\begin{aligned}\mathbf{a}^T\widehat{\mathbf{T}}^T &= \mathbf{a}^T(\mathbf{T}-\bar{\mathbf{T}})^T = \mathbf{a}^T(\mathbf{T}-\frac{1}{N}\mathbf{1}\mathbf{1}^\mathbf{T}\mathbf{T})^T \\ &= \mathbf{a}^T\mathbf{T}^T - \frac{1}{N}\mathbf{a}^T\mathbf{T}^T\mathbf{1}\mathbf{1}^\mathbf{T} = -b\mathbf{1}^T + b\mathbf{1}^T \\ &= \mathbf{0}^T\end{aligned}$$

Problem 4.3 Solution

Suppose there are Q constraints in total. We can write $\mathbf{a_q}^T\mathbf{t_n}+b_q = 0$, $q = 1,2,...,Q$ for all the target vector $\mathbf{t_n}$, $n = 1,2...,N$. Or alternatively, we can group them together:

$$\mathbf{A}^T\mathbf{t_n}+\mathbf{b} = \mathbf{0}$$

Where $\mathbf{A}$ is a $Q \times Q$ matrix, and the qth column of $\mathbf{A}$ is $\mathbf{a_q}$, and meanwhile $\mathbf{b}$ is a $Q \times 1$ column vector, and the qth element is $\mathbf{b_q}$. for every pair of $\{\mathbf{a_q}, b_q\}$ we can follow the same procedure in the previous problem to show that $\mathbf{a_q y(x)} + b_q = 0$. In other words, the proofs will not affect each other. Therefore, it is obvious :

$$\mathbf{A}^T\mathbf{y(x)}+\mathbf{b} = \mathbf{0}$$

Problem 4.4 Solution

We use Lagrange multiplier to enforce the constraint $\mathbf{w}^T\mathbf{w} = 1$. We now need to maximize :

$$L(\lambda,\mathbf{w}) = \mathbf{w}^T(\mathbf{m_2}-\mathbf{m_1}) + \lambda(\mathbf{w}^T\mathbf{w}-1)$$

We calculate the derivatives:

$$\frac{\partial L(\lambda,\mathbf{w})}{\partial\lambda} = \mathbf{w}^T\mathbf{w}-1$$

And

$$\frac{\partial L(\lambda,\mathbf{w})}{\partial\mathbf{w}} = \mathbf{m_2}-\mathbf{m_1}+2\lambda\mathbf{w}$$

We set the derivatives above equals to 0, which gives:

$$\mathbf{w} = -\frac{1}{2\lambda}(\mathbf{m_2}-\mathbf{m_1}) \propto (\mathbf{m_2}-\mathbf{m_1})$$

Problem 4.5 Solution

We expand (4.25) using (4.22), (4.23) and (4.24).

$$\begin{aligned}J(\mathbf{w}) &= \frac{(m_2-m_1)^2}{s_1^2+s_2^2} \\ &= \frac{||\mathbf{w}^T(\mathbf{m_2}-\mathbf{m_1})||^2}{\sum_{n\in C_1}(\mathbf{w}^T\mathbf{x_n}-m_1)^2+\sum_{n\in C_2}(\mathbf{w}^T\mathbf{x_n}-m_2)^2}\end{aligned}$$

The numerator can be further written as:

$$\text{numerator} = \left[\mathbf{w}^T(\mathbf{m_2}-\mathbf{m_1})\right]\left[\mathbf{w}^T(\mathbf{m_2}-\mathbf{m_1})\right]^T = \mathbf{w}^T\mathbf{S_B}\mathbf{w}$$

Where we have defined:

$$\mathbf{S_B} = (\mathbf{m_2}-\mathbf{m_1})(\mathbf{m_2}-\mathbf{m_1})^T$$

And ti is the same for the denominator:

$$\begin{aligned}\text{denominator} &= \sum_{n\in C_1}[\mathbf{w}^T(\mathbf{x_n}-\mathbf{m_1})]^2 + \sum_{n\in C_2}[\mathbf{w}^T(\mathbf{x_n}-\mathbf{m_2})]^2 \\ &= \mathbf{w}^T\mathbf{S_{w1}}\mathbf{w} + \mathbf{w}^T\mathbf{S_{w2}}\mathbf{w} \\ &= \mathbf{w}^T\mathbf{S_w}\mathbf{w}\end{aligned}$$

Where we have defined:

$$\mathbf{S_w} = \sum_{n\in C_1}(\mathbf{x_n}-\mathbf{m_1})(\mathbf{x_n}-\mathbf{m_1})^T + \sum_{n\in C_2}(\mathbf{x_n}-\mathbf{m_2})(\mathbf{x_n}-\mathbf{m_2})^T$$

Just as required.

Problem 4.6 Solution

Let's follow the hint, beginning by expanding (4.33).

$$\begin{aligned}(4.33) &= \sum_{n=1}^{N}\mathbf{w}^T\mathbf{x_n}\mathbf{x_n} + w_0\sum_{n=1}^{N}\mathbf{x_n} - \sum_{n=1}^{N}t_n\mathbf{x_n} \\ &= \sum_{n=1}^{N}\mathbf{x_n}\mathbf{x_n}^T\mathbf{w} - \mathbf{w}^T\mathbf{m}\sum_{n=1}^{N}\mathbf{x_n} - (\sum_{n\in C_1}t_n\mathbf{x_n} + \sum_{n\in C_2}t_n\mathbf{x_n}) \\ &= \sum_{n=1}^{N}\mathbf{x_n}\mathbf{x_n}^T\mathbf{w} - \mathbf{w}^T\mathbf{m}\cdot(N\mathbf{m}) - (\sum_{n\in C_1}\frac{N}{N_1}\mathbf{x_n} + \sum_{n\in C_2}\frac{-N}{N_2}\mathbf{x_n}) \\ &= \sum_{n=1}^{N}\mathbf{x_n}\mathbf{x_n}^T\mathbf{w} - N\mathbf{w}^T\mathbf{m}\mathbf{m} - N(\sum_{n\in C_1}\frac{1}{N_1}\mathbf{x_n} - \sum_{n\in C_2}\frac{1}{N_2}\mathbf{x_n}) \\ &= \sum_{n=1}^{N}\mathbf{x_n}\mathbf{x_n}^T\mathbf{w} - N\mathbf{m}\mathbf{m}^T\mathbf{w} - N(\mathbf{m_1}-\mathbf{m_2}) \\ &= [\sum_{n=1}^{N}(\mathbf{x_n}\mathbf{x_n}^T) - N\mathbf{m}\mathbf{m}^T]\mathbf{w} - N(\mathbf{m_1}-\mathbf{m_2})\end{aligned}$$

If we let the derivative equal to 0, we will see that:

$$[\sum_{n=1}^{N}(\mathbf{x_n}\mathbf{x_n}^T) - N\mathbf{m}\mathbf{m}^T]\mathbf{w} = N(\mathbf{m_1}-\mathbf{m_2})$$

Therefore, now we need to prove:

$$\sum_{n=1}^{N}(\mathbf{x_n}\mathbf{x_n}^T) - N\mathbf{m}\mathbf{m}^T = \mathbf{S_w} + \frac{N_1N_2}{N}\mathbf{S_B}$$

Let's expand the left side of the equation above:

$$\begin{aligned}
\text{left} &= \sum_{n=1}^{N} \mathbf{x_n}\mathbf{x_n}^T - N(\frac{N_1}{N}\mathbf{m_1} + \frac{N_2}{N}\mathbf{m_2})^2 \\
&= \sum_{n=1}^{N} \mathbf{x_n}\mathbf{x_n}^T - N(\frac{N_1^2}{N^2}||\mathbf{m_1}||^2 + \frac{N_2^2}{N^2}||\mathbf{m_2}||^2 + 2\frac{N_1N_2}{N^2}\mathbf{m_1}\mathbf{m_2}^T) \\
&= \sum_{n=1}^{N} \mathbf{x_n}\mathbf{x_n}^T - \frac{N_1^2}{N}||\mathbf{m_1}||^2 - \frac{N_2^2}{N}||\mathbf{m_2}||^2 - 2\frac{N_1N_2}{N}\mathbf{m_1}\mathbf{m_2}^T \\
&= \sum_{n=1}^{N} \mathbf{x_n}\mathbf{x_n}^T + (N_1 + \frac{N_1N_2}{N} - 2N_1)||\mathbf{m_1}||^2 + (N_2 + \frac{N_1N_2}{N} - 2N_2)||\mathbf{m_2}||^2 - 2\frac{N_1N_2}{N}\mathbf{m_1}\mathbf{m_2}^T \\
&= \sum_{n=1}^{N} \mathbf{x_n}\mathbf{x_n}^T + (N_1 - 2N_1)||\mathbf{m_1}||^2 + (N_2 - 2N_2)||\mathbf{m_2}||^2 + \frac{N_1N_2}{N}||\mathbf{m_1} - \mathbf{m_2}||^2 \\
&= \sum_{n=1}^{N} \mathbf{x_n}\mathbf{x_n}^T + N_1||\mathbf{m_1}||^2 - 2\mathbf{m_1}\cdot(N_1\mathbf{m_1}^T) + N_2||\mathbf{m_2}||^2 - 2\mathbf{m_2}\cdot(N_2\mathbf{m_2}^T) + \frac{N_1N_2}{N}\mathbf{S_B} \\
&= \sum_{n=1}^{N} \mathbf{x_n}\mathbf{x_n}^T + N_1||\mathbf{m_1}||^2 - 2\mathbf{m_1}\sum_{n\in C_1} x_n^T + N_2||\mathbf{m_2}||^2 - 2\mathbf{m_2}\sum_{n\in C_2} x_n^T + \frac{N_1N_2}{N}\mathbf{S_B} \\
&= \sum_{n\in C_1} \mathbf{x_n}\mathbf{x_n}^T + N_1||\mathbf{m_1}||^2 - 2\mathbf{m_1}\sum_{n\in C_1} x_n^T \\
&\quad + \sum_{n\in C_2} \mathbf{x_n}\mathbf{x_n}^T + N_2||\mathbf{m_2}||^2 - 2\mathbf{m_2}\sum_{n\in C_2} x_n^T + \frac{N_1N_2}{N}\mathbf{S_B} \\
&= \sum_{n\in C_1} (\mathbf{x_n}\mathbf{x_n}^T + ||\mathbf{m_1}||^2 - 2\mathbf{m_1}x_n^T) + \sum_{n\in C_2} (\mathbf{x_n}\mathbf{x_n}^T + ||\mathbf{m_2}||^2 - 2\mathbf{m_2}\mathbf{x_n}^T) + \frac{N_1N_2}{N}\mathbf{S_B} \\
&= \sum_{n\in C_1} ||\mathbf{x_n} - \mathbf{m_1}||^2 + \sum_{n\in C_2} ||\mathbf{x_n} - \mathbf{m_2}||^2 + \frac{N_1N_2}{N}\mathbf{S_B} \\
&= \mathbf{S_w} + \frac{N_1N_2}{N}\mathbf{S_B}
\end{aligned}$$

Just as required.

Problem 4.7 Solution

This problem is quite simple. We can solve it by definition. We know that logistic sigmoid function has the form:

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

Therefore, we can obtain:

$$\begin{aligned}
\sigma(a) + \sigma(-a) &= \frac{1}{1 + exp(-a)} + \frac{1}{1 + exp(a)} \\
&= \frac{2 + exp(a) + exp(-a)}{[1 + exp(-a)][1 + exp(a)]} \\
&= \frac{2 + exp(a) + exp(-a)}{2 + exp(a) + exp(-a)} = 1
\end{aligned}$$

Next we exchange the dependent and independent variables to obtain its inverse.

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

We first rearrange the equation above, which gives:

$$exp(-y) = \frac{1-a}{a}$$

Then we calculate the logarithm for both sides, which gives:

$$y = \ln(\frac{a}{1-a})$$

Just as required.

Problem 4.8 Solution

According to (4.58) and (4.64), we can write:

$$\begin{aligned}
a &= \ln \frac{p(\mathbf{x}|C_1)p(C_1)}{p(\mathbf{x}|C_2)p(C_2)} \\
&= \ln p(\mathbf{x}|C_1) - \ln p(\mathbf{x}|C_2) + \ln \frac{p(C_1)}{p(C_2)} \\
&= -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu_1})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu_1}) + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu_2})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu_2}) + \ln \frac{p(C_1)}{p(C_2)} \\
&= \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu_1}-\boldsymbol{\mu_2})\mathbf{x} - \frac{1}{2}\boldsymbol{\mu_1}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_1} + \frac{1}{2}\boldsymbol{\mu_2}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_2} + \ln \frac{p(C_1)}{p(C_2)} \\
&= \mathbf{w}^T\mathbf{x} + w_0
\end{aligned}$$

Where in the last second step, we rearrange the term according to $\mathbf{x}$, i.e., its quadratic, linear, constant term. We have also defined :

$$\mathbf{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu_1}-\boldsymbol{\mu_2})$$

And

$$w_0 = -\frac{1}{2}\boldsymbol{\mu_1}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_1} + \frac{1}{2}\boldsymbol{\mu_2}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_2} + \ln \frac{p(C_1)}{p(C_2)}$$

Finally, since $p(C_1|\mathbf{x}) = \sigma(a)$ as stated in (4.57), we have $p(C_1|\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x} + w_0)$ just as required.

Problem 4.9 Solution

We begin by writing down the likelihood function.

$$\begin{aligned}
p(\{\phi_\mathbf{n}, t_n\}|\pi_1, \pi_2, ..., \pi_K) &= \prod_{n=1}^{N}\prod_{k=1}^{K}[p(\boldsymbol{\phi_n}|C_k)\, p(C_k)]^{t_{nk}} \\
&= \prod_{n=1}^{N}\prod_{k=1}^{K}[\pi_k\, p(\boldsymbol{\phi_n}|C_k)]^{t_{nk}}
\end{aligned}$$

Hence we can obtain the expression for the logarithm likelihood:

$$\ln p = \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\left[\ln \pi_k + \ln p(\boldsymbol{\phi_n}|C_k)\right] \propto \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \ln \pi_k$$

Since there is a constraint on π_k, so we need to add a Lagrange Multiplier to the expression, which becomes:

$$L = \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \ln \pi_k + \lambda(\sum_{k=1}^{K} \pi_k - 1)$$

We calculate the derivative of the expression above with regard to π_k:

$$\frac{\partial L}{\partial \pi_k} = \sum_{n=1}^{N} \frac{t_{nk}}{\pi_k} + \lambda$$

And if we set the derivative equal to 0, we can obtain:

$$\pi_k = -(\sum_{n=1}^{N} t_{nk})/\lambda = -\frac{N_k}{\lambda} \qquad (*)$$

And if we preform summation on both sides with regard to k, we can see that:

$$1 = -(\sum_{k=1}^{K} N_k)/\lambda = -\frac{N}{\lambda}$$

Which gives $\lambda = -N$, and substitute it into $(*)$, we can obtain $\pi_k = N_k/N$.

Problem 4.10 Solution

This time, we focus on the term which dependent on $\boldsymbol{\mu_k}$ and $\boldsymbol{\Sigma}$ in the logarithm likelihood.

$$\ln p = \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\left[\ln \pi_k + \ln p(\boldsymbol{\phi_n}|C_k)\right] \propto \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \ln p(\boldsymbol{\phi_n}|C_k)$$

Provided $p(\phi|C_k) = \mathcal{N}(\phi|\boldsymbol{\mu_k}, \boldsymbol{\Sigma})$, we can further derive:

$$\ln p \propto \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\left[-\frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}(\boldsymbol{\phi_n} - \boldsymbol{\mu_k})\boldsymbol{\Sigma}^{-1}(\boldsymbol{\phi_n} - \boldsymbol{\mu_k})^T\right]$$

We first calculate the derivative of the expression above with regard to $\boldsymbol{\mu_k}$:

$$\frac{\partial \ln p}{\partial \boldsymbol{\mu_k}} = \sum_{n=1}^{N} t_{nk}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\phi_n} - \boldsymbol{\mu_k})$$

We set the derivative equals to 0, which gives:

$$\sum_{n=1}^{N} t_{nk}\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi_n} = \sum_{n=1}^{N} t_{nk}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_k} = N_k\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_k}$$

Therefore, if we multiply both sides by $\mathbf{\Sigma}/N_k$, we will obtain (4.161). Now let's calculate the derivative of $\ln p$ with regard to $\mathbf{\Sigma}$, which gives:

$$\begin{aligned}
\frac{\partial \ln p}{\partial \mathbf{\Sigma}} &= \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}(-\frac{1}{2}\mathbf{\Sigma}^{-1}) - \frac{1}{2}\frac{\partial}{\partial \mathbf{\Sigma}}\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}(\boldsymbol{\phi_n}-\boldsymbol{\mu_k})\mathbf{\Sigma}^{-1}(\boldsymbol{\phi_n}-\boldsymbol{\mu_k})^T \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} -\frac{t_{nk}}{2}\mathbf{\Sigma}^{-1} - \frac{1}{2}\frac{\partial}{\partial \mathbf{\Sigma}}\sum_{k=1}^{K}\sum_{n=1}^{N} t_{nk}(\boldsymbol{\phi_n}-\boldsymbol{\mu_k})\mathbf{\Sigma}^{-1}(\boldsymbol{\phi_n}-\boldsymbol{\mu_k})^T \\
&= \sum_{n=1}^{N} -\frac{1}{2}\mathbf{\Sigma}^{-1} - \frac{1}{2}\frac{\partial}{\partial \mathbf{\Sigma}}\sum_{k=1}^{K} N_k \mathrm{Tr}(\mathbf{\Sigma}^{-1}\mathbf{S_k}) \\
&= -\frac{N}{2}\mathbf{\Sigma}^{-1} + \frac{1}{2}\sum_{k=1}^{K} N_k \mathbf{\Sigma}^{-1}\mathbf{S_k}\mathbf{\Sigma}^{-1}
\end{aligned}$$

Where we have denoted

$$\mathbf{S_k} = \frac{1}{N_k}\sum_{n=1}^{N} t_{nk}(\boldsymbol{\phi_n}-\boldsymbol{\mu_k})(\boldsymbol{\phi_n}-\boldsymbol{\mu_k})^T$$

Now we set the derivative equals to 0, and rearrange the equation, which gives:

$$\mathbf{\Sigma} = \sum_{k=1}^{K}\frac{N_k}{N}\mathbf{S_k}$$

Problem 4.11 Solution

Based on definition, we can write down

$$p(\boldsymbol{\phi}|C_k) = \prod_{m=1}^{M}\prod_{l=1}^{L}\mu_{kml}^{\phi_{ml}}$$

Note that here only one of the value among ϕ_{m1}, ϕ_{m2}, ... ϕ_{mL} is 1, and the others are all 0 because we have used a $1-\text{of}-L$ binary coding scheme, and also we have taken advantage of the assumption that the M components of $\boldsymbol{\phi}$ are independent conditioned on the class C_k. We substitute the expression above into (4.63), which gives:

$$a_k = \sum_{m=1}^{M}\sum_{l=1}^{L}\phi_{ml}\mu_{kml} + \ln p(C_k)$$

Hence it is obvious that a_k is a linear function of the components of $\boldsymbol{\phi}$.

Problem 4.12 Solution

Based on definition, i.e., (4.59), we know that logistic sigmoid has the form:

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

Now, we calculate its derivative with regard to a.

$$\frac{d\sigma(a)}{da} = \frac{exp(a)}{[\,1+exp(-a)\,]^2} = \frac{exp(a)}{1+exp(-a)} \cdot \frac{1}{1+exp(-a)} = [\,1-\sigma(a)\,] \cdot \sigma(a)$$

Just as required.

Problem 4.13 Solution

Let's follow the hint.

$$\begin{aligned}
\nabla E(\boldsymbol{w}) &= -\nabla \sum_{n=1}^{N} \{\, t_n \ln y_n + (1-t_n)\ln(1-y_n)\,\} \\
&= -\sum_{n=1}^{N} \nabla \{\, t_n \ln y_n + (1-t_n)\ln(1-y_n)\,\} \\
&= -\sum_{n=1}^{N} \frac{d\{\, t_n \ln y_n + (1-t_n)\ln(1-y_n)\,\}}{dy_n} \frac{dy_n}{da_n} \frac{da_n}{d\mathbf{w}} \\
&= -\sum_{n=1}^{N} (\frac{t_n}{y_n} - \frac{1-t_n}{1-y_n}) \cdot y_n\,(1-y_n) \cdot \boldsymbol{\phi_n} \\
&= -\sum_{n=1}^{N} \frac{t_n - y_n}{y_n(1-y_n)} \cdot y_n\,(1-y_n) \cdot \boldsymbol{\phi_n} \\
&= -\sum_{n=1}^{N} (t_n - y_n)\boldsymbol{\phi_n} \\
&= \sum_{n=1}^{N} (y_n - t_n)\boldsymbol{\phi_n}
\end{aligned}$$

Where we have used $y_n = \sigma(a_n)$, $a_n = \mathbf{w}^T\boldsymbol{\phi_n}$, the chain rules and (4.88).

Problem 4.14 Solution

According to definition, we know that if a dataset is linearly separable, we can find $\mathbf{w}$, for some points $\mathbf{x_n}$, we have $\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x_n}) > 0$, and the others $\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x_m}) < 0$. Then the boundary is given by $\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}) = 0$. Note that for any point $\mathbf{x_0}$ in the dataset, the value of $\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x_0})$ should either be positive or negative, but it can not equal to 0.

Therefore, the maximum likelihood solution for logistic regression is trivial. We suppose for those points $\mathbf{x_n}$ belonging to class C_1, we have $\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x_n}) > 0$ and $\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x_m}) < 0$ for those belonging to class C_2. According to (4.87), if $|\mathbf{w}| \to \infty$, we have

$$p(C_1|\boldsymbol{\phi}(\mathbf{x_n})) = \sigma(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x_n})) \to 1$$

Where we have used $\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x_n}) \to +\infty$. And since $\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x_m}) \to -\infty$, we can also obtain:

$$p(C_2|\boldsymbol{\phi}(\mathbf{x_m})) = 1 - p(C_1|\boldsymbol{\phi}(\mathbf{x_m})) = 1 - \sigma(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x_m})) \to 1$$

In other words, for the likelihood function, i.e.,(4.89), if we have $|\mathbf{w}| \to \infty$, and also we label all the points lying on one side of the boundary as class C_1, and those on the other side as class C_2, the every term in (4.89) can achieve its maximum value, i.e., 1, finally leading to the maximum of the likelihood.

Hence, for a linearly separable dataset, the learning process may prefer to make $|\mathbf{w}| \to \infty$ and use the linear boundary to label the datasets, which can cause severe over-fitting problem.

Problem 4.15 Solution(Waiting for update)

Since y_n is the output of the logistic sigmoid function, we know that $0 < y_n < 1$ and hence $y_n(1-y_n) > 0$. Then we use (4.97), for an arbitrary non-zero real vector $\mathbf{a} \neq \mathbf{0}$, we have:

$$\begin{aligned}
\mathbf{a}^T\mathbf{H}\mathbf{a} &= \mathbf{a}^T\Big[\sum_{n=1}^{N} y_n(1-y_n)\boldsymbol{\phi}_n\boldsymbol{\phi}_n^T\Big]\mathbf{a} \\
&= \sum_{n=1}^{N} y_n(1-y_n)(\boldsymbol{\phi}_n^T\mathbf{a})^T(\boldsymbol{\phi}_n^T\mathbf{a}) \\
&= \sum_{n=1}^{N} y_n(1-y_n)\, b_n^2
\end{aligned}$$

Where we have denoted $b_n = \boldsymbol{\phi}_n^T\mathbf{a}$. What's more, there should be at least one of $\{b_1, b_2, ..., b_N\}$ not equal to zero and then we can see that the expression above is larger than 0 and hence $\mathbf{H}$ is positive definite.

Otherwise, if all the $b_n = 0$, $\mathbf{a} = [a_1, a_2, ..., a_M]^T$ will locate in the null space of matrix $\boldsymbol{\Phi}_{N\times M}$. However, with regard to the *rank-nullity theorem*, we know that Rank($\boldsymbol{\Phi}$) + Nullity($\boldsymbol{\Phi}$) = M, and we have already assumed that those M features are independent, i.e., Rank($\boldsymbol{\Phi}$) = M, which means there is only $\mathbf{0}$ in its null space. Therefore contradictory occurs.

Problem 4.16 Solution

We still denote $y_n = p(t = 1|\boldsymbol{\phi_n})$, and then we can write down the log likelihood by replacing t_n with π_n in (4.89) and (4.90).

$$\ln p(\mathbf{t}|\mathbf{w}) = \sum_{n=1}^{N}\{\pi_n \ln y_n + (1-\pi_n)\ln(1-y_n)\}$$

Problem 4.17 Solution

We should discuss in two situations separately, namely $j = k$ and $j \neq k$. When $j \neq k$, we have:

$$\frac{\partial y_k}{\partial a_j} = \frac{-exp(a_k)\cdot exp(a_j)}{[\sum_j exp(a_j)]^2} = -y_k \cdot y_j$$

And when $j = k$, we have:

$$\frac{\partial y_k}{\partial a_k} = \frac{exp(a_k)\sum_j exp(a_j) - exp(a_k)exp(a_k)}{[\sum_j exp(a_j)]^2} = y_k - y_k^2 = y_k(1-y_k)$$

Therefore, we can obtain:

$$\frac{\partial y_k}{\partial a_j} = y_k(I_{kj} - y_j)$$

Where I_{kj} is the elements of the indentity matrix.

Problem 4.18 Solution

We derive every term $t_{nk} \ln y_{nk}$ with regard to a_j.

$$\begin{aligned}
\frac{\partial t_{nk} \ln y_{nk}}{\partial \mathbf{w_j}} &= \frac{\partial t_{nk} \ln y_{nk}}{\partial y_{nk}} \frac{\partial y_{nk}}{\partial a_j} \frac{\partial a_j}{\partial \mathbf{w_j}} \\
&= t_{nk} \frac{1}{y_{nk}} \cdot y_{nk}(I_{kj} - y_{nj}) \cdot \boldsymbol{\phi_n} \\
&= t_{nk}(I_{kj} - y_{nj})\boldsymbol{\phi_n}
\end{aligned}$$

Where we have used (4.105) and (4.106). Next we perform summation over n and k.

$$\begin{aligned}
\nabla_{\mathbf{w_j}} E &= -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}(I_{kj} - y_{nj})\boldsymbol{\phi_n} \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} y_{nj}\boldsymbol{\phi_n} - \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} I_{kj}\boldsymbol{\phi_n} \\
&= \sum_{n=1}^{N}\Big[(\sum_{k=1}^{K} t_{nk}) y_{nj}\boldsymbol{\phi_n}\Big] - \sum_{n=1}^{N} t_{nj}\boldsymbol{\phi_n} \\
&= \sum_{n=1}^{N} y_{nj}\boldsymbol{\phi_n} - \sum_{n=1}^{N} t_{nj}\boldsymbol{\phi_n} \\
&= \sum_{n=1}^{N}(y_{nj} - t_{nj})\boldsymbol{\phi_n}
\end{aligned}$$

Where we have used the fact that for arbitrary n, we have $\sum_{k=1}^{K} t_{nk} = 1$.

Problem 4.19 Solution

We write down the log likelihood.

$$\ln p(\mathbf{t}|\mathbf{w}) = \sum_{n=1}^{N}\left\{t_n \ln y_n + (1 - t_n)\ln(1 - y_n)\right\}$$

Therefore, we can obtain:

$$\begin{aligned}
\nabla_{\mathbf{w}} \ln p &= \frac{\partial \ln p}{\partial y_n} \cdot \frac{\partial y_n}{\partial a_n} \cdot \frac{\partial a_n}{\partial \mathbf{w}} \\
&= \sum_{n=1}^{N}(\frac{t_n}{y_n} - \frac{1 - t_n}{1 - y_n})\Phi'(a_n)\boldsymbol{\phi_n} \\
&= \sum_{n=1}^{N}\frac{y_n - t_n}{y_n(1 - y_n)}\Phi'(a_n)\boldsymbol{\phi_n}
\end{aligned}$$

Where we have used $y = p(t = 1|a) = \Phi(a)$ and $a_n = \mathbf{w}^T \boldsymbol{\phi_n}$. According to (4.114), we can obtain:

$$\Phi'(a) = \mathcal{N}(\theta|0,1)\big|_{\theta=a} = \frac{1}{\sqrt{2\pi}} exp(-\frac{1}{2}a^2)$$

Hence, we can obtain:

$$\nabla_{\mathbf{w}} \ln p = \sum_{n=1}^{N} \frac{y_n - t_n}{y_n(1-y_n)} \frac{exp(-\frac{a_n^2}{2})}{\sqrt{2\pi}} \boldsymbol{\phi_n}$$

To calculate the Hessian Matrix, we need to first evaluate several derivatives.

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{w}}\{\frac{y_n - t_n}{y_n(1-y_n)}\} &= \frac{\partial}{\partial y_n}\{\frac{y_n - t_n}{y_n(1-y_n)}\} \cdot \frac{\partial y_n}{\partial a_n} \cdot \frac{\partial a_n}{\partial \mathbf{w}} \\
&= \frac{y_n(1-y_n) - (y_n - t_n)(1-2y_n)}{[y_n(1-y_n)]^2} \Phi'(a_n) \boldsymbol{\phi_n} \\
&= \frac{y_n^2 + t_n - 2y_n t_n}{y_n^2(1-y_n)^2} \frac{exp(-\frac{a_n^2}{2})}{\sqrt{2\pi}} \boldsymbol{\phi_n}
\end{aligned}$$

And

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{w}}\{\frac{exp(-\frac{a_n^2}{2})}{\sqrt{2\pi}}\} &= \frac{\partial}{\partial a_n}\{\frac{exp(-\frac{a_n^2}{2})}{\sqrt{2\pi}}\} \frac{\partial a_n}{\partial \mathbf{w}} \\
&= -\frac{a_n}{\sqrt{2\pi}} exp(-\frac{a_n^2}{2}) \boldsymbol{\phi_n}
\end{aligned}$$

Therefore, using the chain rule, we can obtain:

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{w}}\{\frac{y_n - t_n}{y_n(1-y_n)} \frac{exp(-\frac{a_n^2}{2})}{\sqrt{2\pi}}\} &= \frac{\partial}{\partial \mathbf{w}}\{\frac{y_n - t_n}{y_n(1-y_n)}\} \frac{exp(-\frac{a_n^2}{2})}{\sqrt{2\pi}} + \frac{y_n - t_n}{y_n(1-y_n)} \frac{\partial}{\partial \mathbf{w}}\{\frac{exp(-\frac{a_n^2}{2})}{\sqrt{2\pi}}\} \\
&= \big[\frac{y_n^2 + t_n - 2y_n t_n}{y_n(1-y_n)} \frac{exp(-\frac{a_n^2}{2})}{\sqrt{2\pi}} - a_n(y_n - t_n)\big] \frac{exp(-\frac{a_n^2}{2})}{\sqrt{2\pi} y_n(1-y_n)} \boldsymbol{\phi_n}
\end{aligned}$$

Finally if we perform summation over n, we can obtain the Hessian Matrix:

$$\begin{aligned}
\mathbf{H} &= \nabla\nabla_{\mathbf{w}} \ln p \\
&= \sum_{n=1}^{N} \frac{\partial}{\partial \mathbf{w}}\{\frac{y_n - t_n}{y_n(1-y_n)} \frac{exp(-\frac{a_n^2}{2})}{\sqrt{2\pi}}\} \cdot \boldsymbol{\phi_n} \\
&= \sum_{n=1}^{N} \big[\frac{y_n^2 + t_n - 2y_n t_n}{y_n(1-y_n)} \frac{exp(-\frac{a_n^2}{2})}{\sqrt{2\pi}} - a_n(y_n - t_n)\big] \frac{exp(-\frac{a_n^2}{2})}{\sqrt{2\pi} y_n(1-y_n)} \boldsymbol{\phi_n} \boldsymbol{\phi_n}^T
\end{aligned}$$

Problem 4.20 Solution(waiting for update)

We know that the Hessian Matrix is of size $MK \times MK$, and the (j,k)th block with size $M \times M$ is given by (4.110), where $j,k = 1,2,...,K$. Therefore, we can obtain:

$$\mathbf{u^T H u} = \sum_{j=1}^{K}\sum_{k=1}^{K} \mathbf{u_j^T H_{j,k} u_k} \qquad (*)$$

Where we use $\mathbf{u_k}$ to denote the kth block vector of $\mathbf{u}$ with size $M \times 1$, and $\mathbf{H_{j,k}}$ to denote the (j,k)th block matrix of $\mathbf{H}$ with size $M \times M$. Then based on (4.110), we further expand (4.110):

$$\begin{aligned}
(*) &= \sum_{j=1}^{K}\sum_{k=1}^{K} \mathbf{u_j^T}\{-\sum_{n=1}^{N} y_{nk}(I_{kj} - y_{nj})\boldsymbol{\phi_n}\boldsymbol{\phi_n}^T\}\mathbf{u_k} \\
&= \sum_{j=1}^{K}\sum_{k=1}^{K}\sum_{n=1}^{N} \mathbf{u_j^T}\{-y_{nk}(I_{kj} - y_{nj})\boldsymbol{\phi_n}\boldsymbol{\phi_n}^T\}\mathbf{u_k} \\
&= \sum_{j=1}^{K}\sum_{k=1}^{K}\sum_{n=1}^{N} \mathbf{u_j^T}\{-y_{nk}I_{kj}\boldsymbol{\phi_n}\boldsymbol{\phi_n}^T\}\mathbf{u_k} + \sum_{j=1}^{K}\sum_{k=1}^{K}\sum_{n=1}^{N} \mathbf{u_j^T}\{y_{nk}y_{nj}\boldsymbol{\phi_n}\boldsymbol{\phi_n}^T\}\mathbf{u_k} \\
&= \sum_{k=1}^{K}\sum_{n=1}^{N} \mathbf{u_k^T}\{-y_{nk}\boldsymbol{\phi_n}\boldsymbol{\phi_n}^T\}\mathbf{u_k} + \sum_{j=1}^{K}\sum_{k=1}^{K}\sum_{n=1}^{N} y_{nj}\mathbf{u_j^T}\{\boldsymbol{\phi_n}\boldsymbol{\phi_n}^T\}y_{nk}\mathbf{u_k}
\end{aligned}$$

Problem 4.21 Solution

It is quite obvious.

$$\begin{aligned}
\Phi(a) &= \int_{-\infty}^{a} \mathcal{N}(\theta|0,1)d\theta \\
&= \frac{1}{2} + \int_{0}^{a} \mathcal{N}(\theta|0,1)d\theta \\
&= \frac{1}{2} + \int_{0}^{a} \mathcal{N}(\theta|0,1)d\theta \\
&= \frac{1}{2} + \frac{1}{\sqrt{2\pi}}\int_{0}^{a} exp(-\theta^2/2)d\theta \\
&= \frac{1}{2} + \frac{1}{\sqrt{2\pi}}\frac{\sqrt{\pi}}{2}\int_{0}^{a} \frac{2}{\sqrt{\pi}}exp(-\theta^2/2)d\theta \\
&= \frac{1}{2}(1 + \frac{1}{\sqrt{2}}\int_{0}^{a} \frac{2}{\sqrt{\pi}}exp(-\theta^2/2)d\theta) \\
&= \frac{1}{2}\{1 + \frac{1}{\sqrt{2}}erf(a)\}
\end{aligned}$$

Where we have used

$$\int_{-\infty}^{0} \mathcal{N}(\theta|0,1)d\theta = \frac{1}{2}$$

Problem 4.22 Solution

If we denote $f(\boldsymbol{\theta}) = p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})$, we can write:

$$\begin{aligned} p(D) &= \int p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} = \int f(\boldsymbol{\theta})d\boldsymbol{\theta} \\ &= f(\boldsymbol{\theta}_{MAP})\frac{(2\pi)^{M/2}}{|\mathbf{A}|^{1/2}} \\ &= p(D|\boldsymbol{\theta}_{MAP})p(\boldsymbol{\theta}_{MAP})\frac{(2\pi)^{M/2}}{|\mathbf{A}|^{1/2}} \end{aligned}$$

Where $\boldsymbol{\theta}_{MAP}$ is the value of $\boldsymbol{\theta}$ at the mode of $f(\boldsymbol{\theta})$, $\mathbf{A}$ is the Hessian Matrix of $-\ln f(\boldsymbol{\theta})$ and we have also used (4.135). Therefore,

$$\ln p(D) = \ln p(D|\boldsymbol{\theta}_{MAP}) + \ln p(\boldsymbol{\theta}_{MAP}) + \frac{M}{2}\ln 2\pi - \frac{1}{2}\ln|\mathbf{A}|$$

Just as required.

Problem 4.23 Solution

According to (4.137), we can write:

$$\begin{aligned} \ln p(D) &= \ln p(D|\boldsymbol{\theta}_{MAP}) + \ln p(\boldsymbol{\theta}_{MAP}) + \frac{M}{2}\ln 2\pi - \frac{1}{2}\ln|\mathbf{A}| \\ &= \ln p(D|\boldsymbol{\theta}_{MAP}) - \frac{M}{2}\ln 2\pi - \frac{1}{2}\ln|\mathbf{V_0}| - \frac{1}{2}(\boldsymbol{\theta_{MAP}} - \mathbf{m})^T\mathbf{V_0}^{-1}(\boldsymbol{\theta_{MAP}} - \mathbf{m}) \\ &\quad + \frac{M}{2}\ln 2\pi - \frac{1}{2}\ln|\mathbf{A}| \\ &= \ln p(D|\boldsymbol{\theta}_{MAP}) - \frac{1}{2}\ln|\mathbf{V_0}| - \frac{1}{2}(\boldsymbol{\theta_{MAP}} - \mathbf{m})^T\mathbf{V_0}^{-1}(\boldsymbol{\theta_{MAP}} - \mathbf{m}) - \frac{1}{2}\ln|\mathbf{A}| \end{aligned}$$

Where we have used the definition of the multivariate Gaussian Distribution. Then, from (4.138), we can write:

$$\begin{aligned} \mathbf{A} &= -\nabla\nabla\ln p(D|\boldsymbol{\theta_{MAP}})p(\boldsymbol{\theta_{MAP}}) \\ &= -\nabla\nabla\ln p(D|\boldsymbol{\theta_{MAP}}) - \nabla\nabla\ln p(\boldsymbol{\theta_{MAP}}) \\ &= \mathbf{H} - \nabla\nabla\Big\{-\frac{1}{2}(\boldsymbol{\theta_{MAP}} - \mathbf{m})^T\mathbf{V_0}^{-1}(\boldsymbol{\theta_{MAP}} - \mathbf{m})\Big\} \\ &= \mathbf{H} + \nabla\big\{\mathbf{V_0}^{-1}(\boldsymbol{\theta_{MAP}} - \mathbf{m})\big\} \\ &= \mathbf{H} + \mathbf{V_0}^{-1} \end{aligned}$$

Where we have denoted $\mathbf{H} = -\nabla\nabla\ln p(D|\boldsymbol{\theta_{MAP}})$. Therefore, the equation

above becomes:

$$\begin{aligned}
\ln p(D) &= \ln p(D|\boldsymbol{\theta}_{MAP}) - \frac{1}{2}(\boldsymbol{\theta_{MAP}} - \mathbf{m})^T \mathbf{V_0}^{-1}(\boldsymbol{\theta_{MAP}} - \mathbf{m}) - \frac{1}{2}\ln\{|\mathbf{V_0}| \cdot |\mathbf{H} + \mathbf{V_0^{-1}}|\} \\
&= \ln p(D|\boldsymbol{\theta}_{MAP}) - \frac{1}{2}(\boldsymbol{\theta_{MAP}} - \mathbf{m})^T \mathbf{V_0}^{-1}(\boldsymbol{\theta_{MAP}} - \mathbf{m}) - \frac{1}{2}\ln\{|\mathbf{V_0 H} + \mathbf{I}|\} \\
&\approx \ln p(D|\boldsymbol{\theta}_{MAP}) - \frac{1}{2}(\boldsymbol{\theta_{MAP}} - \mathbf{m})^T \mathbf{V_0}^{-1}(\boldsymbol{\theta_{MAP}} - \mathbf{m}) - \frac{1}{2}\ln|\mathbf{V_0}| - \frac{1}{2}\ln|\mathbf{H}| \\
&\approx \ln p(D|\boldsymbol{\theta}_{MAP}) - \frac{1}{2}(\boldsymbol{\theta_{MAP}} - \mathbf{m})^T \mathbf{V_0}^{-1}(\boldsymbol{\theta_{MAP}} - \mathbf{m}) - \frac{1}{2}\ln|\mathbf{H}| + \text{const}
\end{aligned}$$

Where we have used the property of determinant: $|\mathbf{A}| \cdot |\mathbf{B}| = |\mathbf{AB}|$, and the fact that the prior is board, i.e. $\mathbf{I}$ can be neglected with regard to $\mathbf{V_0 H}$. What's more, since the prior is pre-given, we can view $\mathbf{V_0}$ as constant. And if the data is large, we can write:

$$\mathbf{H} = \sum_{n=1}^{N} \mathbf{H_n} = N\widehat{\mathbf{H}}$$

Where $\widehat{\mathbf{H}} = 1/N \sum_{n=1}^{N} \mathbf{H_n}$, and then

$$\begin{aligned}
\ln p(D) &\approx \ln p(D|\boldsymbol{\theta}_{MAP}) - \frac{1}{2}(\boldsymbol{\theta_{MAP}} - \mathbf{m})^T \mathbf{V_0}^{-1}(\boldsymbol{\theta_{MAP}} - \mathbf{m}) - \frac{1}{2}\ln|\mathbf{H}| + \text{const} \\
&\approx \ln p(D|\boldsymbol{\theta}_{MAP}) - \frac{1}{2}(\boldsymbol{\theta_{MAP}} - \mathbf{m})^T \mathbf{V_0}^{-1}(\boldsymbol{\theta_{MAP}} - \mathbf{m}) - \frac{1}{2}\ln|N\widehat{\mathbf{H}}| + \text{const} \\
&\approx \ln p(D|\boldsymbol{\theta}_{MAP}) - \frac{1}{2}(\boldsymbol{\theta_{MAP}} - \mathbf{m})^T \mathbf{V_0}^{-1}(\boldsymbol{\theta_{MAP}} - \mathbf{m}) - \frac{M}{2}\ln N - \frac{1}{2}\ln|\widehat{\mathbf{H}}| + \text{const} \\
&\approx \ln p(D|\boldsymbol{\theta}_{MAP}) - \frac{M}{2}\ln N
\end{aligned}$$

This is because when $N >> 1$, other terms can be neglected.

Problem 4.24 Solution(Waiting for updating)

Problem 4.25 Solution

We first need to obtain the expression for the first derivative of probit function $\Phi(\lambda a)$ with regard to a. According to (4.114), we can write down:

$$\begin{aligned}
\frac{d}{da}\Phi(\lambda a) &= \frac{d\Phi(\lambda a)}{d(\lambda a)} \cdot \frac{d\lambda a}{da} \\
&= \frac{\lambda}{\sqrt{2\pi}} exp\{-\frac{1}{2}(\lambda a)^2\}
\end{aligned}$$

Which further gives:

$$\frac{d}{da}\Phi(\lambda a)\Big|_{a=0} = \frac{\lambda}{\sqrt{2\pi}}$$

And for logistic sigmoid function, according to (4.88), we have

$$\frac{d\sigma}{da} = \sigma(1-\sigma) = 0.5 \times 0.5 = \frac{1}{4}$$

Where we have used $\sigma(0) = 0.5$. Let their derivatives at origin equals, we have:

$$\frac{\lambda}{\sqrt{2\pi}} = \frac{1}{4}$$

i.e., $\lambda = \sqrt{2\pi}/4$. And hence $\lambda^2 = \pi/8$ is obvious.

Problem 4.26 Solution

We will prove (4.152) in a more simple and intuitive way. But firstly, we need to prove a trivial yet useful statement: Suppose we have a random variable satisfied normal distribution denoted as $X \sim \mathcal{N}(X|\mu,\sigma^2)$, the probability of $X \leq x$ is $P(X \leq x) = \Phi(\frac{x-\mu}{\sigma})$, and here x is a given real number. We can see this by writing down the integral:

$$\begin{aligned} P(X \leq x) &= \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} exp\big[-\frac{1}{2\sigma^2}(X-\mu)^2\big] dX \\ &= \int_{-\infty}^{\frac{x-\mu}{\sigma}} \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{1}{2}\gamma^2)\sigma\, d\gamma \\ &= \int_{-\infty}^{\frac{x-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} exp(-\frac{1}{2}\gamma^2) d\gamma \\ &= \Phi(\frac{x-\mu}{\sigma}) \end{aligned}$$

Where we have changed the variable $X = \mu + \sigma\gamma$. Now consider two random variables $X \sim \mathcal{N}(0,\lambda^{-2})$ and $Y \sim \mathcal{N}(\mu,\sigma^2)$. We first calculate the conditional probability $P(X \leq Y \,|\, Y = a)$:

$$P(X \leq Y \,|\, Y = a) = P(X \leq a) = \Phi(\frac{a-0}{\lambda^{-1}}) = \Phi(\lambda a)$$

Together with Bayesian Formula, we can obtain:

$$\begin{aligned} P(X \leq Y) &= \int_{-\infty}^{+\infty} P(X \leq Y \,|\, Y = a)\, pdf(Y = a) dY \\ &= \int_{-\infty}^{+\infty} \Phi(\lambda a)\mathcal{N}(a|\mu,\sigma^2) da \end{aligned}$$

Where $pdf(\cdot)$ denotes the probability density function and we have also used $pdf(Y) = \mathcal{N}(\mu,\sigma^2)$. What's more, we know that $X - Y$ should also satisfy normal distribution, with:

$$E[X-Y] = E[X] - E[Y] = 0 - \mu = -\mu$$

And

$$var[X-Y] = var[X] + var[Y] = \lambda^{-2} + \sigma^2$$

Therefore, $X - Y \sim \mathcal{N}(-\mu, \lambda^{-2} + \sigma^2)$ and it follows that:

$$P(X - Y \leq 0) = \Phi(\frac{0-(-\mu)}{\sqrt{\lambda^{-2}+\sigma^2}}) = \Phi(\frac{\mu}{\sqrt{\lambda^{-2}+\sigma^2}})$$

Since $P(X \leq Y) = P(X - Y \leq 0)$, we obtain what have been required.

0.5 Neural Networks

Problem 5.1 Solution

Based on definition of $tanh(\cdot)$, we can obtain:

$$\begin{aligned} tanh(a) &= \frac{e^a - e^{-a}}{e^a + e^{-a}} \\ &= -1 + \frac{2e^a}{e^a + e^{-a}} \\ &= -1 + 2\frac{1}{1 + e^{-2a}} \\ &= 2\sigma(2a) - 1 \end{aligned}$$

If we have parameters $w_{ji}^{(1s)}, w_{j0}^{(1s)}$ and $w_{kj}^{(2s)}, w_{k0}^{(2s)}$ for a network whose hidden units use logistic sigmoid function as activation and $w_{ji}^{(1t)}, w_{j0}^{(1t)}$ and $w_{kj}^{(2t)}, w_{k0}^{(2t)}$ for another one using $tanh(\cdot)$, for the network using $tanh(\cdot)$ as activation, we can write down the following expression by using (5.4):

$$\begin{aligned} a_k^{(t)} &= \sum_{j=1}^{M} w_{kj}^{(2t)} tanh(a_j^{(t)}) + w_{k0}^{(2t)} \\ &= \sum_{j=1}^{M} w_{kj}^{(2t)} [2\sigma(2a_j^{(t)}) - 1] + w_{k0}^{(2t)} \\ &= \sum_{j=1}^{M} 2w_{kj}^{(2t)} \sigma(2a_j^{(t)}) + \big[-\sum_{j=1}^{M} w_{kj}^{(2t)} + w_{k0}^{(2t)} \big] \end{aligned}$$

What's more, we also have :

$$a_k^{(s)} = \sum_{j=1}^{M} w_{kj}^{(2s)} \sigma(a_j^{(s)}) + w_{k0}^{(2s)}$$

To make the two networks equivalent, i.e., $a_k^{(s)} = a_k^{(t)}$, we should make sure:

$$\begin{cases} a_j^{(s)} = 2a_j^{(t)} \\ w_{kj}^{(2s)} = 2w_{kj}^{(2t)} \\ w_{k0}^{(2s)} = -\sum_{j=1}^{M} w_{kj}^{(2t)} + w_{k0}^{(2t)} \end{cases}$$

Note that the first condition can be achieved by simply enforcing:

$$w_{ji}^{(1s)} = 2w_{ji}^{(1t)}, \quad \text{and} \quad w_{j0}^{(1s)} = 2w_{j0}^{(1t)}$$

Therefore, these two networks are equivalent under a linear transformation.

Problem 5.2 Solution

It is obvious. We write down the likelihood.

$$p(\mathbf{T}|\mathbf{X},\mathbf{w}) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{t_n}|\mathbf{y}(\mathbf{x_n},\mathbf{w}),\beta^{-1}\mathbf{I})$$

Taking the negative logarithm, we can obtain:

$$E(\mathbf{w},\beta) = -\ln p(\mathbf{T}|\mathbf{X},\mathbf{w}) = \frac{\beta}{2}\sum_{n=1}^{N}\left[(\mathbf{y}(\mathbf{x_n},\mathbf{w})-\mathbf{t_n})^T(\mathbf{y}(\mathbf{x_n},\mathbf{w})-\mathbf{t_n})\right] - \frac{NK}{2}\ln\beta + \text{const}$$

Here we have used const to denote the term independent of both $\mathbf{w}$ and β. Note that here we have used the definition of the multivariate Gaussian Distribution. What's more, we see that the covariance matrix $\beta^{-1}\mathbf{I}$ and the weight parameter $\mathbf{w}$ have decoupled, which is distinct from the next problem. We can first solve $\mathbf{w_{ML}}$ by minimizing the first term on the right of the equation above or equivalently (5.11), i.e., imaging β is fixed. Then according to the derivative of $E(\mathbf{w},\beta)$ with regard to β, we can obtain (5.17) and hence β_{ML}.

Problem 5.3 Solution

Following the process in the previous question, we first write down the negative logarithm of the likelihood function.

$$E(\mathbf{w},\mathbf{\Sigma}) = \frac{1}{2}\sum_{n=1}^{N}\left\{[\mathbf{y}(\mathbf{x_n},\mathbf{w})-\mathbf{t_n}]^T\mathbf{\Sigma}^{-1}[\mathbf{y}(\mathbf{x_n},\mathbf{w})-\mathbf{t_n}]\right\} + \frac{N}{2}\ln|\mathbf{\Sigma}| + \text{const} \quad (*)$$

Note here we have assumed $\mathbf{\Sigma}$ is unknown and const denotes the term independent of both $\mathbf{w}$ and $\mathbf{\Sigma}$. In the first situation, if $\mathbf{\Sigma}$ is fixed and known, the equation above will reduce to:

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\left\{[\mathbf{y}(\mathbf{x_n},\mathbf{w})-\mathbf{t_n}]^T\mathbf{\Sigma}^{-1}[\mathbf{y}(\mathbf{x_n},\mathbf{w})-\mathbf{t_n}]\right\} + \text{const}$$

We can simply solve $\mathbf{w_{ML}}$ by minimizing it. If $\mathbf{\Sigma}$ is unknown, since Σ is in the first term on the right of $(*)$, solving $\mathbf{w_{ML}}$ will involve $\mathbf{\Sigma}$. Note that in the previous problem, the main reason that they can decouple is due to the independent assumption, i.e., $\mathbf{\Sigma}$ reduces to $\beta^{-1}\mathbf{I}$, so that we can bring β to the front and view it as a fixed multiplying factor when solving $\mathbf{w_{ML}}$.

Problem 5.4 Solution

Based on (5.20), the current conditional distribution of targets, considering mislabel, given input $\mathbf{x}$ and weight $\mathbf{w}$ is:

$$p(t=1|\mathbf{x},\mathbf{w}) = (1-\epsilon)\cdot p(t_r=1|\mathbf{x},\mathbf{w}) + \epsilon\cdot p(t_r=0|\mathbf{x},\mathbf{w})$$

Note that here we use t to denote the observed target label, t_r to denote its real label, and that our network is aimed to predict the real label t_r not t, i.e., $p(t_r = 1|\mathbf{x}, \mathbf{w}) = y(\mathbf{x}, \mathbf{w})$, hence we see that:

$$p(t = 1|\mathbf{x}, \mathbf{w}) = (1 - \epsilon) \cdot y(\mathbf{x}, \mathbf{w}) + \epsilon \cdot \big[1 - y(\mathbf{x}, \mathbf{w})\big] \qquad (*)$$

Also, it is the same for $p(t = 0|\mathbf{x}, \mathbf{w})$:

$$p(t = 0|\mathbf{x}, \mathbf{w}) = (1 - \epsilon) \cdot \big[1 - y(\mathbf{x}, \mathbf{w})\big] + \epsilon \cdot y(\mathbf{x}, \mathbf{w}) \qquad (**)$$

Combing $(*)$ and $(**)$, we can obtain:

$$p(t|\mathbf{x}, \mathbf{w}) = (1 - \epsilon) \cdot y^t (1 - y)^{1-t} + \epsilon \cdot (1 - y)^t y^{1-t}$$

Where y is short for $y(\mathbf{x}, \mathbf{w})$. Therefore, taking the negative logarithm, we can obtain the error function:

$$E(\mathbf{w}) = -\sum_{n=1}^{N} \ln \big\{(1 - \epsilon) \cdot y_n^{t_n} (1 - y_n)^{1-t_n} + \epsilon \cdot (1 - y_n)^{t_n} y_n^{1-t_n}\big\}$$

When $\epsilon = 0$, it is obvious that the equation above will reduce to (5.21).

Problem 5.5 Solution

It is obvious by using (5.22).

$$\begin{aligned}
E(\mathbf{w}) &= -\ln \prod_{n=1}^{N} p(\mathbf{t}|\mathbf{x_n}, \mathbf{w}) \\
&= -\ln \prod_{n=1}^{N} \prod_{k=1}^{K} y_k(\mathbf{x_n}, \mathbf{w})^{t_{nk}} \big[1 - y_k(\mathbf{x_n}, \mathbf{w})\big]^{1-t_{nk}} \\
&= -\sum_{n=1}^{N} \sum_{k=1}^{K} \ln \big\{y_k(\mathbf{x_n}, \mathbf{w})^{t_{nk}} \big[1 - y_k(\mathbf{x_n}, \mathbf{w})\big]^{1-t_{nk}}\big\} \\
&= -\sum_{n=1}^{N} \sum_{k=1}^{K} \ln \big[y_{nk}^{t_{nk}} (1 - y_{nk})^{1-t_{nk}}\big] \\
&= -\sum_{n=1}^{N} \sum_{k=1}^{K} \big\{t_{nk} \ln y_{nk} + (1 - t_{nk}) \ln(1 - y_{nk})\big\}
\end{aligned}$$

Where we have denoted

$$y_{nk} = y_k(\mathbf{x_n}, \mathbf{w})$$

Problem 5.6 Solution

We know that $y_k = \sigma(a_k)$, where $\sigma(\cdot)$ represents the logistic sigmoid function. Moreover,

$$\frac{d\sigma}{da} = \sigma(1 - \sigma)$$

$$\begin{aligned}\frac{dE(\mathbf{w})}{da_k} &= -t_k\frac{1}{y_k}\big[y_k(1-y_k)\big]+(1-t_k)\frac{1}{1-y_k}\big[y_k(1-y_k)\big]\\ &= \big[y_k(1-y_k)\big]\big[\frac{1-t_k}{1-y_k}-\frac{t_k}{y_k}\big]\\ &= (1-t_k)y_k-t_k(1-y_k)\\ &= y_k-t_k\end{aligned}$$

Just as required.

Problem 5.7 Solution

It is similar to the previous problem. First we denote $y_{kn}=y_k(\mathbf{x_n},\mathbf{w})$. If we use softmax function as activation for the output unit, according to (4.106), we have:

$$\frac{dy_{kn}}{da_j}=y_{kn}(I_{kj}-y_{jn})$$

Therefore,

$$\begin{aligned}\frac{dE(\mathbf{w})}{da_j} &= \frac{d}{da_k}\{-\sum_{n=1}^{N}\sum_{k=1}^{K}t_{kn}\ln y_k(\mathbf{x_n},\mathbf{w})\}\\ &= -\sum_{n=1}^{N}\sum_{k=1}^{K}\frac{d}{da_j}\{t_{kn}\ln y_{kn}\}\\ &= -\sum_{n=1}^{N}\sum_{k=1}^{K}t_{kn}\frac{1}{y_{kn}}\big[y_{kn}(I_{kj}-y_{jn})\big]\\ &= -\sum_{n=1}^{N}\sum_{k=1}^{K}(t_{kn}I_{kj}-t_{kn}y_{jn})\\ &= -\sum_{n=1}^{N}\sum_{k=1}^{K}t_{kn}I_{kj}+\sum_{n=1}^{N}\sum_{k=1}^{K}t_{kn}y_{jn}\\ &= -\sum_{n=1}^{N}t_{jn}+\sum_{n=1}^{N}y_{jn}\\ &= \sum_{n=1}^{N}(y_{jn}-t_{jn})\end{aligned}$$

Where we have used the fact that only when $k=j$, $I_{kj}=1\neq 0$ and that $\sum_{k=1}^{K}t_{kn}=1$.

Problem 5.8 Solution

It is obvious based on definition of 'tanh', i.e., (5.59).

$$\begin{aligned}\frac{d}{da}tanh(a) &= \frac{(e^a+e^{-a})(e^a+e^{-a})-(e^a-e^{-a})(e^a-e^{-a})}{(e^a+e^{-a})^2}\\ &= 1-\frac{(e^a-e^{-a})^2}{(e^a+e^{-a})^2}\\ &= 1-tanh(a)^2\end{aligned}$$

Problem 5.9 Solution

We know that the logistic sigmoid function $\sigma(a) \in [0,1]$, therefore if we perform a linear transformation $h(a) = 2\sigma(a)-1$, we can find a mapping function $h(a)$ from $(-\infty,+\infty)$ to $[-1,1]$. In this case, the conditional distribution of targets given inputs can be similarly written as:

$$p(t|\mathbf{x},\mathbf{w}) = \big[\frac{1+y(\mathbf{x},\mathbf{w})}{2}\big]^{(1+t)/2}\big[\frac{1-y(\mathbf{x},\mathbf{w})}{2}\big]^{(1-t)/2}$$

Where $\big[1+y(\mathbf{x},\mathbf{w})\big]/2$ represents the conditional probability $p(C_1|x)$. Since now $y(\mathbf{x},\mathbf{w}) \in [-1,1]$, we also need to perform the linear transformation to make it satisfy the constraint for probability.Then we can further obtain:

$$\begin{aligned} E(\mathbf{w}) &= -\sum_{n=1}^{N}\{\frac{1+t_n}{2}\ln\frac{1+y_n}{2} + \frac{1-t_n}{2}\ln\frac{1-y_n}{2}\} \\ &= -\frac{1}{2}\sum_{n=1}^{N}\big\{(1+t_n)\ln(1+y_n) + (1-t_n)\ln(1-y_n)\big\} + N\ln 2 \end{aligned}$$

Problem 5.10 Solution

It is obvious. Suppose $\mathbf{H}$ is positive definite, i.e., (5.37) holds. We set $\mathbf{v}$ equals to the eigenvector of $\mathbf{H}$, i.e., $\mathbf{v} = \mathbf{u_i}$ which gives:

$$\mathbf{v}^T\mathbf{H}\mathbf{v} = \mathbf{v}^T(\mathbf{H}\mathbf{v}) = \mathbf{u_i}^T\lambda_i\mathbf{u_i} = \lambda_i||\mathbf{u_i}||^2$$

Therefore, every λ_i should be positive. On the other hand, If all the eigenvalues λ_i are positive, from (5.38) and (5.39), we see that $\mathbf{H}$ is positive definite.

Problem 5.11 Solution

It is obvious. We follow (5.35) and then write the error function in the form of (5.36). To obtain the contour, we enforce $E(\mathbf{w})$ to equal to a constant C.

$$E(\mathbf{w}) = E(\mathbf{w}^*) + \frac{1}{2}\sum_i \lambda_i\alpha_i^2 = C$$

We rearrange the equation above, and then obtain:

$$\sum_i \lambda_i\alpha_i^2 = B$$

Where $B = 2C - 2E(\mathbf{w}^*)$ is a constant. Therefore, the contours of constant error are ellipses whose axes are aligned with the eigenvector $\mathbf{u_i}$ of the Hessian Matrix $\mathbf{H}$. The length for the jth axis is given by setting all $\alpha_i = 0, s.t. i \neq j$:

$$\alpha_j = \sqrt{\frac{B}{\lambda_j}}$$

In other words, the length is inversely proportional to the square root of the corresponding eigenvalue λ_j.

Problem 5.12 Solution

If $\mathbf{H}$ is positive definite, we know the second term on the right side of (5.32) will be positive for arbitrary $\mathbf{w}$. Therefore, $E(\mathbf{w}^*)$ is a local minimum. On the other hand, if $\mathbf{w}^*$ is a local minimum, we have

$$E(\mathbf{w}^*) - E(\mathbf{w}) = -\frac{1}{2}(\mathbf{w}-\mathbf{w}^*)^T\mathbf{H}(\mathbf{w}-\mathbf{w}^*) < 0$$

In other words, for arbitrary $\mathbf{w}$, $(\mathbf{w}-\mathbf{w}^*)^T\mathbf{H}(\mathbf{w}-\mathbf{w}^*) > 0$, according to the previous problem, we know that this means $\mathbf{H}$ is positive definite.

Problem 5.13 Solution

It is obvious. Suppose that there are W adaptive parameters in the network. Therefore, $\mathbf{b}$ has W independent parameters. Since $\mathbf{H}$ is symmetric, there should be $W(W+1)/2$ independent parameters in it. Therefore, there are $W + W(W+1)/2 = W(W+3)/2$ parameters in total.

Problem 5.14 Solution

It is obvious. Since we have

$$E_n(w_{ji}+\epsilon) = E_n(w_{ji}) + \epsilon E_n'(w_{ji}) + \frac{\epsilon^2}{2}E_n''(w_{ji}) + O(\epsilon^3)$$

And

$$E_n(w_{ji}-\epsilon) = E_n(w_{ji}) - \epsilon E_n'(w_{ji}) + \frac{\epsilon^2}{2}E_n''(w_{ji}) + O(\epsilon^3)$$

We combine those two equations, which gives,

$$E_n(w_{ji}+\epsilon) - E_n(w_{ji}-\epsilon) = 2\epsilon E_n'(w_{ji}) + O(\epsilon^3)$$

Rearrange the equation above, we obtain what has been required.

Problem 5.15 Solution

It is obvious. The back propagation formalism starts from performing summation near the input, as shown in (5.73). By symmetry, the forward propagation formalism should start near the output.

$$J_{ki} = \frac{\partial y_k}{\partial x_i} = \frac{\partial h(a_k)}{\partial x_i} = h'(a_k)\frac{\partial a_k}{\partial x_i} \qquad (*)$$

Where $h(\cdot)$ is the activation function at the output node a_k. Considering all the units j, which have links to unit k:

$$\frac{\partial a_k}{\partial x_i} = \sum_j \frac{\partial a_k}{\partial a_j}\frac{\partial a_j}{\partial x_i} = \sum_j w_{kj}h'(a_j)\frac{\partial a_j}{\partial x_i} \qquad (**)$$

Where we have used:

$$a_k = \sum_j w_{kj} z_j, \quad z_j = h(a_j)$$

It is similar for $\partial a_j / \partial x_i$. In this way we have obtained a recursive formula starting from the input node:

$$\frac{\partial a_l}{\partial x_i} = \begin{cases} w_{li}, \text{if there is a link from input unit } i \text{ to } l \\ 0, \text{if there isn't a link from input unit } i \text{ to } l \end{cases}$$

Using recursive formula $(**)$ and then $(*)$, we can obtain the Jacobian Matrix.

Problem 5.16 Solution

It is obvious. We begin by writing down the error function.

$$E = \frac{1}{2} \sum_{n=1}^{N} ||\mathbf{y_n} - \mathbf{t_n}||^2 = \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{M} (y_{n,m} - t_{n,m})^2$$

Where the subscript m denotes the $m$the element of the vector. Then we can write down the Hessian Matrix as before.

$$\mathbf{H} = \nabla\nabla E = \sum_{n=1}^{N} \sum_{m=1}^{M} \nabla \mathbf{y_{n,m}} \nabla \mathbf{y_{n,m}} + \sum_{n=1}^{N} \sum_{m=1}^{M} (y_{n,m} - t_{n,m}) \nabla\nabla \mathbf{y_{n,m}}$$

Similarly, we now know that the Hessian Matrix can be approximated as:

$$\mathbf{H} \simeq \sum_{n=1}^{N} \sum_{m=1}^{M} \mathbf{b}_{n,m} \mathbf{b}_{n,m}^T$$

Where we have defined:

$$\mathbf{b}_{n,m} = \nabla y_{n,m}$$

Problem 5.17 Solution

It is obvious.

$$\begin{aligned} \frac{\partial^2 E}{\partial w_r \partial w_s} &= \frac{\partial}{\partial w_r} \frac{1}{2} \int\int 2(y-t) \frac{\partial y}{\partial w_s} p(\mathbf{x}, t) d\mathbf{x} dt \\ &= \int\int \left[(y-t) \frac{\partial y^2}{\partial w_r \partial w_s} + \frac{\partial y}{\partial w_s} \frac{\partial y}{\partial w_r} \right] p(\mathbf{x}, t) d\mathbf{x} dt \end{aligned}$$

Since we know that

$$\begin{aligned} \int\int (y-t) \frac{\partial y^2}{\partial w_r \partial w_s} p(\mathbf{x}, t) d\mathbf{x} dt &= \int\int (y-t) \frac{\partial y^2}{\partial w_r \partial w_s} p(t|\mathbf{x}) p(\mathbf{x}) d\mathbf{x} dt \\ &= \int \frac{\partial y^2}{\partial w_r \partial w_s} \Big\{ \int (y-t) p(t|\mathbf{x}) dt \Big\} p(\mathbf{x}) d\mathbf{x} \\ &= 0 \end{aligned}$$

Note that in the last step, we have used $y = \int tp(t|\mathbf{x})dt$. Then we substitute it into the second derivative, which gives,

$$\begin{aligned}\frac{\partial^2 E}{\partial w_r \partial w_s} &= \int\int \frac{\partial y}{\partial w_s}\frac{\partial y}{\partial w_r} p(\mathbf{x},t)d\mathbf{x}dt \\ &= \int \frac{\partial y}{\partial w_s}\frac{\partial y}{\partial w_r} p(\mathbf{x})d\mathbf{x}\end{aligned}$$

Problem 5.18 Solution

By analogy with section 5.3.2, we denote w_{ki}^{skip} as those parameters corresponding to skip-layer connections, i.e., it connects the input unit i with the output unit k. Note that the discussion in section 5.3.2 is still correct and now we only need to obtain the derivative of the error function with respect to the additional parameters w_{ki}^{skip}.

$$\frac{\partial E_n}{\partial w_{ki}^{\text{skip}}} = \frac{\partial E_n}{\partial a_k}\frac{\partial a_k}{\partial w_{ki}^{\text{skip}}} = \delta_k x_i$$

Where we have used $a_k = y_k$ due to linear activation at the output unit and:

$$y_k = \sum_{j=0}^{M} w_{kj}^{(2)} z_j + \sum_i w_{ki}^{\text{skip}} x_i$$

Where the first term on the right side corresponds to those information conveying from the hidden unit to the output and the second term corresponds to the information conveying directly from the input to output.

Problem 5.19 Solution

The error function is given by (5.21). Therefore, we can obtain:

$$\begin{aligned}\nabla E(\mathbf{w}) &= \sum_{n=1}^{N} \frac{\partial E}{\partial a_n}\nabla a_n \\ &= -\sum_{n=1}^{N} \frac{\partial}{\partial a_n}\big[t_n \ln y_n + (1-t_n)\ln(1-y_n)\big]\nabla a_n \\ &= -\sum_{n=1}^{N} \Big\{\frac{\partial(t_n \ln y_n)}{\partial y_n}\frac{\partial y_n}{\partial a_n} + \frac{\partial(1-t_n)\ln(1-y_n)}{\partial y_n}\frac{\partial y_n}{\partial a_n}\Big\}\nabla a_n \\ &= -\sum_{n=1}^{N} \Big[\frac{t_n}{y_n}\cdot y_n(1-y_n) + (1-t_n)\frac{-1}{1-y_n}\cdot y_n(1-y_n)\Big]\nabla a_n \\ &= -\sum_{n=1}^{N} \big[t_n(1-y_n) - (1-t_n)y_n\big]\nabla a_n \\ &= \sum_{n=1}^{N} (y_n - t_n)\nabla a_n\end{aligned}$$

Where we have used the conclusion of problem 5.6. Now we calculate the second derivative.

$$\nabla\nabla E(\mathbf{w}) = \sum_{n=1}^{N} \{y_n(1-y_n)\nabla a_n \nabla a_n + (y_n - t_n)\nabla\nabla a_n\}$$

Similarly, we can drop the last term, which gives exactly what has been asked.

Problem 5.20 Solution(waiting for update)

We begin by writing down the error function.

$$E(\mathbf{w}) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \ln y_{nk}$$

Here we assume that the output of the network has K units in total and there are W weights parameters in the network. WE first calculate the first derivative:

$$\begin{aligned}
\nabla E &= \sum_{n=1}^{N} \frac{dE}{d\mathbf{a}_n} \cdot \nabla \mathbf{a}_n \\
&= -\sum_{n=1}^{N} \Big[\frac{d}{d\mathbf{a}_n} (\sum_{k=1}^{K} t_{nk} \ln y_{nk}) \Big] \cdot \nabla \mathbf{a}_n \\
&= \sum_{n=1}^{N} \mathbf{c}_n \cdot \nabla \mathbf{a}_n
\end{aligned}$$

Note that here $\mathbf{c}_n = -dE/d\mathbf{a}_n$ is a vector with size $K \times 1$, $\nabla \mathbf{a}_n$ is a matrix with size $K \times W$. Moreover, the operator $\cdot$ means inner product, which gives ∇E as a vector with size $1 \times W$. According to (4.106), we can obtain the jth element of $\mathbf{c_n}$:

$$\begin{aligned}
c_{n,j} &= -\frac{\partial}{\partial a_j} (\sum_{k=1}^{K} t_{nk} \ln y_{nk}) \\
&= -\sum_{k=1}^{K} \frac{\partial}{\partial a_j} (t_{nk} \ln y_{nk}) \\
&= -\sum_{k=1}^{K} \frac{t_{nk}}{y_{nk}} y_{nk}(I_{kj} - y_{nj}) \\
&= -\sum_{k=1}^{K} t_{nk} I_{kj} + \sum_{k=1}^{K} t_{nk} y_{nj} \\
&= -t_{nj} + y_{nj} (\sum_{k=1}^{K} t_{nk}) \\
&= y_{nj} - t_{nj}
\end{aligned}$$

Now we calculate the second derivative:

$$\nabla\nabla E = \sum_{n=1}^{N}(\frac{d\mathbf{c}_n}{d\mathbf{a}_n}\nabla\mathbf{a}_n)\cdot\nabla\mathbf{a_n} + \mathbf{c}_n\nabla\nabla\mathbf{a}_n$$

Here $d\mathbf{c}_n/d\mathbf{a}_n$ is a matrix with size $K\times K$. Therefore, the second term can be neglected as before, which gives:

$$\mathbf{H} = \sum_{n=1}^{N}(\frac{d\mathbf{c}_n}{d\mathbf{a}_n}\nabla\mathbf{a}_n)\cdot\nabla\mathbf{a_n}$$

Problem 5.21 Solution

We first write down the expression of Hessian Matrix in the case of K outputs.

$$\mathbf{H}_{N,K} = \sum_{n=1}^{N}\sum_{k=1}^{K}\mathbf{b}_{n,k}\mathbf{b}_{n,k}^T$$

Where $\mathbf{b}_{n,k} = \nabla_{\mathbf{w}}\mathbf{a}_{n,k}$. Therefore, we have:

$$\mathbf{H}_{N+1,K} = \mathbf{H}_{N,K} + \sum_{k=1}^{K}\mathbf{b}_{N+1,k}\mathbf{b}_{N+1,k}^T = \mathbf{H}_{N,K} + \mathbf{B}_{N+1}\mathbf{B}_{N+1}^T$$

Where $\mathbf{B}_{N+1} = [\mathbf{b}_{N+1,1}, \mathbf{b}_{N+1,2}, ..., \mathbf{b}_{N+1,K}]$ is a matrix with size $W\times K$, and here W is the total number of the parameters in the network. By analogy with (5.88)-(5.89), we can obtain:

$$\mathbf{H}_{N+1,K}^{-1} = \mathbf{H}_{N,K}^{-1} - \frac{\mathbf{H}_{N,K}^{-1}\mathbf{B}_{N+1}\mathbf{B}_{N+1}^T\mathbf{H}_{N,K}^{-1}}{1+\mathbf{B}_{N+1}^T\mathbf{H}_{N,K}^{-1}\mathbf{B}_{N+1}} \qquad (*)$$

Furthermore, similarly, we have:

$$\mathbf{H}_{N+1,K+1} = \mathbf{H}_{N+1,K} + \sum_{n=1}^{N+1}\mathbf{b}_{n,K+1}\mathbf{b}_{n,K+1}^T = \mathbf{H}_{N+1,K} + \mathbf{B}_{K+1}\mathbf{B}_{K+1}^T$$

Where $\mathbf{B}_{K+1} = [\mathbf{b}_{1,K+1}, \mathbf{b}_{2,K+1}, ..., \mathbf{b}_{N+1,K+1}]$ is a matrix with size $W\times(N+1)$. Also, we can obtain:

$$\mathbf{H}_{N+1,K+1}^{-1} = \mathbf{H}_{N+1,K}^{-1} - \frac{\mathbf{H}_{N+1,K}^{-1}\mathbf{B}_{K+1}\mathbf{B}_{K+1}^T\mathbf{H}_{N+1,K}^{-1}}{1+\mathbf{B}_{K+1}^T\mathbf{H}_{N+1,K}^{-1}\mathbf{B}_{K+1}}$$

Where $\mathbf{H}_{N+1,K}^{-1}$ is defined by $(*)$. If we substitute $(*)$ into the expression above, we can obtain the relationship between $\mathbf{H}_{N+1,K+1}^{-1}$ and $\mathbf{H}_{N,K}^{-1}$.

Problem 5.22 Solution

We begin by handling the first case.

$$\begin{aligned}
\frac{\partial^2 E_n}{\partial w^{(2)}_{kj}\partial w^{(2)}_{k'j'}} &= \frac{\partial}{\partial w^{(2)}_{kj}}(\frac{\partial E_n}{\partial w^{(2)}_{k'j'}}) \\
&= \frac{\partial}{\partial w^{(2)}_{kj}}(\frac{\partial E_n}{\partial a_{k'}}\frac{\partial a_{k'}}{\partial w^{(2)}_{k'j'}}) \\
&= \frac{\partial}{\partial w^{(2)}_{kj}}(\frac{\partial E_n}{\partial a_{k'}}\frac{\partial \sum_{j'} w_{k'j'}z_{j'}}{\partial w^{(2)}_{k'j'}}) \\
&= \frac{\partial}{\partial w^{(2)}_{kj}}(\frac{\partial E_n}{\partial a_{k'}}z_{j'}) \\
&= \frac{\partial}{\partial w^{(2)}_{kj}}(\frac{\partial E_n}{\partial a_{k'}})z_{j'} + \frac{\partial E_n}{\partial a_{k'}}\frac{\partial z_{j'}}{\partial w^{(2)}_{kj}} \\
&= \frac{\partial}{\partial a_k}(\frac{\partial E_n}{\partial a_{k'}})\frac{\partial a_k}{\partial w^{(2)}_{kj}}z_{j'} + 0 \\
&= \frac{\partial}{\partial a_k}(\frac{\partial E_n}{\partial a_{k'}})z_j z_{j'} \\
&= z_j z_{j'} M_{kk'}
\end{aligned}$$

Then we focus on the second case, and if here $j \neq j'$

$$\begin{aligned}
\frac{\partial^2 E_n}{\partial w^{(1)}_{ji}\partial w^{(1)}_{j'i'}} &= \frac{\partial}{\partial w^{(1)}_{ji}}(\frac{\partial E_n}{\partial w^{(1)}_{j'i'}}) \\
&= \frac{\partial}{\partial w^{(1)}_{ji}}(\sum_{k'}\frac{\partial E_n}{\partial a_{k'}}\frac{\partial a_{k'}}{\partial w^{(1)}_{j'i'}}) \\
&= \sum_{k'}\frac{\partial}{\partial w^{(1)}_{ji}}(\frac{\partial E_n}{\partial a_{k'}}w^{(2)}_{k'j'}h'(a_{j'})x_{i'}) \\
&= \sum_{k'}h'(a_{j'})x_{i'}\frac{\partial}{\partial w^{(1)}_{ji}}(\frac{\partial E_n}{\partial a_{k'}}w^{(2)}_{k'j'}) \\
&= \sum_{k'}h'(a_{j'})x_{i'}\sum_k\frac{\partial}{\partial a_k}(\frac{\partial E_n}{\partial a_{k'}}w^{(2)}_{k'j'})\frac{\partial a_k}{\partial w^{(1)}_{ji}} \\
&= \sum_{k'}h'(a_{j'})x_{i'}\sum_k\frac{\partial}{\partial a_k}(\frac{\partial E_n}{\partial a_{k'}}w^{(2)}_{k'j'})\cdot(w^{(2)}_{kj}h'(a_j)x_i) \\
&= \sum_{k'}h'(a_{j'})x_{i'}\sum_k M_{kk'}w^{(2)}_{k'j'}\cdot w^{(2)}_{kj}h'(a_j)x_i \\
&= x_{i'}x_i h'(a_{j'})h'(a_j)\sum_{k'}\sum_k w^{(2)}_{k'j'}\cdot w^{(2)}_{kj}M_{kk'}
\end{aligned}$$

When $j = j'$, similarly we have:

$$\begin{aligned}
\frac{\partial^2 E_n}{\partial w_{ji}^{(1)} \partial w_{ji'}^{(1)}} &= \sum_{k'} \frac{\partial}{\partial w_{ji}^{(1)}} \left(\frac{\partial E_n}{\partial a_{k'}} w_{k'j}^{(2)} h'(a_j) x_{i'}\right) \\
&= x_{i'} \sum_{k'} \frac{\partial}{\partial w_{ji}^{(1)}} \left(\frac{\partial E_n}{\partial a_{k'}} w_{k'j}^{(2)}\right) h'(a_j) + x_{i'} \sum_{k'} \left(\frac{\partial E_n}{\partial a_{k'}} w_{k'j}^{(2)}\right) \frac{\partial h'(a_j)}{\partial w_{ji}^{(1)}} \\
&= x_{i'} x_i h'(a_j) h'(a_j) \sum_{k'} \sum_{k} w_{k'j}^{(2)} \cdot w_{kj}^{(2)} M_{kk'} + x_{i'} \sum_{k'} \left(\frac{\partial E_n}{\partial a_{k'}} w_{k'j}^{(2)}\right) \frac{\partial h'(a_j)}{\partial w_{ji}^{(1)}} \\
&= x_{i'} x_i h'(a_j) h'(a_j) \sum_{k'} \sum_{k} w_{k'j}^{(2)} \cdot w_{kj}^{(2)} M_{kk'} + x_{i'} \sum_{k'} \left(\frac{\partial E_n}{\partial a_{k'}} w_{k'j}^{(2)}\right) h''(a_j) x_i \\
&= x_{i'} x_i h'(a_j) h'(a_j) \sum_{k'} \sum_{k} w_{k'j}^{(2)} \cdot w_{kj}^{(2)} M_{kk'} + h''(a_j) x_i x_{i'} \sum_{k'} \delta_{k'} w_{k'j}^{(2)}
\end{aligned}$$

It seems that what we have obtained is slightly different from (5.94) when $j = j'$. However this is not the case, since the summation over k' in the second term of our formulation and the summation over k in the first term of (5.94) is actually the same (i.e., they both represent the summation over all the output units). Combining the situation when $j = j'$ and $j \neq j'$, we can obtain (5.94) just as required. Finally, we deal with the third case. Similarly we first focus on $j \neq j'$:

$$\begin{aligned}
\frac{\partial^2 E_n}{\partial w_{ji}^{(1)} \partial w_{kj'}^{(2)}} &= \frac{\partial}{\partial w_{ji}^{(1)}} \left(\frac{\partial E_n}{\partial w_{kj'}^{(2)}}\right) \\
&= \frac{\partial}{\partial w_{ji}^{(1)}} \left(\frac{\partial E_n}{\partial a_k} \frac{\partial a_k}{\partial w_{kj'}^{(2)}}\right) \\
&= \frac{\partial}{\partial w_{ji}^{(1)}} \left(\frac{\partial E_n}{\partial a_k} \frac{\partial \sum_{j'} w_{kj'} z_{j'}}{\partial w_{kj'}^{(2)}}\right) \\
&= \frac{\partial}{\partial w_{ji}^{(1)}} \left(\frac{\partial E_n}{\partial a_k} z_{j'}\right) \\
&= z_{j'} \sum_{k'} \frac{\partial}{\partial a_{k'}} \left(\frac{\partial E_n}{\partial a_k}\right) \frac{\partial a_{k'}}{\partial w_{ji}^{(1)}} \\
&= z_{j'} \sum_{k'} M_{kk'} w_{k'j}^{(2)} h'(a_j) x_i \\
&= x_i h'(a_j) z_{j'} \sum_{k'} M_{kk'} w_{k'j}^{(2)}
\end{aligned}$$

Note that in (5.95), there are two typos: (i)$H_{kk'}$ should be $M_{kk'}$. (ii) j should

exchange position with j' in the right side of (5.95). When $j = j'$, we have:

$$\begin{aligned}
\frac{\partial^2 E_n}{\partial w_{ji}^{(1)} \partial w_{kj}^{(2)}} &= \frac{\partial}{\partial w_{ji}^{(1)}}\left(\frac{\partial E_n}{\partial w_{kj}^{(2)}}\right) \\
&= \frac{\partial}{\partial w_{ji}^{(1)}}\left(\frac{\partial E_n}{\partial a_k}\frac{\partial a_k}{\partial w_{kj}^{(2)}}\right) \\
&= \frac{\partial}{\partial w_{ji}^{(1)}}\left(\frac{\partial E_n}{\partial a_k}\frac{\partial \sum_j w_{kj} z_j}{\partial w_{kj}^{(2)}}\right) \\
&= \frac{\partial}{\partial w_{ji}^{(1)}}\left(\frac{\partial E_n}{\partial a_k} z_j\right) \\
&= \frac{\partial}{\partial w_{ji}^{(1)}}\left(\frac{\partial E_n}{\partial a_k}\right) z_j + \frac{\partial E_n}{\partial a_k}\frac{\partial z_j}{w_{ji}^{(1)}} \\
&= x_i h'(a_j) z_j \sum_{k'} M_{kk'} w_{k'j}^{(2)} + \frac{\partial E_n}{\partial a_k}\frac{\partial z_j}{w_{ji}^{(1)}} \\
&= x_i h'(a_j) z_j \sum_{k'} M_{kk'} w_{k'j}^{(2)} + \delta_k h'(a_j) x_i
\end{aligned}$$

Combing these two situations, we obtain (5.95) just as required.

Problem 5.23 Solution

It is similar to the previous problem.

$$\begin{aligned}
\frac{\partial^2 E_n}{\partial w_{k'i'} \partial w_{kj}} &= \frac{\partial}{\partial w_{k'i'}}\left(\frac{\partial E_n}{\partial w_{kj}}\right) \\
&= \frac{\partial}{\partial w_{k'i'}}\left(\frac{\partial E_n}{\partial a_k} z_j\right) \\
&= z_j \frac{\partial w_{k'i'}}{\partial a_{k'}}\frac{\partial}{\partial a_{k'}}\left(\frac{\partial E_n}{\partial a_k}\right) \\
&= z_j x_{i'} M_{kk'}
\end{aligned}$$

And

$$\begin{aligned}
\frac{\partial^2 E_n}{\partial w_{k'i'}\partial w_{ji}} &= \frac{\partial}{\partial w_{k'i'}}(\sum_k \frac{\partial E_n}{\partial a_k}\frac{\partial a_k}{\partial w_{ji}}) \\
&= \frac{\partial}{\partial w_{k'i'}}(\sum_k \frac{\partial E_n}{\partial a_k} w_{kj} h'(a_j) x_i) \\
&= \sum_k h'(a_j) x_i w_{kj} \frac{\partial}{\partial w_{k'i'}}(\frac{\partial E_n}{\partial a_k}) \\
&= \sum_k h'(a_j) x_i w_{kj} \frac{\partial}{\partial a_{k'}}(\frac{\partial E_n}{\partial a_k})\frac{a_{k'}}{w_{k'i'}} \\
&= \sum_k h'(a_j) x_i w_{kj} M_{kk'} x_{i'} \\
&= x_i x_{i'} h'(a_j) \sum_k w_{kj} M_{kk'}
\end{aligned}$$

Finally, we have

$$\begin{aligned}
\frac{\partial^2 E_n}{\partial w_{k'i'} w_{ki}} &= \frac{\partial}{\partial w_{k'i'}}(\frac{\partial E_n}{\partial w_{ki}}) \\
&= \frac{\partial}{\partial w_{k'i'}}(\frac{\partial E_n}{\partial a_k} x_i) \\
&= x_i \frac{\partial}{\partial a_{k'}}(\frac{\partial E_n}{\partial a_k})\frac{\partial a_{k'}}{w_{k'i'}} \\
&= x_i x_{i'} M_{kk'}
\end{aligned}$$

Problem 5.24 Solution

It is obvious. According to (5.113), we have:

$$\begin{aligned}
\widetilde{a}_j &= \sum_i \widetilde{w}_{ji}\widetilde{x}_i + \widetilde{w}_{j0} \\
&= \sum_i \frac{1}{a} w_{ji} \cdot (a x_i + b) + w_{j0} - \frac{b}{a}\sum_i w_{ji} \\
&= \sum_i w_{ji} x_i + w_{j0} = a_j
\end{aligned}$$

Where we have used (5.115), (5.116) and (5.117). Currently, we have proved that under the transformation the hidden unit a_j is unchanged. If the activation function at the hidden unit is also unchanged, we have $\widetilde{z}_j = z_j$. Now we deal with the output unit $\widetilde{y}_k$:

$$\begin{aligned}
\widetilde{y}_k &= \sum_j \widetilde{w}_{kj}\widetilde{z}_j + \widetilde{w}_{k0} \\
&= \sum_j c w_{kj} \cdot z_j + c w_{k0} + d \\
&= c\sum_j \left[w_{kj} \cdot z_j + w_{k0}\right] + d \\
&= c y_k + d
\end{aligned}$$

Where we have used (5.114), (5.119) and (5.120). To be more specific, here we have proved that the linear transformation between $\widetilde{y}_k$ and y_k can be achieved by making transformation (5.119) and (5.120).

Problem 5.25 Solution

Since we know the gradient of the error function with respect to $\mathbf{w}$ is:

$$\nabla E = \mathbf{H}(\mathbf{w} - \mathbf{w}^*)$$

Together with (5.196), we can obtain:

$$\begin{aligned} \mathbf{w}^{(\tau)} &= \mathbf{w}^{(\tau-1)} - \rho \nabla E \\ &= \mathbf{w}^{(\tau-1)} - \rho \mathbf{H}(\mathbf{w}^{(\tau-1)} - \mathbf{w}^*) \end{aligned}$$

Multiplying both sides by $\mathbf{u}_j^T$, using $w_j = \mathbf{w}^T \mathbf{u}_j$, we can obtain:

$$\begin{aligned} w_j^{(\tau)} &= \mathbf{u}_j^T \left[\mathbf{w}^{(\tau-1)} - \rho \mathbf{H}(\mathbf{w}^{(\tau-1)} - \mathbf{w}^*) \right] \\ &= w_j^{(\tau-1)} - \rho \mathbf{u}_j^T \mathbf{H}(\mathbf{w}^{(\tau-1)} - \mathbf{w}^*) \\ &= w_j^{(\tau-1)} - \rho \eta_j \mathbf{u}_j^T (\mathbf{w}^{(\tau-1)} - \mathbf{w}^*) \\ &= w_j^{(\tau-1)} - \rho \eta_j (w_j^{(\tau-1)} - w_j^*) \\ &= (1 - \rho \eta_j) w_j^{(\tau-1)} + \rho \eta_j w_j^* \end{aligned}$$

Where we have used (5.198). Then we use mathematical deduction to prove (5.197), beginning by calculating $w_j^{(1)}$:

$$\begin{aligned} w_j^{(1)} &= (1 - \rho \eta_j) w_j^{(0)} + \rho \eta_j w_j^* \\ &= \rho \eta_j w_j^* \\ &= \left[1 - (1 - \rho \eta_j) \right] w_j^* \end{aligned}$$

Suppose (5.197) holds for τ, we now prove that it also holds for $\tau + 1$.

$$\begin{aligned} w_j^{(\tau+1)} &= (1 - \rho \eta_j w_j^{(\tau)} + \rho \eta_j w_j^* \\ &= (1 - \rho \eta_j) \left[1 - (1 - \rho \eta_j)^\tau \right] w_j^* + \rho \eta_j w_j^* \\ &= \left\{ (1 - \rho \eta_j) \left[1 - (1 - \rho \eta_j)^\tau \right] + \rho \eta_j \right\} w_j^* \\ &= \left[1 - (1 - \rho \eta_j)^{\tau+1} \right] w_j^* \end{aligned}$$

Hence (5.197) holds for $\tau = 1, 2, \ldots$. Provided $|1 - \rho \eta_j| < 1$, we have $(1 - \rho \eta_j)^\tau \to 0$ as $\tau \to \infty$ ans thus $\mathbf{w}^{(\tau)} = \mathbf{w}^*$. If τ is finite and $\eta_j >> (\rho \tau)^{-1}$, the above argument still holds since τ is still relatively large. Conversely, when $\eta_j << (\rho \tau)^{-1}$, we expand the expression above:

$$|w_j^{(\tau)}| = |\left[1 - (1 - \rho \eta_j)^\tau \right] w_j^*| \approx |\tau \rho \eta_j w_j^*| << |w_j^*|$$

We can see that $(\rho\tau)^{-1}$ works as the regularization parameter α in section 3.5.3.

Problem 5.26 Solution

Based on definition or by analogy with (5.128), we have:

$$\begin{aligned}\Omega_n &= \frac{1}{2}\sum_k (\frac{\partial y_{nk}}{\partial \xi}\big|_{\xi=0})^2 \\ &= \frac{1}{2}\sum_k (\sum_i \frac{\partial y_{nk}}{\partial x_i}\frac{\partial x_i}{\partial \xi}\big|_{\xi=0})^2 \\ &= \frac{1}{2}\sum_k (\sum_i \tau_i \frac{\partial}{\partial x_i} y_{nk})^2\end{aligned}$$

Where we have denoted

$$\tau_i = \frac{\partial x_i}{\partial \xi}\big|_{\xi=0}$$

And this is exactly the form given in (5.201) and (5.202) if the nth observation y_{nk} is denoted as y_k in short. Firstly, we define α_j and β_j as (5.205) shows, where z_j and a_j are given by (5.203). Then we will prove (5.204) holds:

$$\begin{aligned}\alpha_j &= \sum_i \tau_i \frac{\partial z_j}{\partial x_i} = \sum_i \tau_i \frac{\partial h(a_j)}{\partial x_i} \\ &= \sum_i \tau_i \frac{\partial h(a_j)}{\partial a_j}\frac{\partial a_j}{\partial x_i} \\ &= h'(a_j)\sum_i \tau_i \frac{\partial}{\partial x_i} a_j = h'(a_j)\beta_j\end{aligned}$$

Moreover,

$$\begin{aligned}\beta_j &= \sum_i \tau_i \frac{\partial a_j}{\partial x_i} = \sum_i \tau_i \frac{\partial \sum_{i'} w_{ji'} z_{i'}}{\partial x_i} \\ &= \sum_i \tau_i \sum_{i'} \frac{\partial w_{ji'} z_{i'}}{\partial x_i} = \sum_i \tau_i \sum_{i'} w_{ji'} \frac{\partial z_{i'}}{\partial x_i} \\ &= \sum_{i'} w_{ji'} \sum_i \tau_i \frac{\partial z_{i'}}{\partial x_i} = \sum_{i'} w_{ji'} \alpha_{i'}\end{aligned}$$

So far we have proved that (5.204) holds and now we aim to find a forward propagation formula to calculate Ω_n. We firstly begin by evaluating $\{\beta_j\}$ at the input units, and then use the first equation in (5.204) to obtain $\{\alpha_j\}$ at the input units, and then the second equation to evaluate $\{\beta_j\}$ at the first hidden layer, and again the first equation to evaluate $\{\alpha_j\}$ at the first hidden layer. We repeatedly evaluate $\{\beta_j\}$ and $\{\alpha_j\}$ in this way until reaching the output

layer. Then we deal with (5.206):

$$\begin{aligned}
\frac{\partial \Omega_n}{\partial w_{rs}} &= \frac{\partial}{\partial w_{rs}}\{\frac{1}{2}\sum_k (\mathcal{G}y_k)^2\} = \frac{1}{2}\sum_k \frac{\partial (\mathcal{G}y_k)^2}{\partial w_{rs}} \\
&= \frac{1}{2}\sum_k \frac{\partial (\mathcal{G}y_k)^2}{\partial (\mathcal{G}y_k)} \frac{\partial (\mathcal{G}y_k)}{\partial w_{rs}} = \sum_k \mathcal{G}y_k \frac{\partial \mathcal{G}y_k}{\partial w_{rs}} \\
&= \sum_k \mathcal{G}y_k \mathcal{G}\big[\frac{\partial y_k}{\partial w_{rs}}\big] = \sum_k \alpha_k \mathcal{G}\big[\frac{\partial y_k}{\partial a_r}\frac{\partial a_r}{\partial w_{rs}}\big] \\
&= \sum_k \alpha_k \mathcal{G}\big[\delta_{kr} z_s\big] = \sum_k \alpha_k \big\{\mathcal{G}[\delta_{kr}] z_s + \mathcal{G}[z_s]\delta_{kr}\big\} \\
&= \sum_k \alpha_k \big\{\phi_{kr} z_s + \alpha_s \delta_{kr}\big\}
\end{aligned}$$

Provided with the idea in section 5.3, the backward propagation formula is easy to derive. We can simply replace E_n with y_k to obtain a backward equation, so we omit it here.

Problem 5.27 Solution

Following the procedure in section 5.5.5, we can obtain:

$$\Omega = \frac{1}{2}\int (\boldsymbol{\tau}^T \nabla y(\mathbf{x}))^2 \, p(\mathbf{x}) d\mathbf{x}$$

Since we have $\boldsymbol{\tau} = \partial \mathbf{s}(\mathbf{x}, \boldsymbol{\xi}) \big/ \partial \boldsymbol{\xi}$ and $\mathbf{s} = \mathbf{x} + \boldsymbol{\xi}$, so we have $\boldsymbol{\tau} = \mathbf{I}$. Therefore, substituting $\boldsymbol{\tau}$ into the equation above, we can obtain:

$$\Omega = \frac{1}{2}\int (\nabla y(\mathbf{x}))^2 \, p(\mathbf{x}) d\mathbf{x}$$

Just as required.

Problem 5.28 Solution

The modifications only affect derivatives with respect to the weights in the convolutional layer. The units within a feature map (indexed m) have different inputs, but all share a common weight vector, $\mathbf{w}^{(m)}$. Therefore, we can write:

$$\frac{\partial E_n}{\partial w_i^{(m)}} = \sum_j \frac{\partial E_n}{\partial a_j^{(m)}} \frac{\partial a_j^{(m)}}{\partial w_i^{(m)}} = \sum_j \delta_j^{(m)} z_{ji}^{(m)}$$

Here $a_j^{(m)}$ denotes the activation of the jth unit in th mth feature map, whereas $w_i^{(m)}$ denotes the ith element of the corresponding feature vector and finally $z_{ij}^{(m)}$ denotes the ith input for the jth unit in the mth feature map. Note that $\delta_j^{(m)}$ can be computed recursively from the units in the following layer.

Problem 5.29 Solution

It is obvious. Firstly, we know that:

$$\frac{\partial}{\partial w_i}\{\pi_j \mathcal{N}(w_i|\mu_j,\sigma_j^2)\} = -\pi_j \frac{w_i-\mu_j}{\sigma_j^2}\mathcal{N}(w_i|\mu_j,\sigma_j^2)$$

We now derive the error function with respect to w_i:

$$\begin{aligned}
\frac{\partial \widetilde{E}}{\partial w_i} &= \frac{\partial E}{\partial w_i} + \frac{\partial \lambda\Omega(\mathbf{w})}{\partial w_i} \\
&= \frac{\partial E}{\partial w_i} - \lambda \frac{\partial}{\partial w_i}\left\{\sum_i \ln\left(\sum_{j=1}^{M}\pi_j \mathcal{N}(w_i|\mu_j,\sigma_j^2)\right)\right\} \\
&= \frac{\partial E}{\partial w_i} - \lambda \frac{\partial}{\partial w_i}\left\{\ln\left(\sum_{j=1}^{M}\pi_j \mathcal{N}(w_i|\mu_j,\sigma_j^2)\right)\right\} \\
&= \frac{\partial E}{\partial w_i} - \lambda \frac{1}{\sum_{j=1}^{M}\pi_j \mathcal{N}(w_i|\mu_j,\sigma_j^2)}\frac{\partial}{\partial w_i}\left\{\sum_{j=1}^{M}\pi_j \mathcal{N}(w_i|\mu_j,\sigma_j^2)\right\} \\
&= \frac{\partial E}{\partial w_i} + \lambda \frac{1}{\sum_{j=1}^{M}\pi_j \mathcal{N}(w_i|\mu_j,\sigma_j^2)}\left\{\sum_{j=1}^{M}\pi_j \frac{w_i-\mu_j}{\sigma_j^2}\mathcal{N}(w_i|\mu_j,\sigma_j^2)\right\} \\
&= \frac{\partial E}{\partial w_i} + \lambda \frac{\sum_{j=1}^{M}\pi_j \frac{w_i-\mu_j}{\sigma_j^2}\mathcal{N}(w_i|\mu_j,\sigma_j^2)}{\sum_k \pi_k \mathcal{N}(w_i|\mu_k,\sigma_k^2)} \\
&= \frac{\partial E}{\partial w_i} + \lambda \sum_{j=1}^{M}\frac{\pi_j \mathcal{N}(w_i|\mu_j,\sigma_j^2)}{\sum_k \pi_k \mathcal{N}(w_i|\mu_k,\sigma_k^2)}\frac{w_i-\mu_j}{\sigma_j^2} \\
&= \frac{\partial E}{\partial w_i} + \lambda \sum_{j=1}^{M}\gamma_j(w_i)\frac{w_i-\mu_j}{\sigma_j^2}
\end{aligned}$$

Where we have used (5.138) and defined (5.140).

Problem 5.30 Solution

Is is similar to the previous problem. Since we know that:

$$\frac{\partial}{\partial \mu_j}\{\pi_j \mathcal{N}(w_i|\mu_j,\sigma_j^2)\} = \pi_j \frac{w_i-\mu_j}{\sigma_j^2}\mathcal{N}(w_i|\mu_j,\sigma_j^2)$$

We can derive:

$$\begin{aligned}
\frac{\partial \widetilde{E}}{\partial \mu_j} &= \frac{\partial \lambda \Omega(\mathbf{w})}{\partial \mu_j} \\
&= -\lambda \frac{\partial}{\partial \mu_j} \left\{ \sum_i \ln \left(\sum_{j=1}^{M} \pi_j \mathcal{N}(w_i | \mu_j, \sigma_j^2) \right) \right\} \\
&= -\lambda \sum_i \frac{\partial}{\partial \mu_j} \left\{ \ln \left(\sum_{j=1}^{M} \pi_j \mathcal{N}(w_i | \mu_j, \sigma_j^2) \right) \right\} \\
&= -\lambda \sum_i \frac{1}{\sum_{j=1}^{M} \pi_j \mathcal{N}(w_i | \mu_j, \sigma_j^2)} \frac{\partial}{\partial \mu_j} \left\{ \sum_{j=1}^{M} \pi_j \mathcal{N}(w_i | \mu_j, \sigma_j^2) \right\} \\
&= -\lambda \sum_i \frac{1}{\sum_{j=1}^{M} \pi_j \mathcal{N}(w_i | \mu_j, \sigma_j^2)} \pi_j \frac{w_i - \mu_j}{\sigma_j^2} \mathcal{N}(w_i | \mu_j, \sigma_j^2) \\
&= \lambda \sum_i \frac{\pi_j \mathcal{N}(w_i | \mu_j, \sigma_j^2)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(w_i | \mu_k, \sigma_k^2)} \frac{\mu_j - w_i}{\sigma_j^2} = \lambda \sum_i \gamma_j(w_i) \frac{\mu_j - w_i}{\sigma_j^2}
\end{aligned}$$

Note that there is a typo in (5.142). The numerator should be $\mu_j - w_i$ instead of $\mu_i - w_j$. This can be easily seen through the fact that the mean and variance of the Gaussian Distribution should have the same subindex and since σ_j is in the denominator, μ_j should occur in the numerator instead of μ_i.

Problem 5.31 Solution

It is similar to the previous problem. Since we know that:

$$\frac{\partial}{\partial \sigma_j} \{ \pi_j \mathcal{N}(w_i | \mu_j, \sigma_j^2) \} = \left(-\frac{1}{\sigma_j} + \frac{(w_i - \mu_j)^2}{\sigma_j^3} \right) \pi_j \mathcal{N}(w_i | \mu_j, \sigma_j^2)$$

We can derive:

$$\begin{aligned}
\frac{\partial \widetilde{E}}{\partial \sigma_j} &= \frac{\partial \lambda \Omega(\mathbf{w})}{\partial \sigma_j} \\
&= -\lambda \frac{\partial}{\partial \sigma_j} \left\{ \sum_i \ln \left(\sum_{j=1}^{M} \pi_j \mathcal{N}(w_i | \mu_j, \sigma_j^2) \right) \right\} \\
&= -\lambda \sum_i \frac{\partial}{\partial \sigma_j} \left\{ \ln \left(\sum_{j=1}^{M} \pi_j \mathcal{N}(w_i | \mu_j, \sigma_j^2) \right) \right\} \\
&= -\lambda \sum_i \frac{1}{\sum_{j=1}^{M} \pi_j \mathcal{N}(w_i | \mu_j, \sigma_j^2)} \frac{\partial}{\partial \sigma_j} \left\{ \sum_{j=1}^{M} \pi_j \mathcal{N}(w_i | \mu_j, \sigma_j^2) \right\} \\
&= -\lambda \sum_i \frac{1}{\sum_{j=1}^{M} \pi_j \mathcal{N}(w_i | \mu_j, \sigma_j^2)} \frac{\partial}{\partial \sigma_j} \left\{ \pi_j \mathcal{N}(w_i | \mu_j, \sigma_j^2) \right\} \\
&= \lambda \sum_i \frac{1}{\sum_{j=1}^{M} \pi_j \mathcal{N}(w_i | \mu_j, \sigma_j^2)} \left(\frac{1}{\sigma_j} - \frac{(w_i - \mu_j)^2}{\sigma_j^3} \right) \pi_j \mathcal{N}(w_i | \mu_j, \sigma_j^2) \\
&= \lambda \sum_i \frac{\pi_j \mathcal{N}(w_i | \mu_j, \sigma_j^2)}{\sum_{k=1}^{M} \pi_k \mathcal{N}(w_i | \mu_k, \sigma_k^2)} \left(\frac{1}{\sigma_j} - \frac{(w_i - \mu_j)^2}{\sigma_j^3} \right) \\
&= \lambda \sum_i \gamma_j(w_i) \left(\frac{1}{\sigma_j} - \frac{(w_i - \mu_j)^2}{\sigma_j^3} \right)
\end{aligned}$$

Just as required.

Problem 5.32 Solution

It is trivial. We begin by verifying (5.208) when $j \neq k$.

$$\begin{aligned}
\frac{\partial \pi_k}{\partial \eta_j} &= \frac{\partial}{\partial \eta_j} \left\{ \frac{exp(\eta_k)}{\sum_k exp(\eta_k)} \right\} \\
&= \frac{-exp(\eta_k) exp(\eta_j)}{\left[\sum_k exp(\eta_k) \right]^2} \\
&= -\pi_j \pi_k
\end{aligned}$$

And if now we have $j = k$:

$$\begin{aligned}
\frac{\partial \pi_k}{\partial \eta_k} &= \frac{\partial}{\partial \eta_k} \left\{ \frac{exp(\eta_k)}{\sum_k exp(\eta_k)} \right\} \\
&= \frac{exp(\eta_k) \left[\sum_k exp(\eta_k) \right] - exp(\eta_k) exp(\eta_k)}{\left[\sum_k exp(\eta_k) \right]^2} \\
&= \pi_k - \pi_k \pi_k
\end{aligned}$$

If we combine these two cases, we can easily see that (5.208) holds. Now

we prove (5.147).

$$\begin{aligned}
\frac{\partial \widetilde{E}}{\partial \eta_j} &= \lambda \frac{\partial \Omega(\mathbf{w})}{\partial \eta_j} \\
&= -\lambda \frac{\partial}{\partial \eta_j} \left\{ \sum_i \ln \left\{ \sum_{j=1}^{M} \pi_j \mathcal{N}(w_i|\mu_j, \sigma_j^2) \right\} \right\} \\
&= -\lambda \sum_i \frac{\partial}{\partial \eta_j} \left\{ \ln \left\{ \sum_{j=1}^{M} \pi_j \mathcal{N}(w_i|\mu_j, \sigma_j^2) \right\} \right\} \\
&= -\lambda \sum_i \frac{1}{\sum_{j=1}^{M} \pi_j \mathcal{N}(w_i|\mu_j, \sigma_j^2)} \frac{\partial}{\partial \eta_j} \left\{ \sum_{k=1}^{M} \pi_k \mathcal{N}(w_i|\mu_k, \sigma_k^2) \right\} \\
&= -\lambda \sum_i \frac{1}{\sum_{j=1}^{M} \pi_j \mathcal{N}(w_i|\mu_j, \sigma_j^2)} \sum_{k=1}^{M} \frac{\partial}{\partial \eta_j} \left\{ \pi_k \mathcal{N}(w_i|\mu_k, \sigma_k^2) \right\} \\
&= -\lambda \sum_i \frac{1}{\sum_{j=1}^{M} \pi_j \mathcal{N}(w_i|\mu_j, \sigma_j^2)} \sum_{k=1}^{M} \frac{\partial}{\partial \pi_k} \left\{ \pi_k \mathcal{N}(w_i|\mu_k, \sigma_k^2) \right\} \frac{\partial \pi_k}{\partial \eta_j} \\
&= -\lambda \sum_i \frac{1}{\sum_{j=1}^{M} \pi_j \mathcal{N}(w_i|\mu_j, \sigma_j^2)} \sum_{k=1}^{M} \mathcal{N}(w_i|\mu_k, \sigma_k^2)(\delta_{jk}\pi_j - \pi_j \pi_k) \\
&= -\lambda \sum_i \frac{1}{\sum_{j=1}^{M} \pi_j \mathcal{N}(w_i|\mu_j, \sigma_j^2)} \left\{ \pi_j \mathcal{N}(w_i|\mu_j, \sigma_j^2) - \pi_j \sum_{k=1}^{M} \pi_k \mathcal{N}(w_i|\mu_k, \sigma_k^2)) \right\} \\
&= -\lambda \sum_i \left\{ \frac{\pi_j \mathcal{N}(w_i|\mu_j, \sigma_j^2)}{\sum_{j=1}^{M} \pi_j \mathcal{N}(w_i|\mu_j, \sigma_j^2)} - \frac{\pi_j \sum_{k=1}^{M} \pi_k \mathcal{N}(w_i|\mu_k, \sigma_k^2))}{\sum_{j=1}^{M} \pi_j \mathcal{N}(w_i|\mu_j, \sigma_j^2)} \right\} \\
&= -\lambda \sum_i \{\gamma_j(w_i) - \pi_j\} = \lambda \sum_i \{\pi_j - \gamma_j(w_i)\}
\end{aligned}$$

Just as required.

Problem 5.33 Solution

It is trivial. We set the attachment point of the lower arm with the ground as the origin of the coordinate. We first aim to find the vertical distance from the origin to the target point, and this is also the value of x_2.

$$\begin{aligned}
x_2 &= L_1 \sin(\pi - \theta_1) + L_2 \sin(\theta_2 - (\pi - \theta_1)) \\
&= L_1 \sin\theta_1 - L_2 \sin(\theta_1 + \theta_2)
\end{aligned}$$

Similarly, we calculate the horizontal distance from the origin to the target point.

$$\begin{aligned}
x_1 &= -L_1 \cos(\pi - \theta_1) + L_2 \cos(\theta_2 - (\pi - \theta_1)) \\
&= L_1 \cos\theta_1 - L_2 \cos(\theta_1 + \theta_2)
\end{aligned}$$

From these two equations, we can clearly see the 'forward kinematics' of the robot arm.

Problem 5.34 Solution

By analogy with (5.208), we can write:

$$\frac{\partial \pi_k(\mathbf{x})}{\partial a_j^\pi} = \delta_{jk}\pi_j(\mathbf{x}) - \pi_j(\mathbf{x})\pi_k(\mathbf{x})$$

Using (5.153), we can see that:

$$E_n = -\ln\left\{\sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2)\right\}$$

Therefore, we can derive:

$$\begin{aligned}
\frac{\partial E_n}{\partial a_j^\pi} &= -\frac{\partial}{\partial a_j^\pi}\ln\left\{\sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2)\right\} \\
&= -\frac{1}{\sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2)}\frac{\partial}{\partial a_j^\pi}\left\{\sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2)\right\} \\
&= -\frac{1}{\sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2)}\sum_{k=1}^{K}\frac{\partial \pi_k}{\partial a_j^\pi}\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2) \\
&= -\frac{1}{\sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2)}\sum_{k=1}^{K}\left[\delta_{jk}\pi_j(\mathbf{x}_n) - \pi_j(\mathbf{x}_n)\pi_k(\mathbf{x}_n)\right]\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2) \\
&= -\frac{1}{\sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2)}\left\{\pi_j(\mathbf{x}_n)\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_j,\sigma_j^2) - \pi_j(\mathbf{x}_n)\sum_{k=1}^{K}\pi_k(\mathbf{x}_n)\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2)\right\} \\
&= \frac{1}{\sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2)}\left\{-\pi_j(\mathbf{x}_n)\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_j,\sigma_j^2) + \pi_j(\mathbf{x}_n)\sum_{k=1}^{K}\pi_k(\mathbf{x}_n)\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2)\right\}
\end{aligned}$$

And if we denoted (5.154), we will have:

$$\frac{\partial E_n}{\partial a_j^\pi} = -\gamma_j + \pi_j$$

Note that our result is slightly different from (5.155) by the subindex. But there are actually the same if we substitute index j by index k in the final expression.

Problem 5.35 Solution

We deal with the derivative of error function with respect to $\boldsymbol{\mu}_k$ instead, which will give a vector as result. Furthermore, the lth element of this vector will be what we have been required. Since we know that:

$$\frac{\partial}{\partial \boldsymbol{\mu}_k}\{\pi_k\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2)\} = \frac{\mathbf{t}_n - \boldsymbol{\mu}_k}{\sigma_k^2}\pi_k\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2)$$

One thing worthy noticing is that here we focus on the isotropic case as stated in page 273 of the textbook. To be more precise, $\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2)$ should be $\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2\mathbf{I})$. Provided with the equation above, we can further obtain:

$$\begin{aligned}
\frac{\partial E_n}{\partial \boldsymbol{\mu}_k} &= \frac{\partial}{\partial \boldsymbol{\mu}_k}\left\{-\ln\sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2)\right\} \\
&= -\frac{1}{\sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2)}\frac{\partial}{\partial \boldsymbol{\mu}_k}\left\{\sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2)\right\} \\
&= -\frac{1}{\sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2)}\cdot\frac{\mathbf{t}_n-\boldsymbol{\mu}_k}{\sigma_k^2}\pi_k\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2) \\
&= -\gamma_k\frac{\mathbf{t}_n-\boldsymbol{\mu}_k}{\sigma_k^2}
\end{aligned}$$

Hence noticing (5.152), the lth element of the result above is what we are required.

$$\frac{\partial E_n}{\partial a_{kl}^{\mu}} = \frac{\partial E_n}{\partial \mu_{kl}} = \gamma_k\frac{\mu_{kl}-\mathbf{t}_l}{\sigma_k^2}$$

Problem 5.36 Solution

Similarly, we know that:

$$\frac{\partial}{\partial \sigma_k}\{\pi_k\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2)\} = \left\{-\frac{D}{\sigma_k}+\frac{||\mathbf{t}_n-\boldsymbol{\mu}_k||^2}{\sigma_k^3}\right\}\pi_k\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2)$$

Therefore, we can obtain:

$$\begin{aligned}
\frac{\partial E_n}{\partial \sigma_k} &= \frac{\partial}{\partial \sigma_k}\left\{-\ln\sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2)\right\} \\
&= -\frac{1}{\sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2)}\frac{\partial}{\partial \sigma_k}\left\{\sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2)\right\} \\
&= -\frac{1}{\sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2)}\cdot\left\{-\frac{D}{\sigma_k}+\frac{||\mathbf{t}_n-\boldsymbol{\mu}_k||^2}{\sigma_k^3}\right\}\pi_k\mathcal{N}(\mathbf{t}_n|\boldsymbol{\mu}_k,\sigma_k^2) \\
&= -\gamma_k\left\{-\frac{D}{\sigma_k}+\frac{||\mathbf{t}_n-\boldsymbol{\mu}_k||^2}{\sigma_k^3}\right\}
\end{aligned}$$

Note that there is a typo in (5.157) and the underlying reason is that: $|\sigma_k^2\mathbf{I}_{D\times D}| = (\sigma_k^2)^D$

Problem 5.37 Solution

First we know two properties for the Gaussian distribution $\mathcal{N}(\mathbf{t}|\boldsymbol{\mu},\sigma^2\mathbf{I})$:

$$\mathbb{E}[\mathbf{t}] = \int \mathbf{t}\mathcal{N}(\mathbf{t}|\boldsymbol{\mu},\sigma^2\mathbf{I})d\mathbf{t} = \boldsymbol{\mu}$$

And

$$\mathbb{E}[||\mathbf{t}||^2] = \int ||\mathbf{t}||^2 \mathcal{N}(\mathbf{t}|\boldsymbol{\mu}, \sigma^2\mathbf{I}) d\mathbf{t} = L\sigma^2 + ||\boldsymbol{\mu}||^2$$

Where we have used $\mathbb{E}[\mathbf{t}^T\mathbf{A}\mathbf{t}] = \mathrm{Tr}[\mathbf{A}\sigma^2\mathbf{I}] + \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu}$ by setting $\mathbf{A} = \mathbf{I}$. This property can be found in *Matrixcookbook* eq(378). Here L is the dimension of $\mathbf{t}$. Noticing (5.148), we can write:

$$\begin{aligned}
\mathbb{E}[\mathbf{t}|\mathbf{x}] &= \int \mathbf{t} p(\mathbf{t}|\mathbf{x}) d\mathbf{t} \\
&= \int \mathbf{t} \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{t}|\boldsymbol{\mu}_k, \sigma_k^2) d\mathbf{t} \\
&= \sum_{k=1}^{K} \pi_k \int \mathbf{t} \mathcal{N}(\mathbf{t}|\boldsymbol{\mu}_k, \sigma_k^2) d\mathbf{t} \\
&= \sum_{k=1}^{K} \pi_k \boldsymbol{\mu}_k
\end{aligned}$$

Then we prove (5.160).

$$\begin{aligned}
s^2(\mathbf{x}) &= \mathbb{E}[||\mathbf{t} - \mathbb{E}[\mathbf{t}|\mathbf{x}]||^2|\mathbf{x}] = \mathbb{E}[\left(\mathbf{t}^2 - 2\mathbf{t}\mathbb{E}[\mathbf{t}|\mathbf{x}] + \mathbb{E}[\mathbf{t}|\mathbf{x}]^2\right)|\mathbf{x}] \\
&= \mathbb{E}[\mathbf{t}^2|\mathbf{x}] - \mathbb{E}[2\mathbf{t}\mathbb{E}[\mathbf{t}|\mathbf{x}]|\mathbf{x}] + \mathbb{E}[\mathbf{t}|\mathbf{x}]^2 = \mathbb{E}[\mathbf{t}^2|\mathbf{x}] - \mathbb{E}[\mathbf{t}|\mathbf{x}]^2 \\
&= \int ||\mathbf{t}||^2 \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{\mu}_k, \sigma_k^2) d\mathbf{t} - ||\sum_{l=1}^{K} \pi_l \boldsymbol{\mu}_l||^2 \\
&= \sum_{k=1}^{K} \pi_k \int ||\mathbf{t}||^2 \mathcal{N}(\boldsymbol{\mu}_k, \sigma_k^2) d\mathbf{t} - ||\sum_{l=1}^{K} \pi_l \boldsymbol{\mu}_l||^2 \\
&= \sum_{k=1}^{K} \pi_k (L\sigma_k^2 + ||\boldsymbol{\mu}_k||^2) - ||\sum_{l=1}^{K} \pi_l \boldsymbol{\mu}_l||^2 \\
&= L\sum_{k=1}^{K} \pi_k \sigma_k^2 + \sum_{k=1}^{K} \pi_k ||\boldsymbol{\mu}_k||^2 - ||\sum_{l=1}^{K} \pi_l \boldsymbol{\mu}_l||^2 \\
&= L\sum_{k=1}^{K} \pi_k \sigma_k^2 + \sum_{k=1}^{K} \pi_k ||\boldsymbol{\mu}_k||^2 - 2 \times ||\sum_{l=1}^{K} \pi_l \boldsymbol{\mu}_l||^2 + 1 \times ||\sum_{l=1}^{K} \pi_l \boldsymbol{\mu}_l||^2 \\
&= L\sum_{k=1}^{K} \pi_k \sigma_k^2 + \sum_{k=1}^{K} \pi_k ||\boldsymbol{\mu}_k||^2 - 2(\sum_{l=1}^{K} \pi_l \boldsymbol{\mu}_l)(\sum_{k=1}^{K} \pi_k \boldsymbol{\mu}_k) + \left(\sum_{k=1}^{K} \pi_k\right) ||\sum_{l=1}^{K} \pi_l \boldsymbol{\mu}_l||^2 \\
&= L\sum_{k=1}^{K} \pi_k \sigma_k^2 + \sum_{k=1}^{K} \pi_k ||\boldsymbol{\mu}_k||^2 - 2(\sum_{l=1}^{K} \pi_l \boldsymbol{\mu}_l)(\sum_{k=1}^{K} \pi_k \boldsymbol{\mu}_k) + \sum_{k=1}^{K} \pi_k ||\sum_{l=1}^{K} \pi_l \boldsymbol{\mu}_l||^2 \\
&= L\sum_{k=1}^{K} \pi_k \sigma_k^2 + \sum_{k=1}^{K} \pi_k ||\boldsymbol{\mu}_k - \sum_{l=1}^{K} \pi_l \boldsymbol{\mu}_l||^2 \\
&= \sum_{k=1}^{K} \pi_k \left(L\sigma_k^2 + ||\boldsymbol{\mu}_k - \sum_{l=1}^{K} \pi_l \boldsymbol{\mu}_l||^2\right)
\end{aligned}$$

Note that there is a typo in (5.160), i.e., the coefficient L in front of σ_k^2 is missing.

Problem 5.38 Solution

From (5.167) and (5.171), we can write down the expression for the predictive distribution:

$$\begin{aligned} p(t|\mathbf{x},D,\alpha,\beta) &= \int p(\mathbf{w}|D,\alpha,\beta)p(t|\mathbf{x},\mathbf{w},\beta)d\mathbf{w} \\ &\approx \int q(\mathbf{w}|D)p(t|\mathbf{x},\mathbf{w},\beta)d\mathbf{w} \\ &= \int \mathcal{N}(\mathbf{w}|\mathbf{w}_{\text{MAP}},\mathbf{A}^{-1})\mathcal{N}(t|\mathbf{g}^T\mathbf{w}-\mathbf{g}^T\mathbf{w}_{\text{MAP}}+y(\mathbf{x},\mathbf{w}_{\text{MAP}}),\beta^{-1})d\mathbf{w} \end{aligned}$$

Note here $p(t|\mathbf{x},\mathbf{w},\beta)$ is given by (5.171) and $q(\mathbf{w}|D)$ is the approximation to the posterior $p(\mathbf{w}|D,\alpha,\beta)$, which is given by (5.167). Then by analogy with (2.115), we first deal with the mean of the predictive distribution:

$$\begin{aligned} \text{mean} &= \mathbf{g}^T\mathbf{w}-\mathbf{g}^T\mathbf{w}_{\text{MAP}}+y(\mathbf{x},\mathbf{w}_{\text{MAP}})|_{\mathbf{w}=\mathbf{w}_{\text{MAP}}} \\ &= y(\mathbf{x},\mathbf{w}_{\text{MAP}}) \end{aligned}$$

Then we deal with the covariance matrix:

$$\text{Covariance matrix} = \beta^{-1}+\mathbf{g}^T\mathbf{A}^{-1}\mathbf{g}$$

Just as required.

Problem 5.39 Solution

Using Laplace Approximation, we can obtain:

$$p(D|\mathbf{w},\beta)p(\mathbf{w}|\alpha) = p(D|\mathbf{w}_{\text{MAP}},\beta)p(\mathbf{w}_{\text{MAP}}|\alpha)\exp\left\{-(\mathbf{w}-\mathbf{w}_{\text{MAP}})^T\mathbf{A}(\mathbf{w}-\mathbf{w}_{\text{MAP}})\right\}$$

Then using (5.174), (5.162) and (5.163), we can obtain:

$$\begin{aligned} p(D|\alpha,\beta) &= \int p(D|\mathbf{w},\beta)p(\mathbf{w},\alpha)d\mathbf{w} \\ &= \int p(D|\mathbf{w}_{\text{MAP}},\beta)p(\mathbf{w}_{\text{MAP}}|\alpha)\exp\left\{-(\mathbf{w}-\mathbf{w}_{\text{MAP}})^T\mathbf{A}(\mathbf{w}-\mathbf{w}_{\text{MAP}})\right\}d\mathbf{w} \\ &= p(D|\mathbf{w}_{\text{MAP}},\beta)p(\mathbf{w}_{\text{MAP}}|\alpha)\frac{(2\pi)^{W/2}}{|\mathbf{A}|^{1/2}} \\ &= \prod_{n=1}^{N}\mathcal{N}(t_n|y(\mathbf{x}_n,\mathbf{w}_{\text{MAP}}),\beta^{-1})\mathcal{N}(\mathbf{w}_{\text{MAP}}|\mathbf{0},\alpha^{-1}\mathbf{I})\frac{(2\pi)^{W/2}}{|\mathbf{A}|^{1/2}} \end{aligned}$$

If we take logarithm of both sides, we will obtain (5.175) just as required.

Problem 5.40 Solution

For a k-class classification problem, we need to use softmax activation function and also the error function is now given by (5.24). Therefore, the

Hessian matrix should be derived from (5.24) and the cross entropy in (5.184) will also be replaced by (5.24).

Problem 5.41 Solution

By analogy to Prob.5.39, we can write:

$$p(D|\alpha) = p(D|\mathbf{w}_{\text{MAP}})p(\mathbf{w}_{\text{MAP}}|\alpha)\frac{(2\pi)^{W/2}}{|\mathbf{A}|^{1/2}}$$

Since we know that the prior $p(\mathbf{w}|\alpha)$ follows a Gaussian distribution, i.e., (5.162), as stated in the text. Therefore we can obtain:

$$\begin{aligned}
\ln p(D|\alpha) &= \ln p(D|\mathbf{w}_{\text{MAP}}) + \ln p(\mathbf{w}_{\text{MAP}}|\alpha) - \frac{1}{2}\ln|\mathbf{A}| + \text{const} \\
&= \ln p(D|\mathbf{w}_{\text{MAP}}) - \frac{\alpha}{2}\mathbf{w}^T\mathbf{w} + \frac{W}{2}\ln\alpha - \frac{1}{2}\ln|\mathbf{A}| + \text{const} \\
&= -E(\mathbf{w}_{\text{MAP}}) + \frac{W}{2}\ln\alpha - \frac{1}{2}\ln|\mathbf{A}| + \text{const}
\end{aligned}$$

Just as required.

0.6 Kernel Methods

Problem 6.1 Solution

Recall that in section.6.1, a_n can be written as (6.4). We can derive:

$$\begin{aligned}
a_n &= -\frac{1}{\lambda}\{\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n) - t_n\} \\
&= -\frac{1}{\lambda}\{w_1\phi_1(\mathbf{x}_n) + w_2\phi_2(\mathbf{x}_n) + \ldots + w_M\phi_M(\mathbf{x}_n) - t_n\} \\
&= -\frac{w_1}{\lambda}\phi_1(\mathbf{x}_n) - \frac{w_2}{\lambda}\phi_2(\mathbf{x}_n) - \ldots - \frac{w_M}{\lambda}\phi_M(\mathbf{x}_n) + \frac{t_n}{\lambda} \\
&= (c_n - \frac{w_1}{\lambda})\phi_1(\mathbf{x}_n) + (c_n - \frac{w_2}{\lambda})\phi_2(\mathbf{x}_n) + \ldots + (c_n - \frac{w_M}{\lambda})\phi_M(\mathbf{x}_n)
\end{aligned}$$

Here we have defined:

$$c_n = \frac{t_n/\lambda}{\phi_1(\mathbf{x}_n) + \phi_2(\mathbf{x}_n) + \ldots + \phi_M(\mathbf{x}_n)}$$

From what we have derived above, we can see that a_n is a linear combination of $\boldsymbol{\phi}(\mathbf{x}_n)$. What's more, we first substitute $\mathbf{K} = \boldsymbol{\Phi}\boldsymbol{\Phi}^T$ into (6.7), and then we will obtain (6.5). Next we substitute (6.3) into (6.5) we will obtain (6.2) just as required.

Problem 6.2 Solution

If we set $\mathbf{w}^{(0)} = \mathbf{0}$ in (4.55), we can obtain:

$$\mathbf{w}^{(\tau+1)} = \sum_{n=1}^{N} \eta c_n t_n \boldsymbol{\phi}_n$$

where N is the total number of samples and c_n is the times that $t_n \boldsymbol{\phi}_n$ has been added from step 0 to step $\tau + 1$. Therefore, it is obvious that we have:

$$\mathbf{w} = \sum_{n=1}^{N} \alpha_n t_n \boldsymbol{\phi}_n$$

We further substitute the expression above into (4.55), which gives:

$$\sum_{n=1}^{N} \alpha_n^{(\tau+1)} t_n \boldsymbol{\phi}_n = \sum_{n=1}^{N} \alpha_n^{(\tau)} t_n \boldsymbol{\phi}_n + \eta t_n \boldsymbol{\phi}_n$$

In other words, the update process is to add learning rate η to the coefficient α_n corresponding to the misclassified pattern $\mathbf{x}_n$, i.e.,

$$\alpha_n^{(\tau+1)} = \alpha_n^{(\tau)} + \eta$$

Now we similarly substitute it into (4.52):

$$\begin{aligned} y(\mathbf{x}) &= f(\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x})) \\ &= f(\sum_{n=1}^{N} \alpha_n t_n \boldsymbol{\phi}_n^T \boldsymbol{\phi}(\mathbf{x})) \\ &= f(\sum_{n=1}^{N} \alpha_n t_n k(\mathbf{x}_n, \mathbf{x})) \end{aligned}$$

Problem 6.3 Solution

We begin by expanding the Euclidean metric.

$$\begin{aligned} ||\mathbf{x} - \mathbf{x}_n||^2 &= (\mathbf{x} - \mathbf{x}_n)^T (\mathbf{x} - \mathbf{x}_n) \\ &= (\mathbf{x}^T - \mathbf{x}_n^T)(\mathbf{x} - \mathbf{x}_n) \\ &= \mathbf{x}^T \mathbf{x} - 2\mathbf{x}_n^T \mathbf{x} + \mathbf{x}_n^T \mathbf{x}_n \end{aligned}$$

Similar to (6.24)-(6.26), we use a nonlinear kernel $k(\mathbf{x}_n, \mathbf{x})$ to replace $\mathbf{x}_n^T \mathbf{x}$, which gives a general nonlinear nearest-neighbor classifier with cost function defined as:

$$k(\mathbf{x}, \mathbf{x}) + k(\mathbf{x}_n, \mathbf{x}_n) - 2k(\mathbf{x}_n, \mathbf{x})$$

Problem 6.4 Solution

To construct such a matrix, let us suppose the two eigenvalues are 1 and 2, and the matrix has form:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

Therefore, based on the definition of eigenvalue, we have two equations:

$$\begin{cases} (a-2)(d-2) = bc & (1) \\ (a-1)(d-1) = bc & (2) \end{cases}$$

(2)-(1), yielding:

$$a + d = 3$$

Therefore, we set $a = 4$ and $d = -1$. Then we substitute them into (1), and thus we see:

$$bc = -6$$

Finally, we choose $b = 3$ and $c = -2$. The constructed matrix is:

$$\begin{bmatrix} 4 & 3 \\ -2 & -1 \end{bmatrix}$$

Problem 6.5 Solution

Since $k_1(\mathbf{x}, \mathbf{x}')$ is a valid kernel, it can be written as:

$$k_1(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^T \phi(\mathbf{x}')$$

We can obtain:

$$k(\mathbf{x}, \mathbf{x}') = ck_1(\mathbf{x}, \mathbf{x}') = \left[\sqrt{c}\phi(\mathbf{x})\right]^T \left[\sqrt{c}\phi(\mathbf{x}')\right]$$

Therefore, (6.13) is a valid kernel. It is similar for (6.14):

$$k(\mathbf{x}, \mathbf{x}') = f(\mathbf{x})k_1(\mathbf{x}, \mathbf{x}')f(\mathbf{x}') = \left[f(\mathbf{x})\phi(\mathbf{x})\right]^T \left[f(\mathbf{x}')\phi(\mathbf{x}')\right]$$

Just as required.

Problem 6.6 Solution

We suppose $q(x)$ can be written as:

$$q(x) = a_n x^n + a_{n-1} x^{n-1} + ... + a_1 x + a_0$$

We now obtain:

$$k(\mathbf{x}, \mathbf{x}') = a_n k_1(\mathbf{x}, \mathbf{x}')^n + a_{n-1} k_1(\mathbf{x}, \mathbf{x}')^{n-1} + ... + a_1 k_1(\mathbf{x}, \mathbf{x}') + a_0$$

By repeatedly using (6.13), (6.17) and (6.18), we can easily verify $k(\mathbf{x}, \mathbf{x}')$ is a valid kernel. For (6.16), we can use Taylor expansion, and since the coefficients of Taylor expansion are all positive, we can similarly prove its validity.

Problem 6.7 Solution

To prove (6.17), we will use the property stated below (6.12). Since we know $k_1(\mathbf{x}, \mathbf{x}')$ and $k_2(\mathbf{x}, \mathbf{x}')$ are valid kernels, their Gram matrix $\mathbf{K}_1$ and $\mathbf{K}_2$

are both positive semidefinite. Given the relation (6.12), it can be easily shown $\mathbf{K} = \mathbf{K}_1 + \mathbf{K}_2$ is also positive semidefinite and thus $k(\mathbf{x},\mathbf{x}')$ is also a valid kernel.

To prove (6.18), we assume the map function for kernel $k_1(\mathbf{x},\mathbf{x}')$ is $\boldsymbol{\phi}^{(1)}(\mathbf{x})$, and similarly $\boldsymbol{\phi}^{(2)}(\mathbf{x})$ for $k_2(\mathbf{x},\mathbf{x}')$. Moreover, we further assume the dimension of $\boldsymbol{\phi}^{(1)}(\mathbf{x})$ is M, and $\boldsymbol{\phi}^{(2)}(\mathbf{x})$ is N. We expand $k(\mathbf{x},\mathbf{x}')$ based on (6.18):

$$\begin{aligned} k(\mathbf{x},\mathbf{x}') &= k_1(\mathbf{x},\mathbf{x}')k_2(\mathbf{x},\mathbf{x}') \\ &= \boldsymbol{\phi}^{(1)}(\mathbf{x})^T\boldsymbol{\phi}^{(1)}(\mathbf{x}')\boldsymbol{\phi}^{(2)}(\mathbf{x})^T\boldsymbol{\phi}^{(2)}(\mathbf{x}') \\ &= \sum_{i=1}^{M}\phi_i^{(1)}(\mathbf{x})\phi_i^{(1)}(\mathbf{x}')\sum_{j=1}^{N}\phi_j^{(2)}(\mathbf{x})\phi_j^{(2)}(\mathbf{x}') \\ &= \sum_{i=1}^{M}\sum_{j=1}^{N}\left[\phi_i^{(1)}(\mathbf{x})\phi_j^{(2)}(\mathbf{x})\right]\left[\phi_i^{(1)}(\mathbf{x}')\phi_j^{(2)}(\mathbf{x}')\right] \\ &= \sum_{k=1}^{MN}\phi_k(\mathbf{x})\phi_k(\mathbf{x}') = \boldsymbol{\phi}(\mathbf{x})^T\boldsymbol{\phi}(\mathbf{x}') \end{aligned}$$

where $\phi_i^{(1)}(\mathbf{x})$ is the ith element of $\boldsymbol{\phi}^{(1)}(\mathbf{x})$, and $\phi_j^{(2)}(\mathbf{x})$ is the jth element of $\boldsymbol{\phi}^{(2)}(\mathbf{x})$. To be more specific, we have proved that $k(\mathbf{x},\mathbf{x}')$ can be written as $\boldsymbol{\phi}(\mathbf{x})^T\boldsymbol{\phi}(\mathbf{x}')$. Here $\boldsymbol{\phi}(\mathbf{x})$ is a $MN\times 1$ column vector, and the kth ($k = 1,2,...,MN$) element is given by $\phi_i^{(1)}(\mathbf{x}) \times \phi_j^{(2)}(\mathbf{x})$. What's more, we can also express i,j in terms of k:

$$i = (k-1)\oslash N + 1 \quad \text{and} \quad j = (k-1)\odot N + 1$$

where $\oslash$ and $\odot$ means integer division and remainder, respectively.

Problem 6.8 Solution

For (6.19) we suppose $k_3(\mathbf{x},\mathbf{x}') = \boldsymbol{g}(\mathbf{x})^T\boldsymbol{g}(\mathbf{x}')$, and thus we have:

$$k(\mathbf{x},\mathbf{x}') = k_3(\boldsymbol{\phi}(\mathbf{x}),\boldsymbol{\phi}(\mathbf{x}')) = \boldsymbol{g}(\boldsymbol{\phi}(\mathbf{x}))^T\boldsymbol{g}(\boldsymbol{\phi}(\mathbf{x}')) = \boldsymbol{f}(\mathbf{x})^T\boldsymbol{f}(\mathbf{x}')$$

where we have denoted $\boldsymbol{g}(\boldsymbol{\phi}(\mathbf{x})) = \boldsymbol{f}(\mathbf{x})$ and now it is obvious that (6.19) holds. To prove (6.20), we suppose $\mathbf{x}$ is a $N\times 1$ column vector and $\mathbf{A}$ is a $N\times N$ symmetric positive semidefinite matrix. We know that $\mathbf{A}$ can be decomposed to $\mathbf{QBQ}^T$. Here $\mathbf{Q}$ is a $N\times N$orthogonal matrix, and $\mathbf{B}$ is a $N\times N$ diagonal matrix whose elements are no less than 0. Now we can derive:

$$\begin{aligned} k(\mathbf{x},\mathbf{x}') &= \mathbf{x}^T\mathbf{A}\mathbf{x}' = \mathbf{x}^T\mathbf{QBQ}^T\mathbf{x}' = (\mathbf{Q}^T\mathbf{x})^T\mathbf{B}(\mathbf{Q}^T\mathbf{x}') = \mathbf{y}^T\mathbf{B}\mathbf{y}' \\ &= \sum_{i=1}^{N}B_{ii}y_iy_i^{'} = \sum_{i=1}^{N}(\sqrt{B_{ii}}y_i)(\sqrt{B_{ii}}y_i^{'}) = \boldsymbol{\phi}(\mathbf{x})^T\boldsymbol{\phi}(\mathbf{x}') \end{aligned}$$

To be more specific, we have proved that $k(\mathbf{x},\mathbf{x}') = \boldsymbol{\phi}(\mathbf{x})^T\boldsymbol{\phi}(\mathbf{x}')$, and here $\boldsymbol{\phi}(\mathbf{x})$ is a $N\times 1$ column vector, whose ith ($i = 1,2,...,N$) element is given by $\sqrt{B_{ii}}y_i$, i.e., $\sqrt{B_{ii}}(\mathbf{Q}^T\mathbf{x})_i$.

Problem 6.9 Solution

To prove (6.21), let's first expand the expression:

$$\begin{aligned} k(\mathbf{x},\mathbf{x}') &= k_a(\mathbf{x}_a,\mathbf{x}_a^{'}) + k_b(\mathbf{x}_b,\mathbf{x}_b^{'}) \\ &= \sum_{i=1}^{M}\phi_i^{(a)}(\mathbf{x}_a)\phi_i^{(a)}(\mathbf{x}_a^{'}) + \sum_{j=1}^{N}\phi_i^{(b)}(\mathbf{x}_b)\phi_i^{(b)}(\mathbf{x}_b^{'}) \\ &= \sum_{k=1}^{M+N}\phi_k(\mathbf{x})\phi_k(\mathbf{x}^{'}) = \boldsymbol{\phi}(\mathbf{x})^T\boldsymbol{\phi}(\mathbf{x}') \end{aligned}$$

where we have assumed the dimension of $\mathbf{x}_a$ is M and the dimension of $\mathbf{x}_b$ is N. The mapping function $\boldsymbol{\phi}(\mathbf{x})$ is a $(M+N)\times 1$ column vector, whose kth ($k = 1,2,...,M+N$) element $\phi_k(\mathbf{x})$ is:

$$\phi_k(\mathbf{x}) = \begin{cases} \phi_k^{(a)}(\mathbf{x}) & 1 \le k \le M \\ \phi_{k-M}^{(b)}(\mathbf{x}_a) & M+1 \le k \le M+N \end{cases}$$

(6.22) is quite similar to (6.18). We follow the same procedure:

$$\begin{aligned} k(\mathbf{x},\mathbf{x}') &= k_a(\mathbf{x}_a,\mathbf{x}_a^{'})k_b(\mathbf{x}_b,\mathbf{x}_b^{'}) \\ &= \sum_{i=1}^{M}\phi_i^{(a)}(\mathbf{x}_a)\phi_i^{(a)}(\mathbf{x}_a^{'})\sum_{j=1}^{N}\phi_j^{(b)}(\mathbf{x}_b)\phi_j^{(b)}(\mathbf{x}_b^{'}) \\ &= \sum_{i=1}^{M}\sum_{j=1}^{N}\left[\phi_i^{(a)}(\mathbf{x}_a)\phi_j^{(b)}(\mathbf{x}_b)\right]\left[\phi_i^{(a)}(\mathbf{x}_a^{'})\phi_j^{(b)}(\mathbf{x}_b^{'})\right] \\ &= \sum_{k=1}^{MN}\phi_k(\mathbf{x})\phi_k(\mathbf{x}') = \boldsymbol{\phi}(\mathbf{x})^T\boldsymbol{\phi}(\mathbf{x}') \end{aligned}$$

By analogy to (6.18), the mapping function $\boldsymbol{\phi}(\mathbf{x})$ is a $MN\times 1$ column vector, whose kth ($k = 1,2,...,MN$) element $\phi_k(\mathbf{x})$ is:

$$\phi_k(\mathbf{x}) = \phi_i^{(a)}(\mathbf{x}_a)\times\phi_j^{(b)}(\mathbf{x}_b)$$

To be more specific, $\mathbf{x}_a$ is the sub-vector of $\mathbf{x}$ made up of the first M element of $\mathbf{x}$, and $\mathbf{x}_b$ is the sub-vector of $\mathbf{x}$ made up of the last N element of $\mathbf{x}$. What's more, we can also express i,j in terms of k:

$$i = (k-1)\oslash N+1 \quad \text{and} \quad j = (k-1)\odot N+1$$

where $\oslash$ and $\odot$ means integer division and remainder, respectively.

Problem 6.10 Solution

According to (6.9), we have:

$$y(\mathbf{x}) = \mathbf{k}(\mathbf{x})^T(\mathbf{K}+\lambda\mathbf{I}_N)^{-1}\mathbf{t} = \mathbf{k}(\mathbf{x})^T\mathbf{a} = \sum_{n=1}^{N}f(\mathbf{x}_n)\cdot f(\mathbf{x})\cdot a_n = \left[\sum_{n=1}^{N}f(\mathbf{x}_n)\cdot\cdot a_n\right]f(\mathbf{x})$$

We see that if we choose $k(\mathbf{x},\mathbf{x}') = f(\mathbf{x})f(\mathbf{x}')$ we will always find a solution $y(\mathbf{x})$ proportional to $f(\mathbf{x})$.

Problem 6.11 Solution

We follow the hint.

$$\begin{aligned} k(\mathbf{x},\mathbf{x}') &= \exp(-\mathbf{x}^T\mathbf{x}/2\sigma^2)\cdot\exp(\mathbf{x}^T\mathbf{x}'/\sigma^2)\cdot\exp(-(\mathbf{x}')^T\mathbf{x}'/2\sigma^2) \\ &= \exp(-\mathbf{x}^T\mathbf{x}/2\sigma^2)\cdot\left(1+\frac{\mathbf{x}^T\mathbf{x}'}{\sigma^2}+\frac{(\frac{\mathbf{x}^T\mathbf{x}'}{\sigma^2})^2}{2!}+\cdots\right)\cdot\exp(-(\mathbf{x}')^T\mathbf{x}'/2\sigma^2) \\ &= \boldsymbol{\phi}(\mathbf{x})^T\boldsymbol{\phi}(\mathbf{x}') \end{aligned}$$

where $\boldsymbol{\phi}(\mathbf{x})$ is a column vector with infinite dimension. To be more specific, (6.12) gives a simple example on how to decompose $(\mathbf{x}^T\mathbf{x}')^2$. In our case, we can also decompose $(\mathbf{x}^T\mathbf{x}')^k, k = 1,2,...,\infty$ in the similar way. However, since $k\to\infty$, i.e., the decomposition will consist monomials with infinite degree. Thus, there will be infinite terms in the decomposition and the feature mapping function $\boldsymbol{\phi}(\mathbf{x})$ will have infinite dimension.

Problem 6.12 Solution

First, let's explain the problem a little bit. According to (6.27), what we need to prove here is:

$$k(A_1,A_2) = 2^{|A_1\cap A_2|} = \boldsymbol{\phi}(A_1)^T\boldsymbol{\phi}(A_2)$$

The biggest difference from the previous problem is that $\boldsymbol{\phi}(A)$ is a $2^{|D|}\times 1$ column vector and instead of indexed by $1,2,...,2^{|D|}$ here we index it by $\{U|U\subseteq D\}$ (Note that $\{U|U\subseteq D\}$ is all the possible subsets of D and thus there are $2^{|D|}$ elements in total). Therefore, according to (6.95), we can obtain:

$$\boldsymbol{\phi}(A_1)^T\boldsymbol{\phi}(A_2) = \sum_{U\subseteq D}\phi_U(A_1)\phi_U(A_2)$$

By using the summation, we actually iterate through all the possible subsets of D. If and only if the current iterating subset U is a subset of both A_1 and A_2 simultaneously, the current adding term equals to 1. Therefore, we actually count how many subsets of D is in the intersection of A_1 and A_2.

Moreover, since A_1 and A_2 are both defined in the subset space of D, what we have deduced above can be written as:

$$\boldsymbol{\phi}(A_1)^T\boldsymbol{\phi}(A_2) = 2^{|A_1\cap A_2|}$$

Just as required.

Problem 6.13 Solution Wait for update

Problem 6.14 Solution

Since the covariance matrix $\mathbf{S}$ is fixed, according to (6.32) we can obtain:

$$\mathbf{g}(\boldsymbol{\mu},\mathbf{x}) = \nabla_{\boldsymbol{\mu}} \ln p(\mathbf{x}|\boldsymbol{\mu}) = \frac{\partial}{\partial \boldsymbol{\mu}} \left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{S}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \right) = \mathbf{S}^{-1}(\mathbf{x}-\boldsymbol{\mu})$$

Therefore, according to (6.34), we can obtain:

$$\mathbf{F} = \mathbb{E}_{\mathbf{x}} \left[\mathbf{g}(\boldsymbol{\mu},\mathbf{x}) \mathbf{g}(\boldsymbol{\mu},\mathbf{x})^T \right] = \mathbf{S}^{-1} \mathbb{E}_{\mathbf{x}} \left[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T \right] \mathbf{S}^{-1}$$

Since $\mathbf{x} \sim \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\mathbf{S})$, we have:

$$\mathbb{E}_{\mathbf{x}} \left[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T \right] = \mathbf{S}$$

So we obtain $\mathbf{F} = \mathbf{S}^{-1}$ and then according to (6.33), we have:

$$k(\mathbf{x},\mathbf{x}') = \mathbf{g}(\boldsymbol{\mu},\mathbf{x})^T \mathbf{F}^{-1} \mathbf{g}(\boldsymbol{\mu},\mathbf{x}') = (\mathbf{x}-\boldsymbol{\mu})^T \mathbf{S}^{-1}(\mathbf{x}'-\boldsymbol{\mu})$$

Problem 6.15 Solution

We rewrite the problem. What we are required to prove is that the Gram matrix $\mathbf{K}$:

$$\mathbf{K} = \begin{bmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{bmatrix},$$

where k_{ij} $(i,j = 1,2)$ is short for $k(x_i,x_j)$, should be positive semidefinite. A positive semidefinite matrix should have positive determinant, i.e.,

$$k_{12}k_{21} \leq k_{11}k_{22}.$$

Using the symmetric property of kernel, i.e.,$k_{12} = k_{21}$, we obtain what has been required.

Problem 6.16 Solution

Based on the total derivative of function f, we have:

$$f\left((\mathbf{w}+\Delta\mathbf{w})^T\boldsymbol{\phi}_1,(\mathbf{w}+\Delta\mathbf{w})^T\boldsymbol{\phi}_2,...,(\mathbf{w}+\Delta\mathbf{w})^T\boldsymbol{\phi}_N\right) = \sum_{n=1}^{N} \frac{\partial f}{\partial(\mathbf{w}^T\boldsymbol{\phi}_n)} \cdot \Delta\mathbf{w}^T\boldsymbol{\phi}_n$$

Which can be further written as:

$$f\left((\mathbf{w}+\Delta\mathbf{w})^T\boldsymbol{\phi}_1,(\mathbf{w}+\Delta\mathbf{w})^T\boldsymbol{\phi}_2,...,(\mathbf{w}+\Delta\mathbf{w})^T\boldsymbol{\phi}_N\right) = \left[\sum_{n=1}^{N} \frac{\partial f}{\partial(\mathbf{w}^T\boldsymbol{\phi}_n)} \cdot \boldsymbol{\phi}_n^T\right] \Delta\mathbf{w}$$

Note that here $\boldsymbol{\phi}_n$ is short for $\boldsymbol{\phi}(\mathbf{x}_n)$. Based on the equation above, we can obtain:

$$\nabla_{\mathbf{w}} f = \sum_{n=1}^{N} \frac{\partial f}{\partial(\mathbf{w}^T\boldsymbol{\phi}_n)} \cdot \boldsymbol{\phi}_n^T$$

Now we focus on the derivative of function g with respect to $\mathbf{w}$:

$$\nabla_{\mathbf{w}} g = \frac{\partial g}{\partial(\mathbf{w}^T\mathbf{w})} \cdot 2\mathbf{w}^T$$

In order to find the optimal $\mathbf{w}$, we set the derivative of J with respect to $\mathbf{w}$ equal to $\mathbf{0}$, yielding:

$$\nabla_{\mathbf{w}} J = \nabla_{\mathbf{w}} f + \nabla_{\mathbf{w}} g = \sum_{n=1}^{N} \frac{\partial f}{\partial(\mathbf{w}^T\boldsymbol{\phi}_n)} \cdot \boldsymbol{\phi}_n^T + \frac{\partial g}{\partial(\mathbf{w}^T\mathbf{w})} \cdot 2\mathbf{w}^T = \mathbf{0}$$

Rearranging the equation above, we can obtain:

$$\mathbf{w} = \frac{1}{2a} \sum_{n=1}^{N} \frac{\partial f}{\partial(\mathbf{w}^T\boldsymbol{\phi}_n)} \cdot \boldsymbol{\phi}_n$$

Where we have defined: $a = 1 \div \frac{\partial g}{\partial(\mathbf{w}^T\mathbf{w})}$, and since g is a monotonically increasing function, we have $a > 0$.

Problem 6.17 Solution

We consider a variation in the function $y(\mathbf{x})$ of the form:

$$y(\mathbf{x}) \rightarrow y(\mathbf{x}) + \epsilon\eta(\mathbf{x})$$

Substituting it into (6.39) yields:

$$\begin{aligned} E[y+\epsilon\eta] &= \frac{1}{2}\sum_{n=1}^{N}\int \{y+\epsilon\eta-t_n\}^2 v(\boldsymbol{\xi})d\boldsymbol{\xi} \\ &= \frac{1}{2}\sum_{n=1}^{N}\int \{(y-t_n)^2 + 2\cdot(\epsilon\eta)\cdot(y-t_n) + (\epsilon\eta)^2\} v(\boldsymbol{\xi})d\boldsymbol{\xi} \\ &= E[y] + \epsilon\sum_{n=1}^{N}\int \{y-t_n\}\eta v d\boldsymbol{\xi} + O(\epsilon^2) \end{aligned}$$

Note that here y is short for $y(\mathbf{x}_n+\boldsymbol{\xi})$, η is short for $\eta(\mathbf{x}_n+\boldsymbol{\xi})$ and v is short for $v(\boldsymbol{\xi})$ respectively. Several clarifications must be made here. What we have done is that we vary the function y by a little bit (i.e., $\epsilon\eta$) and then we expand the corresponding error with respect to the small variation ϵ. The coefficient before ϵ is actually the first derivative of the error $E[y+\epsilon\eta]$ with respect to ϵ at $\epsilon = 0$. Since we know that y is the optimal function that can make E the smallest, the first derivative of the error $E[y+\epsilon\eta]$ should equal to zero at $\epsilon = 0$, which gives:

$$\sum_{n=1}^{N}\int \{y(\mathbf{x}_n+\boldsymbol{\xi}) - t_n\}\eta(\mathbf{x}_n+\boldsymbol{\xi})v(\boldsymbol{\xi})d\boldsymbol{\xi} = 0$$

Now we are required to find a function y that can satisfy the equation above no matter what η is. We choose:

$$\eta(\mathbf{x}) = \delta(\mathbf{x}-\mathbf{z})$$

This allows us to evaluate the integral:

$$\sum_{n=1}^{N}\int \{y(\mathbf{x}_n+\boldsymbol{\xi})-t_n\}\eta(\mathbf{x}_n+\boldsymbol{\xi})v(\boldsymbol{\xi})d\boldsymbol{\xi} = \sum_{n=1}^{N}\{y(\mathbf{z})-t_n\}v(\mathbf{z}-\mathbf{x}_n)$$

We set it to zero and rearrange it, which finally gives (6.40) just as required.

Problem 6.18 Solution

According to the main text below Eq (6.48), we know that $f(x,t)$, i.e., $f(\mathbf{z})$, follows a zero-mean isotropic Gaussian:

$$f(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0},\sigma^2\mathbf{I})$$

Then $f(x-x_m,t-t_m)$, i.e., $f(\mathbf{z}-\mathbf{z}_m)$ should also satisfy a Gaussian distribution:

$$f(\mathbf{z}-\mathbf{z}_m) = \mathcal{N}(\mathbf{z}|\mathbf{z}_m,\sigma^2\mathbf{I})$$

Where we have defined:

$$\mathbf{z}_m = (x_m,t_m)$$

The integral $\int f(\mathbf{z}-\mathbf{z}_m)dt$ corresponds to the marginal distribution with respect to the remaining variable x and, thus, we obtain:

$$\int f(\mathbf{z}-\mathbf{z}_m)\,dt = \mathcal{N}(x|x_m,\sigma^2)$$

We substitute all the expressions into Eq (6.48), which gives:

$$\begin{aligned}
p(t|x) &= \frac{p(t,x)}{\int p(t,x)dt} = \frac{\sum_n \mathcal{N}(\mathbf{z}|\mathbf{z}_m,\sigma^2\mathbf{I})}{\sum_m \mathcal{N}(x|x_m,\sigma^2)} \\
&= \frac{\sum_n \frac{1}{2\pi\sigma^2} exp\left(-\frac{1}{2}(\mathbf{z}-\mathbf{z}_n)^T(\sigma^2\mathbf{I})^{-1}(\mathbf{z}-\mathbf{z}_n)\right)}{\sum_m \frac{1}{(2\pi\sigma^2)^{1/2}} exp\left(-\frac{1}{2\sigma^2}(x-x_m)^2\right)} \\
&= \frac{\sum_n \frac{1}{2\pi\sigma^2} exp\left(-\frac{1}{2\sigma^2}(x-x_n)^2\right) exp\left(-\frac{1}{2\sigma^2}(t-t_n)^2\right)}{\sum_m \frac{1}{(2\pi\sigma^2)^{1/2}} exp\left(-\frac{1}{2\sigma^2}(x-x_m)^2\right)} \\
&= \sum_n \frac{\frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{1}{2\sigma^2}(x-x_n)^2\right)}{\sum_m \frac{1}{(2\pi\sigma^2)^{1/2}} exp\left(-\frac{1}{2\sigma^2}(x-x_m)^2\right)} \cdot \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{1}{2\sigma^2}(t-t_n)^2\right) \\
&= \sum_n \pi_n \cdot \mathcal{N}(t|t_n,\sigma^2)
\end{aligned}$$

Where we have defined:

$$\pi_n = \frac{exp\left(-\frac{1}{2\sigma^2}(x-x_n)^2\right)}{\sum_m exp\left(-\frac{1}{2\sigma^2}(x-x_m)^2\right)}$$

We also observe that:

$$\sum_n \pi_n = 1$$

Therefore, the conditional distribution $p(t|x)$ is given by a Gaussian Mixture. Similarly, we attempt to find a specific form for Eq (6.46):

$$\begin{aligned} k(x,x_n) &= \frac{\int f(x-x_n,t)\,dt}{\sum_m \int f(x-x_m,t)\,dt} \\ &= \frac{\mathcal{N}(x|x_n,\sigma^2)}{\sum_m \mathcal{N}(x|x_m,\sigma^2)} \\ &= \pi_n \end{aligned}$$

In other words, the conditional distribution can be more precisely written as:

$$p(t|x) = \sum_n k(x,x_n)\cdot\mathcal{N}(t|t_n,\sigma^2)$$

Thus its mean is given by:

$$\mathbb{E}[t|x] = \sum_n k(x,x_n)\cdot t_n$$

Its variance is given by:

$$\begin{aligned} \text{var}[t|x] &= \mathbb{E}[(t|x)^2] - \mathbb{E}[t|x]^2 \\ &= \sum_n k(x,x_n)\cdot(t_n^2+\sigma^2) - \left(\sum_n k(x,x_n)\cdot t_n\right)^2 \end{aligned}$$

Problem 6.19 Solution

Similar to Prob.6.17, it is straightforward to show that:

$$y(\mathbf{x}) = \sum_n t_n\, k(\mathbf{x},\mathbf{x}_n)$$

Where we have defined:

$$k(\mathbf{x},\mathbf{x}_n) = \frac{g(\mathbf{x}_n-\mathbf{x})}{\sum_n g(\mathbf{x}_n-\mathbf{x})}$$

Problem 6.20 Solution

Since we know that $\mathbf{t}_{N+1} = (t_1,t_2,...,t_N,t_{N+1})^T$ follows a Gaussian distribution, i.e., $\mathbf{t}_{N+1} \sim \mathcal{N}(\mathbf{t}_{N+1}|\mathbf{0},\mathbf{C}_{N+1})$ given in Eq (6.64), if we rearrange its

order by putting the last element (i.e., t_{N+1}) to the first position, denoted as $\bar{\mathbf{t}}_{N+1}$, it should also satisfy a Gaussian distribution:

$$\bar{\mathbf{t}}_{N+1} = (t_{N+1}, t_1, ..., t_2, t_N)^T \sim \mathcal{N}(\bar{\mathbf{t}}_{N+1}|\mathbf{0}, \bar{\mathbf{C}}_{N+1})$$

Where we have defined:

$$\bar{\mathbf{C}}_{N+1} = \begin{pmatrix} c & \mathbf{k}^T \\ \mathbf{k} & \mathbf{C}_N \end{pmatrix}$$

Where $\mathbf{k}$ and c have been given in the main text below Eq (6.65). The conditional distribution $p(t_{N+1}|\mathbf{t}_N)$ should also be a Gaussian. By analogy to Eq (2.94)-(2.98), we can simply treat t_{N+1} as $\mathbf{x}_a$, $\mathbf{t}_N$ as $\mathbf{x}_b$, c as $\mathbf{\Sigma}_{aa}$, $\mathbf{k}$ as $\mathbf{\Sigma}_{ba}$, $\mathbf{k}^T$ as $\mathbf{\Sigma}_{ab}$ and $\mathbf{C}_N$ as $\mathbf{\Sigma}_{bb}$. Substituting them into Eq (2.79) and Eq (2.80) yields:

$$\mathbf{\Lambda}_{aa} = (c - \mathbf{k}^T \mathbf{C}_N^{-1} \mathbf{k})^{-1}$$

And:

$$\mathbf{\Lambda}_{ab} = -(c - \mathbf{k}^T \mathbf{C}_N^{-1} \mathbf{k})^{-1} \mathbf{k}^T \mathbf{C}_N^{-1}$$

Then we substitute them into Eq (2.96) and (2.97), yields:

$$p(t_{N+1}|\mathbf{t}_N) = \mathcal{N}(\boldsymbol{\mu}_{a|b}, \mathbf{\Lambda}_{aa}^{-1})$$

For its mean $\boldsymbol{\mu}_{a|b}$, we have:

$$\begin{aligned} \boldsymbol{\mu}_{a|b} &= 0 - \left(c - \mathbf{k}^T \mathbf{C}_N^{-1} \mathbf{k}\right) \cdot \left[-(c - \mathbf{k}^T \mathbf{C}_N^{-1} \mathbf{k})^{-1} \mathbf{k}^T \mathbf{C}_N^{-1}\right] \cdot (\mathbf{t}_N - \mathbf{0}) \\ &= \mathbf{k}^T \mathbf{C}_N^{-1} \mathbf{t}_N = m(\mathbf{x}_{N+1}) \end{aligned}$$

Similarly, for its variance $\mathbf{\Lambda}_{aa}^{-1}$ (Note that here since t_{N+1} is a scalar, the mean and the covariance matrix actually degenerate to one dimension case), we have:

$$\mathbf{\Lambda}_{aa}^{-1} = c - \mathbf{k}^T \mathbf{C}_N^{-1} \mathbf{k} = \sigma^2(\mathbf{x}_{N+1})$$

Problem 6.21 Solution

We follow the hint beginning by verifying the mean. We write Eq (6.62) in a matrix form:

$$\mathbf{C}_N = \frac{1}{\alpha} \mathbf{\Phi} \mathbf{\Phi}^T + \beta^{-1} \mathbf{I}_N$$

Where we have used Eq (6.54). Here $\mathbf{\Phi}$ is the design matrix defined below Eq (6.51) and $\mathbf{I}_N$ is an identity matrix. Before we use Eq (6.66), we need to obtain $\mathbf{k}$:

$$\begin{aligned} \mathbf{k} &= [k(\mathbf{x}_1, \mathbf{x}_{N+1}), k(\mathbf{x}_2, \mathbf{x}_{N+1}), ..., k(\mathbf{x}_N, \mathbf{x}_{N+1})]^T \\ &= \frac{1}{\alpha} [\boldsymbol{\phi}(\mathbf{x}_1)^T \boldsymbol{\phi}(\mathbf{x}_{N+1}), \boldsymbol{\phi}(\mathbf{x}_2)^T \boldsymbol{\phi}(\mathbf{x}_{N+1}), ..., \boldsymbol{\phi}(\mathbf{x}_n)^T \boldsymbol{\phi}(\mathbf{x}_{N+1})]^T \\ &= \frac{1}{\alpha} \mathbf{\Phi} \boldsymbol{\phi}(\mathbf{x}_{N+1})^T \end{aligned}$$

Now we substitute all the expressions into Eq (6.66), yielding:

$$m(\mathbf{x}_{N+1}) = \alpha^{-1}\boldsymbol{\phi}(\mathbf{x}_{N+1})^T\boldsymbol{\Phi}^T\left[\alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^T + \beta^{-1}\mathbf{I}_N\right]^{-1}\mathbf{t}$$

Next using matrix identity (C.6), we obtain:

$$\boldsymbol{\Phi}^T\left[\alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^T + \beta^{-1}\mathbf{I}_N\right]^{-1} = \alpha\beta\left[\beta\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \alpha\mathbf{I}_M\right]^{-1}\boldsymbol{\Phi}^T = \alpha\beta\mathbf{S}_N\boldsymbol{\Phi}^T$$

Where we have used Eq (3.54). Substituting it into $\mathbf{m}(\mathbf{x}_{N+1})$, we obtain:

$$m(\mathbf{x}_{N+1}) = \beta\boldsymbol{\phi}(\mathbf{x}_{N+1})^T\mathbf{S}_N\boldsymbol{\Phi}^T\mathbf{t} = <\boldsymbol{\phi}(\mathbf{x}_{N+1})^T, \beta\mathbf{S}_N\boldsymbol{\Phi}^T\mathbf{t}>$$

Where $<\cdot,\cdot>$ represents the inner product. Comparing the result above with Eq (3.58), (3.54) and (3.53), we conclude that the means are equal. It is similar for the variance. We substitute c, $\mathbf{k}$ and $\mathbf{C}_N$ into Eq (6.67). Then we simplify the expression using matrix identity (C.7). Finally, we will observe that it is equal to Eq (3.59).

Problem 6.22 Solution

Based on Eq (6.64) and (6.65), We first write down the joint distribution for $\mathbf{t}_{N+L} = [t_1(\mathbf{x}), t_2(\mathbf{x}), ..., t_{N+L}(\mathbf{x})]^T$:

$$p(\mathbf{t}_{N+L}) = \mathcal{N}(\mathbf{t}_{N+L}|\mathbf{0}, \mathbf{C}_{N+L})$$

Where $\mathbf{C}_{N+L}$ is similarly given by:

$$\mathbf{C}_{N+L} = \begin{pmatrix} \mathbf{C}_{1,N} & \mathbf{K} \\ \mathbf{K}^T & \mathbf{C}_{N+1,N+L} \end{pmatrix}$$

The expression above has already implicitly divided the vector $\mathbf{t}_{N+L}$ into two parts. Similar to Prob.6.20, for later simplicity we rearrange the order of $\mathbf{t}_{N+L}$denoted as $\bar{\mathbf{t}}_{N+L} = [t_{N+1}, ..., t_{N+L}, t_1, ..., t_N]^T$. Moreover, $\bar{\mathbf{t}}_{N+L}$ should also follows a Gaussian distribution:

$$p(\bar{\mathbf{t}}_{N+L}) = \mathcal{N}(\bar{\mathbf{t}}_{N+L}|\mathbf{0}, \bar{\mathbf{C}}_{N+L})$$

Where we have defined:

$$\bar{\mathbf{C}}_{N+L} = \begin{pmatrix} \mathbf{C}_{N+1,N+L} & \mathbf{K}^T \\ \mathbf{K} & \mathbf{C}_{1,N} \end{pmatrix}$$

Now we use Eq (2.94)-(2.98) and Eq (2.79)-(2.80) to derive the conditional distribution, beginning by calculate $\boldsymbol{\Lambda}_{aa}$:

$$\boldsymbol{\Lambda}_{aa} = (\mathbf{C}_{N+1,N+L} - \mathbf{K}^T \cdot \mathbf{C}_{1,N}^{-1} \cdot \mathbf{K})^{-1}$$

and $\boldsymbol{\Lambda}_{ab}$:

$$\boldsymbol{\Lambda}_{ab} = -(\mathbf{C}_{N+1,N+L} - \mathbf{K}^T \cdot \mathbf{C}_{1,N}^{-1} \cdot \mathbf{K})^{-1} \cdot \mathbf{K}^T \cdot \mathbf{C}_{1,N}^{-1}$$

Now we can obtain:

$$p(t_{N+1},...,t_{N+L}|\mathbf{t}_N) = \mathcal{N}(\boldsymbol{\mu}_{a|b}, \boldsymbol{\Lambda}_{aa}^{-1})$$

Where we have defined:

$$\boldsymbol{\mu}_{a|b} = \mathbf{0} + \mathbf{K}^T \cdot \mathbf{C}_{1,N}^{-1} \cdot \mathbf{t}_N = \mathbf{K}^T \cdot \mathbf{C}_{1,N}^{-1} \cdot \mathbf{t}_N$$

If now we want to find the conditional distribution $p(t_j|\mathbf{t}_N)$, where $N+1 \leq j \leq N+L$, we only need to find the corresponding entry in the mean (i.e., the $(j-N)$-th entry) and covariance matrix (i.e., the $(j-N)$-th diagonal entry) of $p(t_{N+1},...,t_{N+L}|\mathbf{t}_N)$. In this case, it will degenerate to Eq (6.66) and (6.67) just as required.

Problem 6.24 Solution

By definition, we only need to prove that for arbitrary vector $\mathbf{x} \neq \mathbf{0}$, $\mathbf{x}^T\mathbf{W}\mathbf{x}$ is positive. Here suppose that $\mathbf{W}$ is a $M \times M$ matrix. We expand the multiplication:

$$\mathbf{x}^T\mathbf{W}\mathbf{x} = \sum_{i=1}^{M}\sum_{j=1}^{M} W_{ij} \cdot x_i \cdot x_j = \sum_{i=1}^{M} W_{ii} \cdot x_i^2$$

where we have used the fact that $\mathbf{W}$ is a diagonal matrix. Since $W_{ii} > 0$, we obtain $\mathbf{x}^T\mathbf{W}\mathbf{x} > 0$ just as required. Suppose we have two positive definite matrix, denoted as $\mathbf{A}_1$ and $\mathbf{A}_2$, i.e., for arbitrary vector $\mathbf{x}$, we have $\mathbf{x}^T\mathbf{A}_1\mathbf{x} > 0$ and $\mathbf{x}^T\mathbf{A}_2\mathbf{x} > 0$. Therefore, we can obtain:

$$\mathbf{x}^T(\mathbf{A}_1 + \mathbf{A}_2)\mathbf{x} = \mathbf{x}^T\mathbf{A}_1\mathbf{x} + \mathbf{x}^T\mathbf{A}_2\mathbf{x} > 0$$

Just as required.

Problem 6.25 Solution

Based on Newton-Raphson formula, Eq(6.81) and Eq(6.82), we have:

$$\begin{aligned}
\mathbf{a}_N^{new} &= \mathbf{a}_N - (-\mathbf{W}_N - \mathbf{C}_N^{-1})^{-1}(\mathbf{t}_N - \sigma_N - \mathbf{C}_N^{-1}\mathbf{a}_N) \\
&= \mathbf{a}_N + (\mathbf{W}_N + \mathbf{C}_N^{-1})^{-1}(\mathbf{t}_N - \sigma_N - \mathbf{C}_N^{-1}\mathbf{a}_N) \\
&= (\mathbf{W}_N + \mathbf{C}_N^{-1})^{-1}\left[(\mathbf{W}_N + \mathbf{C}_N^{-1})\mathbf{a}_N + \mathbf{t}_N - \sigma_N - \mathbf{C}_N^{-1}\mathbf{a}_N\right] \\
&= \mathbf{C}_N\mathbf{C}_N^{-1}(\mathbf{W}_N + \mathbf{C}_N^{-1})^{-1}(\mathbf{t}_N - \sigma_N + \mathbf{W}_N\mathbf{a}_N) \\
&= \mathbf{C}_N(\mathbf{C}_N\mathbf{W}_N + \mathbf{I})^{-1}(\mathbf{t}_N - \sigma_N + \mathbf{W}_N\mathbf{a}_N)
\end{aligned}$$

Just as required.

Problem 6.26 Solution

Using Eq(6.77), (6.78) and (6.86), we can obtain:

$$\begin{aligned}
p(a_{N+1}|\mathbf{t}_N) &= \int p(a_{N+1}|\mathbf{a}_N)p(\mathbf{a}_N|\mathbf{t}_N)d\mathbf{a}_N \\
&= \int N(a_{N+1}|\mathbf{k}^T\mathbf{C}_N^{-1}\mathbf{a}_N, c - \mathbf{k}^T\mathbf{C}_N^{-1}\mathbf{k}) \cdot N(\mathbf{a}_N|\mathbf{a}_N^\star, \mathbf{H}^{-1})d\mathbf{a}_N
\end{aligned}$$

By analogy to Eq (2.115), i.e.,

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

We can obtain:

$$p(a_{N+1}|\mathbf{t}_N) = N(\mathbf{A}\boldsymbol{\mu}+\mathbf{b}, \mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T) \qquad (*)$$

Where we have defined:

$$\mathbf{A} = \mathbf{k}^T\mathbf{C}_N^{-1}, \mathbf{b} = \mathbf{0}, \mathbf{L}^{-1} = c - \mathbf{k}^T\mathbf{C}_N^{-1}\mathbf{k}$$

And

$$\boldsymbol{\mu} = \mathbf{a}_N^\star, \boldsymbol{\Lambda} = \mathbf{H}$$

Therefore, the mean is given by:

$$\mathbf{A}\boldsymbol{\mu}+\mathbf{b} = \mathbf{k}^T\mathbf{C}_N^{-1}\mathbf{a}_N^\star = \mathbf{k}^T\mathbf{C}_N^{-1}\mathbf{C}_N(\mathbf{t}_N - \sigma_N) = \mathbf{k}^T(\mathbf{t}_N - \sigma_N)$$

Where we have used Eq (6.84). The covariance matrix is given by:

$$\begin{aligned}
\mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T &= c - \mathbf{k}^T\mathbf{C}_N^{-1}\mathbf{k} + \mathbf{k}^T\mathbf{C}_N^{-1}\mathbf{H}^{-1}(\mathbf{k}^T\mathbf{C}_N^{-1})^T \\
&= c - \mathbf{k}^T(\mathbf{C}_N^{-1} - \mathbf{C}_N^{-1}\mathbf{H}^{-1}\mathbf{C}_N^{-1})\mathbf{k} \\
&= c - \mathbf{k}^T\Big(\mathbf{C}_N^{-1} - \mathbf{C}_N^{-1}(\mathbf{W}_N + \mathbf{C}_N^{-1})^{-1}\mathbf{C}_N^{-1}\Big)\mathbf{k} \\
&= c - \mathbf{k}^T\Big(\mathbf{C}_N^{-1} - (\mathbf{C}_N\mathbf{W}_N\mathbf{C}_N + \mathbf{C}_N^{-1})^{-1}\Big)\mathbf{k}
\end{aligned}$$

Where we have used Eq (6.85) and the fact that $\mathbf{C}_N$ is symmetric. Then we use matrix identity (C.7) to further reduce the expression, which will finally give Eq (6.88).

Problem 6.27 Solution(Wait for update) This problem is really complicated.

What's more, I find that Eq (6.91) seems not right.

0.7 Sparse Kernel Machines

Problem 7.1 Solution

By analogy to Eq (2.249), we can obtain:

$$p(\mathbf{x}|t) = \begin{cases} \dfrac{1}{N_{+1}}\displaystyle\sum_{n=1}^{N_{+1}}\frac{1}{Z_k}\cdot k(\mathbf{x},\mathbf{x}_n) & t = +1 \\ \dfrac{1}{N_{-1}}\displaystyle\sum_{n=1}^{N_{-1}}\frac{1}{Z_k}\cdot k(\mathbf{x},\mathbf{x}_n) & t = -1 \end{cases}$$

where N_{+1} represents the number of samples with label $t = +1$ and it is the same for N_{-1}. Z_k is a normalization constant representing the volume of the hypercube. Since we have equal prior for the class, i.e.,

$$p(t) = \begin{cases} 0.5 & t = +1 \\ 0.5 & t = -1 \end{cases}$$

Based on Bayes' Theorem, we have $p(t|\mathbf{x}) \propto p(\mathbf{x}|t) \cdot p(t)$, yielding:

$$p(t|\mathbf{x}) = \begin{cases} \dfrac{1}{Z} \cdot \dfrac{1}{N_{+1}} \displaystyle\sum_{n=1}^{N_{+1}} \cdot k(\mathbf{x}, \mathbf{x}_n) & t = +1 \\ \dfrac{1}{Z} \cdot \dfrac{1}{N_{-1}} \displaystyle\sum_{n=1}^{N_{-1}} \cdot k(\mathbf{x}, \mathbf{x}_n) & t = -1 \end{cases}$$

Where $1/Z$ is a normalization constant to guarantee the integration of the posterior equal to 1. To classify a new sample $\mathbf{x}^\star$, we try to find the value $t^\star$ that can maximize $p(t|\mathbf{x})$. Therefore, we can obtain:

$$t^\star = \begin{cases} +1 & \text{if} \dfrac{1}{N_{+1}} \displaystyle\sum_{n=1}^{N_{+1}} \cdot k(\mathbf{x}, \mathbf{x}_n) \geq \dfrac{1}{N_{-1}} \displaystyle\sum_{n=1}^{N_{-1}} \cdot k(\mathbf{x}, \mathbf{x}_n) \\ -1 & \text{if} \dfrac{1}{N_{+1}} \displaystyle\sum_{n=1}^{N_{+1}} \cdot k(\mathbf{x}, \mathbf{x}_n) \leq \dfrac{1}{N_{-1}} \displaystyle\sum_{n=1}^{N_{-1}} \cdot k(\mathbf{x}, \mathbf{x}_n) \end{cases} \tag{$*$}$$

If we now choose the kernel function as $k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T \mathbf{x}'$,we have:

$$\frac{1}{N_{+1}} \sum_{n=1}^{N_{+1}} k(\mathbf{x}, \mathbf{x}_n) = \frac{1}{N_{+1}} \sum_{n=1}^{N_{+1}} \mathbf{x}^T \mathbf{x}_n = \mathbf{x}^T \tilde{\mathbf{x}}_{+1}$$

Where we have denoted:

$$\tilde{\mathbf{x}}_{+1} = \frac{1}{N_{+1}} \sum_{n=1}^{N_{+1}} \mathbf{x}_n$$

and similarly for $\tilde{\mathbf{x}}_{-1}$. Therefore, the classification criterion $(*)$ can be written as:

$$t^\star = \begin{cases} +1 & \text{if} \quad \tilde{\mathbf{x}}_{+1} \geq \tilde{\mathbf{x}}_{-1} \\ -1 & \text{if} \quad \tilde{\mathbf{x}}_{+1} \leq \tilde{\mathbf{x}}_{-1} \end{cases}$$

When we choose the kernel function as $k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^T \phi(\mathbf{x}')$, we can similarly obtain the classification criterion:

$$t^\star = \begin{cases} +1 & \text{if} \quad \tilde{\phi}(\mathbf{x}_{+1}) \geq \tilde{\phi}(\mathbf{x}_{-1}) \\ -1 & \text{if} \quad \tilde{\phi}(\mathbf{x}_{+1}) \leq \tilde{\phi}(\mathbf{x}_{-1}) \end{cases}$$

Where we have defined:

$$\tilde{\phi}(\mathbf{x}_{+1}) = \frac{1}{N_{+1}} \sum_{n=1}^{N_{+1}} \phi(\mathbf{x}_n)$$

Problem 7.2 Solution

Suppose we have find $\mathbf{w}_0$ and b_0, which can let all points satisfy Eq (7.5) and simultaneously minimize Eq (7.3). This hyperlane decided by $\mathbf{w}_0$ and b_0 is the optimal classification margin. Now if the constraint in Eq (7.5) becomes:

$$t_n(\mathbf{w}^T\phi(\mathbf{x}_n)+b) \geq \gamma$$

We can conclude that if we perform change of variables: $\mathbf{w}_0 -> \gamma\mathbf{w}_0$ and $b -> \gamma b$, the constraint will still satisfy and Eq (7.3) will be minimize. In other words, if the right side of the constraint changes from 1 to γ, The new hyperlane decided by $\gamma\mathbf{w}_0$ and γb_0 is the optimal classification margin. However, the minimum distance from the points to the classification margin is still the same.

Problem 7.3 Solution

Suppose we have $\mathbf{x}_1$ belongs to class one and we denote its target value $t_1 = 1$, and similarly $\mathbf{x}_2$ belongs to class two and we denote its target value $t_2 = -1$. Since we only have two points, they must have $t_i \cdot y(\mathbf{x}_i) = 1$ as shown in Fig. 7.1. Therefore, we have an equality constrained optimization problem:

$$\text{minimize}\,\frac{1}{2}||\mathbf{w}||^2 \quad \text{s.t.}\begin{cases}\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_1)+b = 1\\ \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_2)+b = -1\end{cases}$$

This is an convex optimization problem and it has been proved that global optimal exists.

Problem 7.4 Solution

Since we know that

$$\rho = \frac{1}{||\mathbf{w}||}$$

Therefore, we have:

$$\frac{1}{\rho^2} = ||\mathbf{w}||^2$$

In other words, we only need to prove that

$$||\mathbf{w}||^2 = \sum_{n=1}^{N} a_n$$

When we find th optimal solution, the second term on the right hand side of Eq (7.7) vanishes. Based on Eq (7.8) and Eq (7.10), we also observe that its dual is given by:

$$\tilde{L}(\mathbf{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2}||\mathbf{w}||^2$$

Therefore, we have:

$$\frac{1}{2}||\mathbf{w}||^2 = L(\mathbf{a}) = \tilde{L}(\mathbf{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2}||\mathbf{w}||^2$$

Rearranging it, we will obtain what we are required.

Problem 7.5 Solution

We have already proved this problem in the previous one.

Problem 7.6 Solution

If the target variable can only choose from $\{-1,1\}$, and we know that

$$p(t=1|y) = \sigma(y)$$

We can obtain:

$$p(t=-1|y) = 1 - p(t=1|y) = 1 - \sigma(y) = \sigma(-y)$$

Therefore, combining these two situations, we can derive:

$$p(t|y) = \sigma(yt)$$

Consequently, we can obtain the negative log likelihood:

$$-\ln p(\mathbf{D}) = -\ln \prod_{n=1}^{N} \sigma(y_n t_n) = -\sum_{n=1}^{N} \ln \sigma(y_n t_n) = \sum_{n=1}^{N} E_{LR}(y_n t_n)$$

Here $\mathbf{D}$ represents the dataset, i.e.,$\mathbf{D} = \{(\mathbf{x}_n, t_n); n = 1,2,...,N\}$, and $E_{LR}(yt)$ is given by Eq (7.48). With the addition of a quadratic regularization, we obtain exactly Eq (7.47).

Problem 7.7 Solution

The derivatives are easy to obtain. Our main task is to derive Eq (7.61)

using Eq (7.57)-(7.60).

$$\begin{aligned}
L &= C\sum_{n=1}^{N}(\xi_n+\widehat{\xi}_n)+\frac{1}{2}||\mathbf{w}||^2-\sum_{n=1}^{N}(\mu_n\xi_n+\widehat{\mu}_n\widehat{\xi}_n) \\
&\quad -\sum_{n=1}^{N}a_n(\epsilon+\xi_n+y_n-t_n)-\sum_{n=1}^{N}\widehat{a}_n(\epsilon+\widehat{\xi}_n+y_n-t_n) \\
&= C\sum_{n=1}^{N}(\xi_n+\widehat{\xi}_n)+\frac{1}{2}||\mathbf{w}||^2-\sum_{n=1}^{N}(a_n+\mu_n)\xi_n-\sum_{n=1}^{N}(\widehat{a}_n+\widehat{\mu}_n)\widehat{\xi}_n \\
&\quad -\sum_{n=1}^{N}a_n(\epsilon+y_n-t_n)-\sum_{n=1}^{N}\widehat{a}_n(\epsilon+y_n-t_n) \\
&= C\sum_{n=1}^{N}(\xi_n+\widehat{\xi}_n)+\frac{1}{2}||\mathbf{w}||^2-\sum_{n=1}^{N}C\xi_n-\sum_{n=1}^{N}C\widehat{\xi}_n \\
&\quad -\sum_{n=1}^{N}(a_n+\widehat{a}_n)\epsilon-\sum_{n=1}^{N}(a_n-\widehat{a}_n)(y_n-t_n) \\
&= \frac{1}{2}||\mathbf{w}||^2-\sum_{n=1}^{N}(a_n+\widehat{a}_n)\epsilon-\sum_{n=1}^{N}(a_n-\widehat{a}_n)(y_n-t_n) \\
&= \frac{1}{2}||\mathbf{w}||^2-\sum_{n=1}^{N}(a_n-\widehat{a}_n)(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n)+b-t_n)-\sum_{n=1}^{N}(a_n+\widehat{a}_n)\epsilon+\sum_{n=1}^{N} \\
&= \frac{1}{2}||\mathbf{w}||^2-\sum_{n=1}^{N}(a_n-\widehat{a}_n)(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n)+b)-\sum_{n=1}^{N}(a_n+\widehat{a}_n)\epsilon+\sum_{n=1}^{N}(a_n-\widehat{a}_n)t_n \\
&= \frac{1}{2}||\mathbf{w}||^2-\sum_{n=1}^{N}(a_n-\widehat{a}_n)\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n)-\sum_{n=1}^{N}(a_n+\widehat{a}_n)\epsilon+\sum_{n=1}^{N}(a_n-\widehat{a}_n)t_n \\
&= \frac{1}{2}||\mathbf{w}||^2-||\mathbf{w}||^2-\sum_{n=1}^{N}(a_n+\widehat{a}_n)\epsilon+\sum_{n=1}^{N}(a_n-\widehat{a}_n)t_n \\
&= -\frac{1}{2}||\mathbf{w}||^2-\sum_{n=1}^{N}(a_n+\widehat{a}_n)\epsilon+\sum_{n=1}^{N}(a_n-\widehat{a}_n)t_n
\end{aligned}$$

Just as required.

Problem 7.8 Solution

This obviously follows from the KKT condition, described in Eq (7.67) and (7.68).

Problem 7.9 Solution

The prior is given by Eq (7.80).

$$p(\mathbf{w}|\boldsymbol{\alpha})=\prod_{i=1}^{M}\mathcal{N}(0,\alpha_i^{-1})=\mathcal{N}(\mathbf{w}|\mathbf{0},\mathbf{A}^{-1})$$

Where we have defined:

$$\mathbf{A}=diag(\alpha_i)$$

The likelihood is given by Eq (7.79).

$$\begin{aligned} p(\mathbf{t}|\mathbf{X},\mathbf{w},\beta) &= \prod_{n=1}^{N} p(t_n|\mathbf{x}_n,\mathbf{w},\beta^{-1}) \\ &= \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n),\beta^{-1}) \\ &= \mathcal{N}(\mathbf{t}|\boldsymbol{\Phi}\mathbf{w},\beta^{-1}\mathbf{I}) \end{aligned}$$

Where we have defined:

$$\boldsymbol{\Phi} = [\boldsymbol{\phi}(\mathbf{x}_1), \boldsymbol{\phi}(\mathbf{x}_2), ..., \boldsymbol{\phi}(\mathbf{x}_n)]^T$$

Our definitions of $\boldsymbol{\Phi}$ and $\mathbf{A}$ as consistent with the main text. Therefore, according to Eq (2.113)-Eq (2.117), we have:

$$p(\mathbf{w}|\mathbf{t},\mathbf{X},\boldsymbol{\alpha},\beta) = \mathcal{N}(\mathbf{m},\boldsymbol{\Sigma})$$

Where we have defined:

$$\boldsymbol{\Sigma} = (\mathbf{A} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}$$

And

$$\mathbf{m} = \beta\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\mathbf{t}$$

Just as required.

Problem 7.10&7.11 Solution

It is quite similar to the previous problem. We begin by writting down the prior:

$$p(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{i=1}^{M} \mathcal{N}(0,\alpha_i^{-1}) = \mathcal{N}(\mathbf{w}|\mathbf{0},\mathbf{A}^{-1})$$

Then we write down the likelihood:

$$\begin{aligned} p(\mathbf{t}|\mathbf{X},\mathbf{w},\beta) &= \prod_{n=1}^{N} p(t_n|\mathbf{x}_n,\mathbf{w},\beta^{-1}) \\ &= \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n),\beta^{-1}) \\ &= \mathcal{N}(\mathbf{t}|\boldsymbol{\Phi}\mathbf{w},\beta^{-1}\mathbf{I}) \end{aligned}$$

Since we know that:

$$p(\mathbf{t}|\mathbf{X},\boldsymbol{\alpha},\beta) = \int p(\mathbf{t}|\mathbf{X},\mathbf{w},\beta)p(\mathbf{w}|\boldsymbol{\alpha})d\mathbf{w}$$

First as required by Prob.7.10, we will solve it by completing the square. We begin by write down the expression for $p(\mathbf{t}|\mathbf{X},\mathbf{w},\beta)$:

$$\begin{aligned} p(\mathbf{t}|\mathbf{X},\boldsymbol{\alpha},\beta) &= \int \mathcal{N}(\mathbf{w}|\mathbf{0},\mathbf{A}^{-1})\mathcal{N}(\mathbf{t}|\boldsymbol{\Phi}\mathbf{w},\beta^{-1}\mathbf{I})d\mathbf{w} \\ &= (\frac{\beta}{2\pi})^{N/2}\cdot\frac{1}{(2\pi)^{M/2}}\cdot\prod_{m=1}^{M}\alpha_i^{1/2}\cdot\int exp\{-E(\mathbf{w})\}d\mathbf{w} \end{aligned}$$

Where we have defined:

$$E(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{A}\mathbf{w} + \frac{\beta}{2}||\mathbf{t}-\boldsymbol{\Phi}\mathbf{w}||^2$$

We expand $E(\mathbf{w})$ with respect to $\mathbf{w}$:

$$\begin{aligned} E(\mathbf{w}) &= \frac{1}{2}\Big\{\mathbf{w}^T(\mathbf{A}+\beta\boldsymbol{\Phi}^T\boldsymbol{\Phi})\mathbf{w} - 2\beta\mathbf{t}^T(\boldsymbol{\Phi}\mathbf{w}) + \beta\mathbf{t}^T\mathbf{t}\Big\} \\ &= \frac{1}{2}\Big\{\mathbf{w}^T\boldsymbol{\Sigma}^{-1}\mathbf{w} - 2\mathbf{m}^T\boldsymbol{\Sigma}^{-1}\mathbf{w} + \beta\mathbf{t}^T\mathbf{t}\Big\} \\ &= \frac{1}{2}\Big\{(\mathbf{w}-\mathbf{m})^T\boldsymbol{\Sigma}^{-1}(\mathbf{w}-\mathbf{m}) + \beta\mathbf{t}^T\mathbf{t} - \mathbf{m}^T\boldsymbol{\Sigma}^{-1}\mathbf{m}\Big\} \end{aligned}$$

Where we have used Eq (7.82) and Eq (7.83). Substituting $E(\mathbf{w})$ into the integral, we will obtain:

$$\begin{aligned} p(\mathbf{t}|\mathbf{X},\boldsymbol{\alpha},\beta) &= (\frac{\beta}{2\pi})^{N/2}\cdot\frac{1}{(2\pi)^{M/2}}\cdot\prod_{m=1}^{M}\alpha_i^{1/2}\cdot\int exp\{-E(\mathbf{w})\}d\mathbf{w} \\ &= (\frac{\beta}{2\pi})^{N/2}\cdot\frac{1}{(2\pi)^{M/2}}\cdot\prod_{m=1}^{M}\alpha_i^{1/2}\cdot(2\pi)^{M/2}\cdot|\boldsymbol{\Sigma}|^{1/2}exp\Big\{-\frac{1}{2}(\beta\mathbf{t}^T\mathbf{t}-\mathbf{m}^T\boldsymbol{\Sigma}^{-1}\mathbf{m})\Big\} \\ &= (\frac{\beta}{2\pi})^{N/2}\cdot|\boldsymbol{\Sigma}|^{1/2}\cdot\prod_{m=1}^{M}\alpha_i^{1/2}\cdot exp\Big\{-\frac{1}{2}(\beta\mathbf{t}^T\mathbf{t}-\mathbf{m}^T\boldsymbol{\Sigma}^{-1}\mathbf{m})\Big\} \\ &= (\frac{\beta}{2\pi})^{N/2}\cdot|\boldsymbol{\Sigma}|^{1/2}\cdot\prod_{m=1}^{M}\alpha_i^{1/2}\cdot exp\Big\{-E(\mathbf{t})\Big\} \end{aligned}$$

We further expand $E(\mathbf{t})$:

$$\begin{aligned} E(\mathbf{t}) &= \frac{1}{2}(\beta\mathbf{t}^T\mathbf{t}-\mathbf{m}^T\boldsymbol{\Sigma}^{-1}\mathbf{m}) \\ &= \frac{1}{2}(\beta\mathbf{t}^T\mathbf{t}-(\beta\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\mathbf{t})^T\boldsymbol{\Sigma}^{-1}(\beta\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\mathbf{t})) \\ &= \frac{1}{2}(\beta\mathbf{t}^T\mathbf{t}-\beta^2\mathbf{t}^T\boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\mathbf{t}) \\ &= \frac{1}{2}(\beta\mathbf{t}^T\mathbf{t}-\beta^2\mathbf{t}^T\boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\mathbf{t}) \\ &= \frac{1}{2}\mathbf{t}^T(\beta\mathbf{I}-\beta^2\boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T)\mathbf{t} \\ &= \frac{1}{2}\mathbf{t}^T\Big[\beta\mathbf{I}-\beta\boldsymbol{\Phi}(\mathbf{A}+\beta\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\beta\Big]\mathbf{t} \\ &= \frac{1}{2}\mathbf{t}^T(\beta^{-1}\mathbf{I}+\boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}^T)^{-1}\mathbf{t} = \frac{1}{2}\mathbf{t}^T\mathbf{C}^{-1}\mathbf{t} \end{aligned}$$

Note that in the last step we have used matrix identity Eq (C.7). Therefore, as we know that the pdf is Gaussian and the exponential term has been given by $E(\mathbf{t})$, we can easily write down Eq (7.85) considering those normalization constant.

What's more, as required by Prob.7.11, the evaluation of the integral can be easily performed using Eq(2.113)- Eq(2.117).

Problem 7.12 Solution

According to the previous problem, we can explicitly write down the log marginal likelihood in an alternative form:

$$\ln p(\mathbf{t}|\mathbf{X},\boldsymbol{\alpha},\beta) = \frac{N}{2}\ln\beta - \frac{N}{2}\ln 2\pi + \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \frac{1}{2}\sum_{i=1}^{M}\ln\alpha_i - E(\mathbf{t})$$

We first derive:

$$\begin{aligned}
\frac{dE(\mathbf{t})}{d\alpha_i} &= -\frac{1}{2}\frac{d}{d\alpha_i}(\mathbf{m}^T\boldsymbol{\Sigma}^{-1}\mathbf{m}) \\
&= -\frac{1}{2}\frac{d}{d\alpha_i}(\beta^2\mathbf{t}^T\boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\mathbf{t}) \\
&= -\frac{1}{2}\frac{d}{d\alpha_i}(\beta^2\mathbf{t}^T\boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\mathbf{t}) \\
&= -\frac{1}{2}Tr\Big[\frac{d}{d\boldsymbol{\Sigma}^{-1}}(\beta^2\mathbf{t}^T\boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\mathbf{t})\cdot\frac{d\boldsymbol{\Sigma}^{-1}}{d\alpha_i}\Big] \\
&= \frac{1}{2}\beta^2 Tr\big[\boldsymbol{\Sigma}(\boldsymbol{\Phi}^T\mathbf{t})(\boldsymbol{\Phi}^T\mathbf{t})^T\boldsymbol{\Sigma}\cdot\mathbf{I}_i\big] = \frac{1}{2}m_{ii}^2
\end{aligned}$$

In the last step, we have utilized the following equation:

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

Moreover, here $\mathbf{I}_i$ is a matrix with all elements equal to zero, expect the i-th diagonal element, and the i-th diagonal element equals to 1. Then we utilize matrix identity Eq (C.22) to derive:

$$\begin{aligned}
\frac{d\ln|\boldsymbol{\Sigma}|}{d\alpha_i} &= -\frac{d\ln|\boldsymbol{\Sigma}^{-1}|}{d\alpha_i} \\
&= -Tr\Big[\boldsymbol{\Sigma}\frac{d}{d\alpha_i}(\mathbf{A}+\beta\boldsymbol{\Phi}^T\boldsymbol{\Phi})\Big] \\
&= -\Sigma_{ii}
\end{aligned}$$

Therefore, we can obtain:

$$\frac{d\ln p}{d\alpha_i} = \frac{1}{2\alpha_i} - \frac{1}{2}m_i^2 - \frac{1}{2}\Sigma_{ii}$$

Set it to zero and obtain:

$$\alpha_i = \frac{1-\alpha_i \Sigma_{ii}}{m_i} = \frac{\gamma_i}{m_i^2}$$

Then we calculate the derivatives of $\ln p$ with respect to β beginning by:

$$\begin{aligned}
\frac{d\ln|\mathbf{\Sigma}|}{d\beta} &= -\frac{d\ln|\mathbf{\Sigma}^{-1}|}{d\beta} \\
&= -Tr\Big[\mathbf{\Sigma}\frac{d}{d\beta}(\mathbf{A}+\beta\mathbf{\Phi}^T\mathbf{\Phi})\Big] \\
&= -Tr\Big[\mathbf{\Sigma}\mathbf{\Phi}^T\mathbf{\Phi}\Big]
\end{aligned}$$

Then we continue:

$$\begin{aligned}
\frac{dE(\mathbf{t})}{d\beta} &= \frac{1}{2}\mathbf{t}^T\mathbf{t} - \frac{1}{2}\frac{d}{d\beta}(\mathbf{m}^T\mathbf{\Sigma}^{-1}\mathbf{m}) \\
&= \frac{1}{2}\mathbf{t}^T\mathbf{t} - \frac{1}{2}\frac{d}{d\beta}(\beta^2\mathbf{t}^T\mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Sigma}^{-1}\mathbf{\Sigma}\mathbf{\Phi}^T\mathbf{t}) \\
&= \frac{1}{2}\mathbf{t}^T\mathbf{t} - \frac{1}{2}\frac{d}{d\beta}(\beta^2\mathbf{t}^T\mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}^T\mathbf{t}) \\
&= \frac{1}{2}\mathbf{t}^T\mathbf{t} - \beta\mathbf{t}^T\mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}^T\mathbf{t} - \frac{1}{2}\beta^2\frac{d}{d\beta}(\mathbf{t}^T\mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}^T\mathbf{t}) \\
&= \frac{1}{2}\Big\{\mathbf{t}^T\mathbf{t} - 2\beta\mathbf{t}^T\mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}^T\mathbf{t} - \beta^2\frac{d}{d\beta}(\mathbf{t}^T\mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}^T\mathbf{t})\Big\} \\
&= \frac{1}{2}\Big\{\mathbf{t}^T\mathbf{t} - 2\mathbf{t}^T(\mathbf{\Phi}\mathbf{m}) - \beta^2\frac{d}{d\beta}(\mathbf{t}^T\mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}^T\mathbf{t})\Big\} \\
&= \frac{1}{2}\Big\{\mathbf{t}^T\mathbf{t} - 2\mathbf{t}^T(\mathbf{\Phi}\mathbf{m}) - \beta^2 Tr\big[\frac{d}{d\mathbf{\Sigma}^{-1}}(\mathbf{t}^T\mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}^T\mathbf{t})\cdot\frac{d\mathbf{\Sigma}^{-1}}{d\beta}\big]\Big\} \\
&= \frac{1}{2}\Big\{\mathbf{t}^T\mathbf{t} - 2\mathbf{t}^T(\mathbf{\Phi}\mathbf{m}) + \beta^2 Tr\big[\mathbf{\Sigma}(\mathbf{\Phi}^T\mathbf{t})(\mathbf{\Phi}^T\mathbf{t})^T\mathbf{\Sigma}\cdot\mathbf{\Phi}^T\mathbf{\Phi}\big]\Big\} \\
&= \frac{1}{2}\Big\{\mathbf{t}^T\mathbf{t} - 2\mathbf{t}^T(\mathbf{\Phi}\mathbf{m}) + Tr\big[\mathbf{m}\mathbf{m}^T\cdot\mathbf{\Phi}^T\mathbf{\Phi}\big]\Big\} \\
&= \frac{1}{2}\Big\{\mathbf{t}^T\mathbf{t} - 2\mathbf{t}^T(\mathbf{\Phi}\mathbf{m}) + Tr\big[\mathbf{\Phi}\mathbf{m}\mathbf{m}^T\cdot\mathbf{\Phi}^T\big]\Big\} \\
&= \frac{1}{2}||\mathbf{t}-\mathbf{\Phi}\mathbf{m}||^2
\end{aligned}$$

Therefore, we have obtained:

$$\frac{d\ln p}{d\beta} = \frac{1}{2}\Big(\frac{N}{\beta} - ||\mathbf{t}-\mathbf{\Phi}\mathbf{m}||^2 - Tr[\mathbf{\Sigma}\mathbf{\Phi}^T\mathbf{\Phi}]\Big)$$

Using Eq (7.83), we can obtain:

$$\begin{aligned}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\boldsymbol{\Phi} &= \boldsymbol{\Sigma}\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \beta^{-1}\boldsymbol{\Sigma}\mathbf{A} - \beta^{-1}\boldsymbol{\Sigma}\mathbf{A} \\ &= \boldsymbol{\Sigma}(\beta\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \mathbf{A})\beta^{-1} - \beta^{-1}\boldsymbol{\Sigma}\mathbf{A} \\ &= \mathbf{I}\beta^{-1} - \beta^{-1}\boldsymbol{\Sigma}\mathbf{A} \\ &= (\mathbf{I} - \boldsymbol{\Sigma}\mathbf{A})\beta^{-1}\end{aligned}$$

Setting the derivative equal to zero, we can obtain:

$$\beta^{-1} = \frac{||\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}||^2}{N - Tr(\mathbf{I} - \boldsymbol{\Sigma}\mathbf{A})} = \frac{||\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}||^2}{N - \sum_i \gamma_i}$$

Just as required.

Problem 7.13 Solution

This problem is quite confusing. In my point of view, the posterior should be denoted as $p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \{a_i, b_i\}, a_\beta, b_\beta)$, where a_β, b_β controls the Gamma distribution of β, and a_i, b_i controls the Gamma distribution of α_i. What we should do is to maximize the marginal likelihood $p(\mathbf{t}|\mathbf{X}, \{a_i, b_i\}, a_\beta, b_\beta)$ with respect to $\{a_i, b_i\}, a_\beta, b_\beta$. Now we do not have a point estimation for the hyperparameters β and α_i. We have a distribution (controled by the hyper priors, i.e., $\{a_i, b_i\}, a_\beta, b_\beta$) instead.

Problem 7.14 Solution

We begin by writing down $p(t|\mathbf{x}, \mathbf{w}, \beta^*)$. Using Eq (7.76) and Eq (7.77), we can obtain:

$$p(t|\mathbf{x}, \mathbf{w}, \beta^*) = \mathcal{N}(t|\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}), (\beta^*)^{-1})$$

Then we write down $p(\mathbf{w}|\mathbf{X}, \mathbf{t}, \alpha^*, \beta^*)$. Using Eq (7.81), (7.82) and (7.83), we can obtain:

$$p(\mathbf{w}|\mathbf{X}, \mathbf{t}, \alpha^*, \beta^*) = \mathcal{N}(\mathbf{w}|\mathbf{m}, \boldsymbol{\Sigma})$$

Where $\mathbf{m}$ and $\boldsymbol{\Sigma}$ are evaluated using Eq (7.82) and (7.83) given $\alpha = \alpha^*$ and $\beta = \beta^*$. Then we utilize Eq (7.90) and obtain:

$$\begin{aligned}p(t|\mathbf{x}, \mathbf{X}, \mathbf{t}, \alpha^*, \beta^*) &= \int \mathcal{N}(t|\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}), (\beta^*)^{-1})\mathcal{N}(\mathbf{w}|\mathbf{m}, \boldsymbol{\Sigma})\, d\mathbf{w} \\ &= \int \mathcal{N}(t|\boldsymbol{\phi}(\mathbf{x})^T\mathbf{w}, (\beta^*)^{-1})\mathcal{N}(\mathbf{w}|\mathbf{m}, \boldsymbol{\Sigma})\, d\mathbf{w}\end{aligned}$$

Using Eq (2.113)-(2.117), we can obtain:

$$p(t|\mathbf{x}, \mathbf{X}, \mathbf{t}, \alpha^*, \beta^*) = \mathcal{N}(\mu, \sigma^2)$$

Where we have defined:

$$\mu = \mathbf{m}^T\boldsymbol{\phi}(\mathbf{x})$$

And

$$\sigma^2 = (\beta^*)^{-1} + \boldsymbol{\phi}(\mathbf{x})^T \boldsymbol{\Sigma} \boldsymbol{\phi}(\mathbf{x})$$

Just as required.

Problem 7.15 Solution

We just follow the hint.

$$\begin{aligned}
L(\boldsymbol{\alpha}) &= -\frac{1}{2}\{N\ln 2\pi + \ln|\mathbf{C}| + \mathbf{t}^T\mathbf{C}^{-1}\mathbf{t}\} \\
&= -\frac{1}{2}\Big\{N\ln 2\pi + \ln|\mathbf{C}_{-i}| + \ln|1+\alpha_i^{-1}\boldsymbol{\varphi}_i^T\mathbf{C}_{-i}^{-1}\boldsymbol{\varphi}_i| \\
&\quad + \mathbf{t}^T(\mathbf{C}_{-i}^{-1} - \frac{\mathbf{C}_{-i}^{-1}\boldsymbol{\varphi}_i\boldsymbol{\varphi}_i^T\mathbf{C}_{-i}^{-1}}{\alpha_i + \boldsymbol{\varphi}_i^T\mathbf{C}_{-i}^{-1}\boldsymbol{\varphi}_i})\mathbf{t}\Big\} \\
&= L(\boldsymbol{\alpha}_{-i}) - \frac{1}{2}\ln|1+\alpha_i^{-1}\boldsymbol{\varphi}_i^T\mathbf{C}_{-i}^{-1}\boldsymbol{\varphi}_i| + \frac{1}{2}\mathbf{t}^T\frac{\mathbf{C}_{-i}^{-1}\boldsymbol{\varphi}_i\boldsymbol{\varphi}_i^T\mathbf{C}_{-i}^{-1}}{\alpha_i + \boldsymbol{\varphi}_i^T\mathbf{C}_{-i}^{-1}\boldsymbol{\varphi}_i}\mathbf{t} \\
&= L(\boldsymbol{\alpha}_{-i}) - \frac{1}{2}\ln|1+\alpha_i^{-1}s_i| + \frac{1}{2}\frac{q_i^2}{\alpha_i + s_i} \\
&= L(\boldsymbol{\alpha}_{-i}) - \frac{1}{2}\ln\frac{\alpha_i + s_i}{\alpha_i} + \frac{1}{2}\frac{q_i^2}{\alpha_i + s_i} \\
&= L(\boldsymbol{\alpha}_{-i}) + \frac{1}{2}\Big[\ln\alpha_i - \ln(\alpha_i + s_i) + \frac{q_i^2}{\alpha_i + s_i}\Big] = L(\boldsymbol{\alpha}_{-i}) + \lambda(\alpha_i)
\end{aligned}$$

Where we have defined $\lambda(\alpha_i)$, s_i and q_i as shown in Eq (7.97)-(7.99).

Problem 7.16 Solution

We first calculate the first derivative of Eq(7.97) with respect to α_i:

$$\frac{\partial\lambda}{\partial\alpha_i} = \frac{1}{2}[\frac{1}{\alpha_i} - \frac{1}{\alpha_i + s_i} - \frac{q_i^2}{(\alpha_i + s_i)^2}]$$

Then we calculate the second derivative:

$$\frac{\partial^2\lambda}{\partial\alpha_i^2} = \frac{1}{2}[-\frac{1}{\alpha_i^2} + \frac{1}{(\alpha_i + s_i)^2} + \frac{2q_i^2}{(\alpha_i + s_i)^3}]$$

Next we aim to prove that when α_i is given by Eq (7.101), i.e., setting the first derivative equal to 0, the second derivative (i.e., the expression above) is negative. First we can obtain:

$$\alpha_i + s_i = \frac{s_i^2}{q_i^2 - s_i} + s_i = \frac{s_i q_i^2}{q_i^2 - s_i}$$

Therefore, substituting $\alpha_i + s_i$ and α_i into the second derivative, we can obtain:

$$\begin{aligned}
\frac{\partial^2 \lambda}{\partial \alpha_i^2} &= \frac{1}{2}[-\frac{(q_i^2 - s_i)^2}{s_i^4} + \frac{(q_i^2 - s_i)^2}{s_i^2 q_i^4} + \frac{2q_i^2(q_i^2 - s_i)^3}{s_i^3 q_i^6}] \\
&= \frac{1}{2}[-\frac{q_i^4(q_i^2 - s_i)^2}{q_i^4 s_i^4} + \frac{s_i^2(q_i^2 - s_i)^2}{s_i^4 q_i^4} + \frac{2s_i(q_i^2 - s_i)^3}{s_i^4 q_i^4}] \\
&= \frac{1}{2}\frac{(q_i^2 - s_i)^2}{q_i^4 s_i^4}[-q_i^4 + s_i^2 + 2s_i(q_i^2 - s_i)] \\
&= \frac{1}{2}\frac{(q_i^2 - s_i)^2}{q_i^4 s_i^4}[-(q_i^2 - s_i)^2] \\
&= -\frac{1}{2}\frac{(q_i^2 - s_i)^4}{q_i^4 s_i^4} < 0
\end{aligned}$$

Just as required.

Problem 7.17 Solution

We just follow the hint. According to Eq (7.102), Eq (7.86) and matrix identity (C.7), we have:

$$\begin{aligned}
Q_i &= \boldsymbol{\varphi}_i^T \mathbf{C}^{-1} \mathbf{t} \\
&= \boldsymbol{\varphi}_i^T (\beta^{-1}\mathbf{I} + \boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}^T)^{-1}\mathbf{t} \\
&= \boldsymbol{\varphi}_i^T (\beta\mathbf{I} - \beta\mathbf{I}\boldsymbol{\Phi}(\mathbf{A} + \boldsymbol{\Phi}^T\beta\mathbf{I}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\beta\mathbf{I})\mathbf{t} \\
&= \boldsymbol{\varphi}_i^T (\beta - \beta^2\boldsymbol{\Phi}(\mathbf{A} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T)\mathbf{t} \\
&= \boldsymbol{\varphi}_i^T (\beta - \beta^2\boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T)\mathbf{t} \\
&= \beta\boldsymbol{\varphi}_i^T\mathbf{t} - \beta^2\boldsymbol{\varphi}_i^T\boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\mathbf{t}
\end{aligned}$$

Similarly, we can obtain:

$$\begin{aligned}
S_i &= \boldsymbol{\varphi}_i^T \mathbf{C}^{-1} \boldsymbol{\varphi}_i \\
&= \boldsymbol{\varphi}_i^T (\beta - \beta^2\boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T)\boldsymbol{\varphi}_i \\
&= \beta\boldsymbol{\varphi}_i^T\boldsymbol{\varphi}_i - \beta^2\boldsymbol{\varphi}_i^T\boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\boldsymbol{\varphi}_i
\end{aligned}$$

Just as required.

Problem 7.18 Solution

We begin by deriving the first term in Eq (7.109) with respect to $\mathbf{w}$. This can be easily evaluate based on Eq (4.90)-(4.91).

$$\frac{\partial}{\partial \mathbf{w}}\Big\{\sum_{n=1}^{N} t_n \ln y_n + (1 - t_n)\ln(1 - y_n)\Big\} = \sum_{n=1}^{N}(t_n - y_n)\boldsymbol{\phi}_n = \boldsymbol{\Phi}^T(\mathbf{t} - \mathbf{y})$$

Since the derivative of the second term in Eq (7.109) with respect to $\mathbf{w}$ is rather simple to obtain. Therefore, The first derivative of Eq (7.109) with respect to $\mathbf{w}$ is:

$$\frac{\partial \ln p}{\partial \mathbf{w}} = \mathbf{\Phi}^T(\mathbf{t}-\mathbf{y}) - \mathbf{A}\mathbf{w}$$

For the Hessian matrix, we can first obtain:

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{w}}\Big\{\mathbf{\Phi}^T(\mathbf{t}-\mathbf{y})\Big\} &= \sum_{n=1}^{N}\frac{\partial}{\partial \mathbf{w}}\Big\{(t_n-y_n)\boldsymbol{\phi}_n\Big\} \\
&= -\sum_{n=1}^{N}\frac{\partial}{\partial \mathbf{w}}\Big\{y_n\cdot\boldsymbol{\phi}_n\Big\} \\
&= -\sum_{n=1}^{N}\frac{\partial \sigma(\mathbf{w}^T\boldsymbol{\phi}_n)}{\partial \mathbf{w}}\cdot\boldsymbol{\phi}_n^T \\
&= -\sum_{n=1}^{N}\frac{\partial \sigma(a)}{\partial a}\cdot\frac{\partial a}{\partial \mathbf{w}}\cdot\boldsymbol{\phi}_n^T
\end{aligned}$$

Where we have defined $a = \mathbf{w}^T\boldsymbol{\phi}_n$. Then we can utilize Eq (4.88) to derive:

$$\frac{\partial}{\partial \mathbf{w}}\Big\{\mathbf{\Phi}^T(\mathbf{t}-\mathbf{y})\Big\} = -\sum_{n=1}^{N}\sigma(1-\sigma)\cdot\boldsymbol{\phi}_n\cdot\boldsymbol{\phi}_n^T = -\mathbf{\Phi}^T\mathbf{B}\mathbf{\Phi}$$

Where $\mathbf{B}$ is a diagonal $N\times N$ matrix with elements $b_n = y_n(1-y_n)$. Therefore, we can obtain the Hessian matrix:

$$\mathbf{H} = \frac{\partial}{\partial \mathbf{w}}\Big\{\frac{\partial \ln p}{\partial \mathbf{w}}\Big\} = -(\mathbf{\Phi}^T\mathbf{B}\mathbf{\Phi}+\mathbf{A})$$

Just as required.

Problem 7.19 Solution

We begin from Eq (7.114).

$$\begin{aligned}
p(\mathbf{t}|\alpha) &= p(\mathbf{t}|\mathbf{w}^*)p(\mathbf{w}^*|\alpha)(2\pi)^{M/2}|\mathbf{\Sigma}|^{1/2} \\
&= \Big[\prod_{n=1}^{N}p(t_n|x_n,\mathbf{w})\Big]\Big[\prod_{i=1}^{M}\mathcal{N}(w_i|0,\alpha_i^{-1})\Big](2\pi)^{M/2}|\mathbf{\Sigma}|^{1/2}\Big|_{\mathbf{w}=\mathbf{w}^*} \\
&= \Big[\prod_{n=1}^{N}p(t_n|x_n,\mathbf{w})\Big]\cdot\mathcal{N}(\mathbf{w}|\mathbf{0},\mathbf{A})\cdot(2\pi)^{M/2}|\mathbf{\Sigma}|^{1/2}\Big|_{\mathbf{w}=\mathbf{w}^*}
\end{aligned}$$

We further take logarithm for both sides.

$$\begin{aligned}
\ln p(\mathbf{t}|\alpha) &= \Big[\sum_{n=1}^{N}\ln p(t_n|x_n,\mathbf{w}) + \ln\mathcal{N}(\mathbf{w}|\mathbf{0},\mathbf{A}) + \frac{M}{2}\ln 2\pi + \frac{1}{2}\ln|\mathbf{\Sigma}|\Big]\Big|_{\mathbf{w}=\mathbf{w}^*} \\
&= \Big[\sum_{n=1}^{N}\big[t_n\ln y_n + (1-t_n)\ln(1-y_n)\big] - \frac{1}{2}\mathbf{w}^T\mathbf{A}\mathbf{w} - \frac{1}{2}\ln|\mathbf{A}| + \frac{1}{2}\ln|\mathbf{\Sigma}| + const\Big]\Big|_{\mathbf{w}=\mathbf{w}^*} \\
&= \Big[\sum_{n=1}^{N}\big[t_n\ln y_n + (1-t_n)\ln(1-y_n)\big] - \frac{1}{2}\mathbf{w}^T\mathbf{A}\mathbf{w}\Big] + \Big[\frac{1}{2}\ln|\mathbf{\Sigma}| - \frac{1}{2}\ln|\mathbf{A}| + const\Big]\Big|_{\mathbf{w}=\mathbf{w}^*}
\end{aligned}$$

Using the Chain rule, we can obtain:

$$\frac{\partial \ln p(\mathbf{t}|\alpha)}{\partial \alpha_i}\Big|_{\mathbf{w}=\mathbf{w}^*} = \frac{\partial \ln p(\mathbf{t}|\alpha)}{\partial \mathbf{w}} \frac{\partial \mathbf{w}}{\partial \alpha_i}\Big|_{\mathbf{w}=\mathbf{w}^*}$$

Observing Eq (7.109), (7.110) and that (7.110) will equal 0 at $\mathbf{w}^*$, we can conclude that the first term on the right hand side of $\ln p(\mathbf{t}|\alpha)$ will have zero derivative with respect to $\mathbf{w}$ at $\mathbf{w}^*$. Therefore, we only need to focus on the second term:

$$\frac{\partial \ln p(\mathbf{t}|\alpha)}{\partial \alpha_i}\Big|_{\mathbf{w}=\mathbf{w}^*} = \frac{\partial}{\partial \alpha_i}\left[\frac{1}{2}\ln|\mathbf{\Sigma}| - \frac{1}{2}\ln|\mathbf{A}|\right]\Big|_{\mathbf{w}=\mathbf{w}^*}$$

It is rather easy to obtain:

$$\frac{\partial}{\partial \alpha_i}[-\frac{1}{2}\ln|\mathbf{A}|] = -\frac{1}{2}\frac{\partial}{\partial \alpha_i}\big[\sum_i \ln \alpha_i^{-1}\big] = \frac{1}{2\alpha_i}$$

Then we follow the same procedure as in Prob.7.12, we can obtain:

$$\frac{\partial}{\partial \alpha_i}[\frac{1}{2}\ln|\mathbf{\Sigma}|] = -\frac{1}{2}\Sigma_{ii}$$

Therefore, we obtain:

$$\frac{\partial \ln p(\mathbf{t}|\alpha)}{\partial \alpha_i} = \frac{1}{2\alpha_i} - \frac{1}{2}\Sigma_{ii}$$

Note: here I draw a different conclusion as the main text. I have also verified my result in another way. You can write the prior as the product of $\mathcal{N}(w_i|0,\alpha_i^{-1})$ instead of $\mathcal{N}(\mathbf{w}|\mathbf{0},\mathbf{A})$. In this form, since we know that:

$$\frac{\partial}{\partial \alpha_i}\sum_{i=1}^{M}\ln \mathcal{N}(w_i|0,\alpha_i^{-1}) = \frac{\partial}{\partial \alpha_i}(\frac{1}{2}\ln \alpha_i - \frac{\alpha_i}{2}w_i^2) = \frac{1}{2\alpha_i} - \frac{1}{2}(w_i^*)^2$$

The above expression can be used to replace the derivative of $-1/2\mathbf{w}^T\mathbf{A}\mathbf{w} - 1/2\ln|\mathbf{A}|$. Since the derivative of the likelihood with respect to α_i is not zero at $\mathbf{w}^*$, (7.115) seems not right anyway.

0.8 Graphical Models

Problem 8.1 Solution

We are required to prove:

$$\int_{\mathbf{x}} p(\mathbf{x})d\mathbf{x} = \int_{\mathbf{x}}\prod_{k=1}^{K} p(x_k|pa_k)d\mathbf{x} = 1$$

Here we adopt the same assumption as in the main text: No arrows lead from a higher numbered node to a According to Eq(8.5), we can write:

$$\begin{aligned}
\int_{\mathbf{x}} p(\mathbf{x})d\mathbf{x} &= \int_{\mathbf{x}} \prod_{k=1}^{K} p(x_k|pa_k)d\mathbf{x} \\
&= \int_{\mathbf{x}} p(x_K|pa_K) \prod_{k=1}^{K-1} p(x_k|pa_k)d\mathbf{x} \\
&= \int_{[x_1,x_2,...,x_{K-1}]} \int_{x_K} \Big[p(x_K|pa_K) \prod_{k=1}^{K-1} p(x_k|pa_k)dx_K \Big] dx_1 dx_2,...dx_{K-1} \\
&= \int_{[x_1,x_2,...,x_{K-1}]} \Big[\prod_{k=1}^{K-1} p(x_k|pa_k) \int_{x_K} p(x_K|pa_K)dx_K \Big] dx_1 dx_2,...dx_{K-1} \\
&= \int_{[x_1,x_2,...,x_{K-1}]} \Big[\prod_{k=1}^{K-1} p(x_k|pa_k) \Big] dx_1 dx_2,...dx_{K-1} \\
&= \int_{[x_1,x_2,...,x_{K-1}]} \prod_{k=1}^{K-1} p(x_k|pa_k)dx_1 dx_2,...dx_{K-1}
\end{aligned}$$

Note that from the third line to the fourth line, we have used the fact that $x_1, x_2, ...x_{K-1}$ do not depend on x_K, and thus the product from $k = 1$ to $K-1$ can be moved to the outside of the integral with respect to x_K, and that we have used the fact that the conditional probability is correctly normalized from the fourth line to the fifth line. The aforementioned procedure will be repeated for K times until all the variables have been integrated out.

Problem 8.2 Solution

This statement is obvious. Suppose that there exists an ordered numbering of the nodes such that for each node there are no links going to a lower-numbered node, and that there is a directed cycle in the graph:

$$a_1 \to a_2 \to ... \to a_N$$

To make it a real cycle, we also require $a_N \to a_1$. According to the assumption, we have $a_1 \le a_2 \le ... \le a_N$. Therefore, the last link $a_N \to a_1$ is invalid since $a_N \ge a_1$.

Problem 8.3 Solution

Based on definition, we can obtain:

$$p(a,b) = p(a,b,c=0) + p(a,b,c=1) = \begin{cases} 0.336, \text{ if } a = 0, b = 0 \\ 0.264, \text{ if } a = 0, b = 1 \\ 0.256, \text{ if } a = 1, b = 0 \\ 0.144, \text{ if } a = 1, b = 1 \end{cases}$$

Similarly, we can obtain:

$$p(a) = p(a,b=0) + p(a,b=1) = \begin{cases} 0.6, \text{ if } a = 0 \\ 0.4, \text{ if } a = 1 \end{cases}$$

And

$$p(b) = p(a=0,b) + p(a=1,b) = \begin{cases} 0.592, \text{if } b = 0 \\ 0.408, \text{if } b = 1 \end{cases}$$

Therefore, we conclude that $p(a,b) \neq p(a)p(b)$. For instance, we have $p(a=1,b=1) = 0.144$, $p(a=1) = 0.4$ and $p(b=1) = 0.408$. It is obvious that:

$$0.144 = p(a=1,b=1) \neq p(a=1)p(b=1) = 0.4 \times 0.408$$

To prove the conditional dependency, we first calculate $p(c)$:

$$p(c) = \sum_{a,b=0,1} p(a,b,c) = \begin{cases} 0.480, \text{if } c = 0 \\ 0.520, \text{if } c = 1 \end{cases}$$

According to Bayes' Theorem, we have:

$$p(a,b|c) = \frac{p(a,b,c)}{p(c)} = \begin{cases} 0.400, \text{if } a = 0, b = 0, c = 0 \\ 0.277, \text{if } a = 0, b = 0, c = 1 \\ 0.100, \text{if } a = 0, b = 1, c = 0 \\ 0.415, \text{if } a = 0, b = 1, c = 1 \\ 0.400, \text{if } a = 1, b = 0, c = 0 \\ 0.123, \text{if } a = 1, b = 0, c = 1 \\ 0.100, \text{if } a = 1, b = 1, c = 0 \\ 0.185, \text{if } a = 1, b = 1, c = 1 \end{cases}$$

Similarly, we also have:

$$p(a|c) = \frac{p(a,c)}{p(c)} = \begin{cases} 0.240/0.480 = 0.500, \text{if } a = 0, c = 0 \\ 0.360/0.520 = 0.692, \text{if } a = 0, c = 1 \\ 0.240/0.480 = 0.500, \text{if } a = 1, c = 0 \\ 0.160/0.520 = 0.308, \text{if } a = 1, c = 1 \end{cases}$$

Where we have used $p(a,c) = p(a,b=0,c) + p(a,b=1,c)$. Similarly, we can obtain:

$$p(b|c) = \frac{p(b,c)}{p(c)} = \begin{cases} 0.384/0.480 = 0.800, \text{if } b = 0, c = 0 \\ 0.208/0.520 = 0.400, \text{if } b = 0, c = 1 \\ 0.096/0.480 = 0.200, \text{if } b = 1, c = 0 \\ 0.312/0.520 = 0.600, \text{if } b = 1, c = 1 \end{cases}$$

Now we can easily verify the statement $p(a,b|c) = p(a|c)p(b|c)$. For instance, we have:

$$0.1 = p(a=1,b=1|c=0) = p(a=1|c=0)p(b=1|c=0) = 0.5 \times 0.2 = 0.1$$

Problem 8.4 Solution

This problem follows the previous one. We have already calculated $p(a)$ and $p(b|c)$, we rewrite it here.

$$p(a) = p(a, b = 0) + p(a, b = 1) = \begin{cases} 0.6, \text{if } a = 0 \\ 0.4, \text{if } a = 1 \end{cases}$$

And

$$p(b|c) = \frac{p(b,c)}{p(c)} = \begin{cases} 0.384/0.480 = 0.800, \text{if } b = 0, c = 0 \\ 0.208/0.520 = 0.400, \text{if } b = 0, c = 1 \\ 0.096/0.480 = 0.200, \text{if } b = 1, c = 0 \\ 0.312/0.520 = 0.600, \text{if } b = 1, c = 1 \end{cases}$$

We can also obtain $p(c|a)$:

$$p(c|a) = \frac{p(a,c)}{p(a)} = \begin{cases} 0.24/0.6 = 0.4, \text{if } a = 0, c = 0 \\ 0.36/0.6 = 0.6, \text{if } a = 0, c = 1 \\ 0.24/0.4 = 0.6, \text{if } a = 1, c = 0 \\ 0.16/0.4 = 0.4, \text{if } a = 1, c = 1 \end{cases}$$

Now we can easily verify the statement that $p(a,b,c) = p(a)p(c|a)p(b|c)$ given Table 8.2. The directed graph looks like:

$$a \rightarrow c \rightarrow b$$

Problem 8.5 Solution

It looks quite like Figure 8.6. The difference is that we introduce α_i for each w_i, where $i = 1, 2, ..., M$.

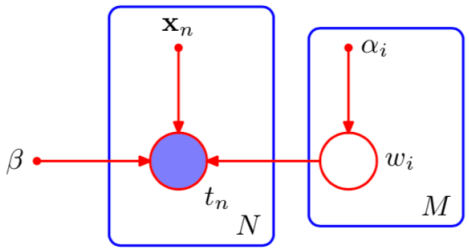

Figure 1: probabilistic graphical model corresponding to the RVM described in (7.79) and (7.80).

Problem 8.6 Solution(Wait for update)

Problem 8.7 Solution

Let's just follow the hint. We begin by calculating the mean $\boldsymbol{\mu}$.

$$\mathbb{E}[x_1] = b_1$$

According to Eq (8.15), we can obtain:

$$\mathbb{E}[x_2] = \sum_{j \in pa_2} w_{2j}\mathbb{E}[x_j] + b_2 = w_{21}b_1 + b_2$$

Then we can obtain:

$$\begin{aligned} \mathbb{E}[x_3] &= w_{32}\mathbb{E}[x_2] + b_3 \\ &= w_{32}(w_{21}b_1 + b_2) + b_3 \\ &= w_{32}w_{21}b_1 + w_{32}b_2 + b_3 \end{aligned}$$

Therefore, we obtain Eq (8.17) just as required. Next, we deal with the covariance matrix.

$$\text{cov}[x_1, x_1] = v_1$$

Then we can obtain:

$$\text{cov}[x_1, x_2] = \sum_{k=1} w_{2k}\text{cov}[x_1, x_k] + I_{12}v_2 = w_{21}\text{cov}[x_1, x_1] = w_{21}v_1$$

And also $\text{cov}[x_2, x_1] = \text{cov}[x_1, x_2] = w_{21}v_1$. Hence, we can obtain:

$$\text{cov}[x_2, x_2] = \sum_{k=1} w_{2k}\text{cov}[x_2, x_k] + I_{22}v_2 = w_{21}^2 v_1 + v_2$$

Next, we can obtain:

$$\text{cov}[x_1, x_3] = \sum_{k=2} w_{3k}\text{cov}[x_1, x_k] + I_{31}v_1 = w_{32}w_{21}v_1$$

Then, we can obtain:

$$\text{cov}[x_2, x_3] = \sum_{k=2} w_{3k}\text{cov}[x_2, x_k] + I_{23}v_3 = w_{32}(v_2 + w_{21}^2 v_1)$$

Finally, we can obtain:

$$\begin{aligned} \text{cov}[x_3, x_3] &= \sum_{k=2} w_{3k}\text{cov}[x_3, x_k] + I_{33}v_3 \\ &= w_{32}\Big[w_{32}(v_2 + w_{21}^2 v_1)\Big] + v_3 \end{aligned}$$

Where we have used the fact that $\text{cov}[x_3, x_2] = \text{cov}[x_2, x_3]$. By now, we have obtained Eq (8.18) just as required.

Problem 8.8 Solution

According to the definition, we can write:

$$p(a, b, c|d) = p(a|d)\, p(b, c|d)$$

We marginalize both sides with respect to c, yielding:

$$p(a,b|d) = p(a|d)p(b|d)$$

Just as required.

Problem 8.9 Solution

This statement is easy to see but a little bit difficult to prove. We put Fig 8.26 here to give a better illustration.

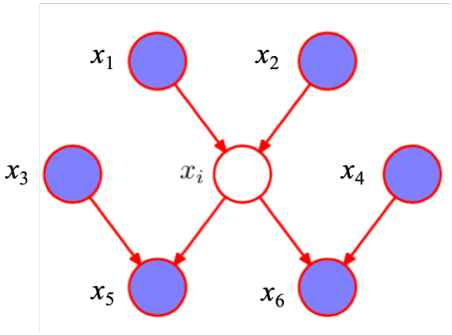

Figure 2: Markov blanket of a node x_i

Markov blanket Φ of node x_i is made up of three kinds of nodes:(i) the set Φ_1 containing all the parents of node x_i (x_1 and x_2 in Fig.2), (ii) the set Φ_2 containing all the children of node x_i (x_5 and x_6 in Fig.2), and (iii) the set Φ_3 containing all the co-parents of node x_i (x_3 and x_4 in Fig.2). According to the d-separation criterion, we need to show that all the paths from node x_i to an arbitrary node $\hat{x} \notin \Phi = \{\Phi_1 \cup \Phi_2 \cup \Phi_3\}$ are blocked given that the Markov blanket Φ are observed.

It is obvious that $\hat{x}$ can only connect to the target node x_i via two kinds of node: Φ_1,Φ_2. First, suppose that $\hat{x}$ connects to x_i via some node $x^\star \in \Phi_1$. The arrows definitely meet head-to-tail or tail-to-tail at node $x^\star$ because the link from a parent node $x^\star$ to x_i has its tail connected to the parent node $x^\star$, and since $x^\star$ is in $\Phi_1 \subseteq \Phi$, we see that this path is blocked.

In the second case, suppose that $\hat{x}$ connects to x_i via some node $x^\star \in \Phi_2$. We need to further divide this situation. If the path from $\hat{x}$ to x_i also goes through a node $x^{\star\star}$ from Φ_3 (e.g., in Fig.2, some node $\hat{x}$ connects to node x_3, and in this example $x^{\star\star} = x_3$, $x^\star = x_5$), it is clearly that the arrows meet head-to-tail or tail-to-tail at the node $x^{\star\star} \in \Phi_3 \subseteq \Phi$, this path is blocked.

In the final case, suppose that $\hat{x}$ connects to x_i via some node $x^\star \in \Phi_2$ and the path doesn't go through any node from Φ_3. An important observation is that the arrows cannot meet head-to-head at node $x^\star$ (otherwise, this path will go through a node from Φ_3). Thus, the arrows must meet either head-to-tail or tail-to-tail at node $x^\star \in \Phi_2 \subseteq \Phi$. Therefore, the path is also blocked.

Problem 8.10 Solution

By examining Fig.8.54, we can obtain:

$$p(a,b,c,d) = p(a)p(b)p(c|a,b)p(d|c)$$

Next we performing summation on both sides with respect to c and d, we can obtain:

$$\begin{aligned} p(a,b) &= p(a)p(b)\sum_c\sum_d p(c|a,b)p(d|c) \\ &= p(a)p(b)\sum_c p(c|a,b)\Big[\sum_d p(d|c)\Big] \\ &= p(a)p(b)\sum_c p(c|a,b)\times 1 \\ &= p(a)p(b)\times 1 \\ &= p(a)p(b) \end{aligned}$$

If we want to prove that a and b are dependent conditioned on d, we only need to prove:

$$p(a,b|d) = p(a|d)p(b|d)$$

We multiply both sides by $p(d)$ and use Bayes' Theorem, yielding:

$$p(a,b,d) = p(a)p(b|d) \qquad (*)$$

In other words, we can equivalently prove the expression above instead. Recall that we have:

$$p(a,b,c,d) = p(a)p(b)p(c|a,b)p(d|c)$$

We perform summation on both sides with respect to c, yielding:

$$p(a,b,d) = p(a)p(b)\sum_c p(c|a,b)p(d|c)$$

Combining with $(*)$, we only need to prove:

$$p(b|d) = p(b)\sum_c p(c|a,b)p(d|c)$$

However, we can see that the value of the right hand side depends on a,b and d, while the left hand side only depends on b and d. In general, this expression will not hold, and, thus, a and b are not dependent conditioned on d.

Problem 8.11 Solution

This problem is quite straightforward, but it needs some patience. According to the Bayes' Theorem, we have:

$$p(F=0|D=0) = \frac{p(D=0|F=0)\,p(F=0)}{p(D=0)} \qquad (*)$$

We will calculate each of the term on the right hand side. Let's begin from the numerator $p(D=0)$. According to the sum rule, we have:

$$\begin{aligned} p(D=0) &= p(D=0,G=0)+p(D=0,G=1) \\ &= p(D=0|G=0)p(G=0)+p(D=0|G=1)p(G=1) \\ &= 0.9\times 0.315+(1-0.9)\times(1-0.315) \\ &= 0.352 \end{aligned}$$

Where we have used Eq(8.30), Eq(8.105) and Eq(8.106). Note that the second term in the denominator, i.e., $p(F=0)$, equals 0.1, which can be easily derived from the main test above Eq(8.30). We now only need to calculate $p(D=0|F=0)$. Similarly, according to the sum rule, we have:

$$\begin{aligned} p(D=0|F=0) &= \sum_{G=0,1} p(D=0,G|F=0) \\ &= \sum_{G=0,1} p(D=0|G,F=0)\,p(G|F=0) \\ &= \sum_{G=0,1} p(D=0|G)\,p(G|F=0) \\ &= 0.9\times 0.81+(1-0.9)\times(1-0.81) \\ &= 0.748 \end{aligned}$$

Several clarifications must be made here. First, from the second line to the third line, we simply eliminate the dependence on $F=0$ because we know that D only depends on G according to Eq(8.105) and Eq(8.106). Second,from the third line to the fourth line, we have used Eq(8.31), Eq(8.105) and Eq(8.106). Now, we substitute all of them back to $(*)$, yielding:

$$p(F=0|D=0)=\frac{p(D=0|F=0)\,p(F=0)}{p(D=0)}=\frac{0.748\times 0.1}{0.352}=0.2125$$

Next, we are required to calculate the probability conditioned on both $D=0$ and $B=0$. Similarly, we can write:

$$\begin{aligned} p(F=0|D=0,B=0) &= \frac{p(D=0,B=0,F=0)}{p(D=0,B=0)} \\ &= \frac{\sum_G p(D=0,B=0,F=0,G)}{\sum_G p(D=0,B=0,G)} \\ &= \frac{\sum_G p(B=0,F=0,G)\,p(D=0|B=0,F=0,G)}{\sum_G p(B=0,G)\,p(D=0|B=0,G)} \\ &= \frac{\sum_G p(B=0,F=0,G)\,p(D=0|G)}{\sum_G p(B=0,G)\,p(D=0|G)} \quad (**) \end{aligned}$$

We need to calculate $p(B=0,F=0,G)$ and $p(B=0,G)$, where $G=0,1$.

We begin by calculating $p(B=0,F=0,G=0)$:

$$\begin{aligned} p(B=0,F=0,G=0) &= p(G=0|B=0,F=0)\times p(B=0,F=0) \\ &= p(G=0|B=0,F=0)\times p(B=0)\times p(F=0) \\ &= (1-0.1)\times(1-0.9)\times(1-0.9) \\ &= 0.009 \end{aligned}$$

Similarly, we can obtain $p(B=0,F=0,G=1)=0.001$. Next we calculate $p(B=0,G)$:

$$\begin{aligned} p(B=0,G=0) &= \sum_{F=0,1} p(B=0,G=0,F) \\ &= \sum_{F=0,1} p(G=0|B=0,F)\times p(B=0,F) \\ &= \sum_{F=0,1} p(G=0|B=0,F)\times p(B=0)\times p(F) \\ &= (1-0.1)\times(1-0.9)\times(1-0.9)+(1-0.2)\times(1-0.9)\times 0.9 \\ &= 0.081 \end{aligned}$$

Similarly, we can obtain $p(B=0,G=1)=0.019$. We substitute them back into $(**)$, yielding:

$$\begin{aligned} p(F=0|D=0,B=0) &= \frac{\sum_G p(B=0,F=0,G)\,p(D=0|G)}{\sum_G p(B=0,G)\,p(D=0|G)} \\ &= \frac{0.009\times 0.9+0.001\times(1-0.9)}{0.081\times 0.9+0.019\times(1-0.9)} \\ &= 0.1096 \end{aligned}$$

Just as required. The intuition behind this result coincides with the common sense. Moreover, by analogy to Fig.8.54, the node a and b in Fig.8.54 represents B and F in our case. Node c represents G, while node d represents D. You can use d-separation criterion to verify the conditional properties.

Problem 8.12 Solution

An intuitive solution is that we construct a matrix $\mathbf{A}$ with size of $M\times M$. If there is a link from node i to node j, the entry on the i-th row and j-th column of matrix $\mathbf{A}$, i.e., $A_{i,j}$, will equal to 1. Otherwise, it will equal to 0. Since the graph is undirected, the matrix $\mathbf{A}$ will be symmetric. What's more, the element on the diagonal is 0 by definition. For a undirected graph, we can use a matrix $\mathbf{A}$ to represent it. It is also a one-to-one mapping.

In other words, we equivalently count the number of possible matrix $\mathbf{A}$ satisfying the following criteria: (i) each of the entry is either 0 or 1, (ii) it is symmetric, and (iii) all of the entries on the diagonal are already determined (i.e., they all equal 0).

Using the property of symmetry, we only need to count the free variables on the lower triangle of the matrix. In the first column, there are $(M-1)$ free variables. In the second column, there are $(M-2)$ free variables. Therefore, the total free variables are given by:

$$(M-1)+(M-2)+...+0 = \frac{M(M-1)}{2}$$

Each value of these free variables has two choices, i.e., 1 or 0. Therefore, the total number of such matrix is $2^{M(M-1)/2}$. In the case of $M=3$, there are 8 possible undirected graphs:

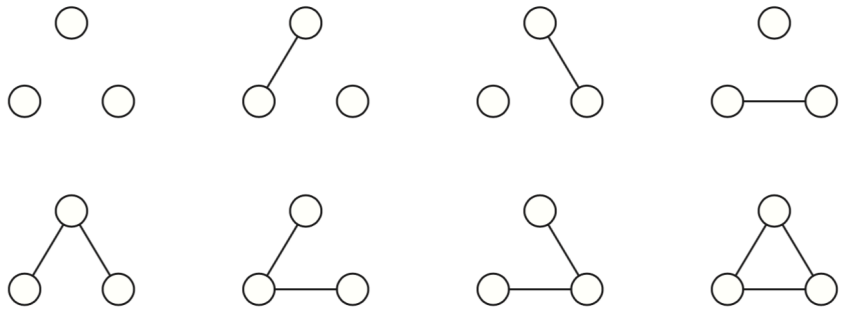

Figure 3: the undirected graph when $M=3$

Problem 8.13 Solution

It is straightforward. Suppose that x_k is the target variable whose state may be $\{-1.1\}$ while all other variables are fixed. According to Eq (8.42), we can obtain:

$$\begin{aligned} E(\mathbf{x},\mathbf{y}) &= h\sum_{i\neq k} x_i - \beta \sum_{i,j\neq k} x_i x_j - \eta \sum_{i\neq k} x_i y_i \\ &\quad + h x_k - \beta \sum_m x_k x_m - \eta x_k y_k \end{aligned}$$

Note that we write down the dependence of $E(\mathbf{x},\mathbf{y})$ on x_k explicitly, which is expressed via the second line. Moreover, the $x_i x_j$ term in the first line doesn't include the pairs $\{x_i, x_j\}$, which one of them is x_k. These terms are considered by $x_k x_m$ in the secone line. To be more specific, here x_m represents the neighbor of x_k. Noticing that the first line doesn't depend on x_k, we can obtain:

$$E(\mathbf{x},\mathbf{y})|_{x_k=1} - E(\mathbf{x},\mathbf{y})|_{x_k=-1} = 2h - 2\beta\sum_m x_m - 2\eta y_k$$

Obviously, the difference depends locally on x_k, implied by h, the neighbors x_m and its observed value y_k.

Problem 8.14 Solution

It is quite obvious. When $h = 0$, $\beta = 0$, the energy function reduces to

$$E(\mathbf{x},\mathbf{y}) = -\eta \sum_i x_i y_i$$

If there exists some index j which satisfies $x_j \neq y_j$, considering that $x_j, y_j \in \{-1.1\}$, then $x_j y_j$ will equal to -1. By changing the sign of x_j, we can always increase the value of $x_j y_j$ from -1 to 1, and, thus, decrease the energy function $E(\mathbf{x},\mathbf{y})$.

Therefore, given the observed binary pixels $y_i \in \{-1.1\}$, where $i = 1,2,...,D$, in order to obtain the minimum of energy, the optimal choice for x_i is to set it equal to y_i.

Problem 8.15 Solution

This problem can be solved by analogy to Eq (8.49) - Eq(8.54). We begin by noticing:

$$p(x_{n-1},x_n) = \sum_{x_1} ... \sum_{x_{n-2}} \sum_{x_{n+1}} ... \sum_{x_N} p(\mathbf{x})$$

We also have:

$$p(\mathbf{x}) = \frac{1}{Z}\psi_{1,2}(x_1,x_2)\psi_{2,3}(x_2,x_3)...\psi_{N-1,N}(x_{N-1},x_N)$$

By analogy to Eq(8.52), we can obtain:

$$\begin{aligned} p(x_{n-1},x_n) &= \frac{1}{Z}\Big[\sum_{x_{n-2}} \psi_{n-2,n-1}(x_{n-2},x_{n-1})...\Big[\sum_{x_2}\psi_{2,3}(x_2,x_3)\Big[\sum_{x_1}\psi_{1,2}(x_1,x_2)\Big]\Big]...\Big] \\ &\times \quad \psi_{n-1,n}(x_{n-1,x_n}) \\ &\times \quad \Big[\sum_{x_{n+1}} \psi_{n,n+1}(x_n,x_{n+1})...\Big[\sum_{x_N}\psi_{N-1,N}(x_{N-1},x_N)\Big]...\Big] \\ &= \frac{1}{Z}\times\mu_\alpha(x_{n-1})\times\psi_{n-1,n}(x_{n-1},x_n)\times\mu_\beta(x_n) \end{aligned}$$

Just as required.

Problem 8.16 Solution

We can simply obtain $p(x_N)$ using Eq(8.52) and Eq(8.54):

$$p(x_N) = \frac{1}{Z}\mu_\alpha(x_N) \qquad (*)$$

According to Bayes' Theorem, we have:

$$p(x_n|x_N) = \frac{p(x_n,x_N)}{p(x_N)}$$

Therefore, now we only need to derive an expression for $p(x_n,x_N)$, where $n = 1,2,...,N-1$. We follow the same procedure as in the previous problem. Since we know that:

$$p(x_n,x_N) = \sum_{x_1} ... \sum_{x_{n-1}} \sum_{x_{n+1}} ... \sum_{x_{N-1}} p(\mathbf{x})$$

We can obtain:

$$\begin{aligned} p(x_n, x_N) &= \frac{1}{Z}\Big[\sum_{x_{n-1}} \psi_{n-1,n}(x_{n-1}, x_n)...\Big[\sum_{x_2}\psi_{2,3}(x_2,x_3)\Big[\sum_{x_1}\psi_{1,2}(x_1,x_2)\Big]\Big]...\Big] \\ &\times \Big[\sum_{x_{n+1}} \psi_{n,n+1}(x_n, x_{n+1})...\Big[\sum_{x_{N-1}} \psi_{N-2,N-1}(x_{N-2}, x_{N-1})\psi_{N-1,N}(x_{N-1}, x_N)\Big]...\Big] \end{aligned}$$

Note that in the second line, the summation term with respect to x_{N-1} is the product of $\psi_{N-2,N-1}(x_{N-2}, x_{N-1})$ and $\psi_{N-1,N}(x_{N-1}, x_N)$. So here we can actually draw an undirected graph with $N-1$ nodes, and adopt the proposed algorithm to solve $p(x_n, x_N)$. If we use $x_n^\star$ to represent the new node, then the joint distribution can be written as:

$$p(\mathbf{x}^\star) = \frac{1}{Z^\star}\psi_{1,2}^\star(x_1^\star, x_2^\star)\psi_{2,3}^\star(x_2^\star, x_3^\star)...\psi_{N-2,N-1}^\star(x_{N-2}^\star, x_{N-1}^\star)$$

Where $\psi_{n,n+1}^\star(x_n^\star, x_{n+1}^\star)$ is defined as:

$$\psi_{n,n+1}^\star(x_n^\star, x_{n+1}^\star) = \begin{cases} \psi_{n,n+1}(x_n, x_{n+1}), & n = 1,2,...,N-3 \\ \psi_{N-2,N-1}(x_{N-2}, x_{N-1})\psi_{N-1,N}(x_{N-1}, x_N), & n = N-2 \end{cases}$$

In other words, we have combined the original node x_{N-1} and x_N. Moreover, we have the relationship:

$$p(x_n, x_N) = p(x_n^\star) = \frac{1}{Z^\star}\mu_\alpha^\star(x_n^\star)\mu_\beta^\star(x_n^\star) \quad n = 1,2,...,N-1$$

By adopting the proposed algorithm to the new undirected graph, $p(x_n^\star)$ can be easily evaluated, and so is $p(x_n, x_N)$.

Problem 8.17 Solution

It is straightforward to see that for every path connecting node x_2 and x_5 in Fig.8.38, it must pass through node x_3. Therefore, all paths are blocked and the conditional property holds. For more details, you should read section 8.3.1. According to Bayes' Theorem, we can obtain:

$$p(x_2|x_3, x_5) = \frac{p(x_2, x_3, x_5)}{p(x_2)}$$

Using the proposed algorithm in section 8.4.1, we can obtain:

$$\begin{aligned} p(x_2|x_3, x_5) &= \frac{p(x_2, x_3, x_5)}{p(x_3, x_5)} = \frac{\sum_{x_1}\sum_{x_4} p(\mathbf{x})}{\sum_{x_1}\sum_{x_2}\sum_{x_4} p(\mathbf{x})} \\ &= \frac{\sum_{x_1}\sum_{x_4}\psi_{1,2}\psi_{2,3}\psi_{3,4}\psi_{4,5}}{\sum_{x_1}\sum_{x_2}\sum_{x_4}\psi_{1,2}\psi_{2,3}\psi_{3,4}\psi_{4,5}} \\ &= \frac{\Big(\sum_{x_1}\psi_{1,2}\Big)\cdot\psi_{2,3}\cdot\Big(\sum_{x_4}\psi_{3,4}\psi_{4,5}\Big)}{\sum_{x_2}\Big[\Big(\sum_{x_1}\psi_{1,2}\Big)\psi_{2,3}\Big]\cdot\Big(\sum_{x_4}\psi_{3,4}\psi_{4,5}\Big)} \\ &= \frac{\Big(\sum_{x_1}\psi_{1,2}\Big)\cdot\psi_{2,3}}{\sum_{x_2}\Big[\Big(\sum_{x_1}\psi_{1,2}\Big)\psi_{2,3}\Big]} \end{aligned}$$

It is obvious that the right hand side doesn't depend on x_5.

Problem 8.18 Solution

First, the distribution represented by a directed tree can be trivially be written as an equivalent distribution over an undirected tree by moralization. You can find more details in section 8.4.2.

Alternatively, now we want to represent a distribution, which is given by a directed graph, via a directed graph. For example, the distribution defined by the undirected tree in Fig.4 can be written as:

$$p(\mathbf{x}) = \frac{1}{Z}\psi_{1,3}(x_1,x_3)\,\psi_{2,3}(x_2,x_3)\,\psi_{3,4}(x_3,x_4)\,\psi_{4,5}(x_4,x_5)$$

We simply choose x_4 as the root and the corresponding directed tree is well defined by working outwards. In this case, the distribution defined by the directed tree is:

$$p(\mathbf{x}) = p(x_4)\,p(x_5|x_4)\,p(x_3|x_4)\,p(x_1|x_3)\,p(x_2|x_3)$$

Thus it is not difficult to change an undirected tree to a directed on if performing:

$$p(x_4)p(x_5|x_4) \propto \psi_{5,4},\ p(x_3|x_4) \propto \psi_{3,4},\ p(x_2|x_3) \propto \psi_{2,3},\ p(x_1|x_3) \propto \psi_{1,3},$$

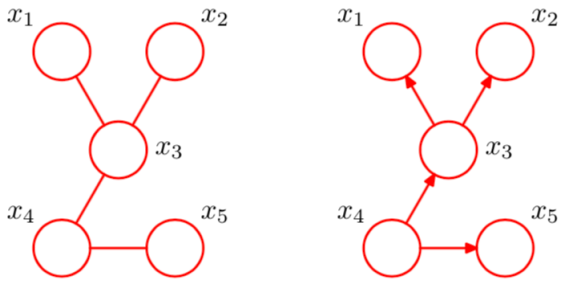

Figure 4: Example of changing an undirected tree to a directed one x_i

The symbol $\propto$ is used to represent a normalization term, which is used to guarantee the integral of PDF equal to 1. In summary, in the particular case of an undirected tree, there is only one path between any pair of nodes, and thus the maximal clique is given by a pair of two nodes in an undirected tree. This is because if we choose any three nodes x_1, x_2, x_3, according to the definition there cannot exist a loop. Otherwise there are two paths between x_1 and x_3: (i) $x_1 -> x_3$ and (ii) $x_1 -> x_2 -> x_3$. In the directed tree, each node

only depends on only one node (except the root), i.e., its parent. Thus we can easily change a undirected tree to a directed one by matching the potential function with the corresponding conditional PDF, as shown in the example.

Moreover, we can choose any node in the undirected tree to be the root and then work outwards to obtain a directed tree. Therefore, in an undirected tree with n nodes, there is n corresponding directed trees in total.

Problem 8.19-8.29 Solution (Waiting for update)

I am quite confused by the deduction in Eq(8.66). I do not understand the sum-prodcut algorithm and the max-sum algorithm very well.

0.9 Mixture Models and EM

Problem 9.1 Solution

For each r_{nk} when n is fixed and $k = 1,2,...,K$, only one of them equals 1 and others are all 0. Therefore, there are K possible choices. When N data are given, there are K^N possible assignments for $\{r_{nk}; n = 1,2,...,N; k = 1,2,...,K\}$. For each assignments, the optimal $\{\boldsymbol{\mu}_k; k = 1,2,...,K\}$ are well determined by Eq (9.4).

As discussed in the main text, by iteratively performing E-step and M-step, the distortion measure in Eq (9.1) is gradually minimized. The worst case is that we find the optimal assignment and $\{\boldsymbol{\mu}_k\}$ in the last iteration. In other words, K^N iterations are required. However, it is guaranteed to converge because the assignments are finite and the optimal $\{\boldsymbol{\mu}_k\}$ is determined once the assignment is given.

Problem 9.2 Solution

By analogy to Eq (9.1), we can write down:

$$J_N = J_{N-1} + \sum_{k=1}^{K} r_{Nk} ||\mathbf{x}_N - \boldsymbol{\mu}_k||^2$$

In the E-step, we still assign the N-th data $\mathbf{x}_N$ to the closet center and suppose that this cloest center is $\boldsymbol{\mu}_m$. Therefore, the expression above will reduce to:

$$J_N = J_{N-1} + ||\mathbf{x}_n - \boldsymbol{\mu}_m||^2$$

In the M-step, we set the derivative of J_N with respect to $\boldsymbol{\mu}_k$ to 0, where $k = 1,2,...,K$. We can observe that for those $\boldsymbol{\mu}_k$, $k \neq m$, we have:

$$\frac{\partial J_N}{\partial \boldsymbol{\mu}_k} = \frac{\partial J_{N-1}}{\partial \boldsymbol{\mu}_k}$$

In other words, we will only update $\boldsymbol{\mu}_m$ in the M-step by setting the derivative of J_N equal to 0. Utilizing Eq (9.4), we can obtain:

$$\begin{aligned}
\boldsymbol{\mu}_m^{(N)} &= \frac{\sum_{n=1}^{N-1} r_{nk}\mathbf{x}_n + \mathbf{x}_N}{\sum_{n=1}^{N-1} r_{nk} + 1} \\
&= \frac{\frac{\sum_{n=1}^{N-1} r_{nk}\mathbf{x}_n}{\sum_{n=1}^{N-1} r_{nk}} + \frac{\mathbf{x}_N}{\sum_{n=1}^{N-1} r_{nk}}}{1 + \frac{1}{\sum_{n=1}^{N-1} r_{nk}}} \\
&= \frac{\boldsymbol{\mu}_m^{(N-1)} + \frac{\mathbf{x}_N}{\sum_{n=1}^{N-1} r_{nk}}}{1 + \frac{1}{\sum_{n=1}^{N-1} r_{nk}}} \\
&= \boldsymbol{\mu}_m^{(N-1)} + \frac{\frac{\mathbf{x}_N}{\sum_{n=1}^{N-1} r_{nk}} - \frac{\boldsymbol{\mu}_m^{(N-1)}}{\sum_{n=1}^{N-1} r_{nk}}}{1 + \frac{1}{\sum_{n=1}^{N-1} r_{nk}}} \\
&= \boldsymbol{\mu}_m^{(N-1)} + \frac{\mathbf{x}_N - \boldsymbol{\mu}_m^{(N-1)}}{1 + \sum_{n=1}^{N-1} r_{nk}}
\end{aligned}$$

So far we have obtained a sequential on-line update formula just as required.

Problem 9.3 Solution

We simply follow the hint.

$$\begin{aligned}
p(\mathbf{x}) &= \sum_{\mathbf{z}} p(\mathbf{z}) p(\mathbf{x}|\mathbf{z}) \\
&= \sum_{\mathbf{z}} \prod_{k=1}^{K} \left[(\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right]^{z_k}
\end{aligned}$$

Note that we have used 1-of-K coding scheme for $\mathbf{z} = [z_1, z_2, ..., z_K]^T$. To be more specific, only one of $z_1, z_2, ..., z_K$ will be 1 and all others will equal 0. Therefore, the summation over $\mathbf{z}$ actually consists of K terms and the k-th term corresponds to z_k equal to 1 and others 0. Moreover, for the k-th term, the product will reduce to $\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. Therefore, we can obtain:

$$p(\mathbf{x}) = \sum_{\mathbf{z}} \prod_{k=1}^{K} \left[(\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right]^{z_k} = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Just as required.

Problem 9.4 Solution

According to Bayes' Theorem, we can write:

$$p(\boldsymbol{\theta}|\mathbf{X}) \propto p(\mathbf{X}|\boldsymbol{\theta}) p(\boldsymbol{\theta})$$

Taking logarithm on both sides, we can write:

$$\ln p(\boldsymbol{\theta}|\mathbf{X}) \propto \ln p(\mathbf{X}|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta})$$

Further utilizing Eq (9.29), we can obtain:

$$\begin{aligned}
\ln p(\boldsymbol{\theta}|\mathbf{X}) &\propto \ln\Big\{\sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})\Big\} + \ln p(\boldsymbol{\theta}) \\
&= \ln\Big\{[\sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})] \cdot p(\boldsymbol{\theta})\Big\} \\
&= \ln\Big\{\sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta}) p(\boldsymbol{\theta})\Big\}
\end{aligned}$$

In other words, in thise case, the only modification is that the term $p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})$ in Eq (9.29) will be replaced by $p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. Therefore, in the E-step, we still need to calculate the posterior $p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{old})$ and then in the M-step, we are required to maximize $Q^{'}(\boldsymbol{\theta},\boldsymbol{\theta}^{old})$. In this case, by analogy to Eq (9.30), we can write down $Q^{'}(\boldsymbol{\theta},\boldsymbol{\theta}^{old})$:

$$\begin{aligned}
Q^{'}(\boldsymbol{\theta},\boldsymbol{\theta}^{old}) &= \sum_{Z} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{old}) \ln\Big[p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})p(\boldsymbol{\theta})\Big] \\
&= \sum_{Z} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{old}) \Big[\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta})\Big] \\
&= \sum_{Z} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{old}) \ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta}) + \sum_{Z} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{old}) \ln p(\boldsymbol{\theta}) \\
&= \sum_{Z} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{old}) \ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) \cdot \sum_{Z} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{old}) \\
&= \sum_{Z} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{old}) \ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) \\
&= Q(\boldsymbol{\theta},\boldsymbol{\theta}^{old}) + \ln p(\boldsymbol{\theta})
\end{aligned}$$

Just as required.

Problem 9.5 Solution

Notice that the condition on $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\pi}$ can be omitted here, and we only need to prove $p(\mathbf{Z}|\mathbf{X})$ can be written as the product of $p(\mathbf{z}_n|\mathbf{x}_n)$. Correspondingly, the small dots representing $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\pi}$ can also be omitted in Fig 9.6. Observing Fig 9.6 and based on definition, we can write :

$$p(\mathbf{X},\mathbf{Z}) = p(\mathbf{x}_1,\mathbf{z}_1)p(\mathbf{z}_1)...p(\mathbf{x}_N,\mathbf{z}_N)p(\mathbf{z}_N) = p(\mathbf{x}_1,\mathbf{z}_1)...p(\mathbf{x}_N,\mathbf{z}_N)$$

Moreover, since there is no link from $\mathbf{z}_m$ to $\mathbf{z}_n$, from $\mathbf{x}_m$ to $\mathbf{x}_n$, and from $\mathbf{z}_m$ to $\mathbf{x}_n$ $(m \neq n)$, we can obtain:

$$p(\mathbf{Z}) = p(\mathbf{z}_1)...p(\mathbf{z}_N), \quad p(\mathbf{X}) = p(\mathbf{x}_1)...p(\mathbf{x}_N)$$

These can also be verified by calculating the marginal distribution from $p(\mathbf{X},\mathbf{Z})$, for example:

$$p(\mathbf{Z}) = \sum_{\mathbf{X}} p(\mathbf{X},\mathbf{Z}) = \sum_{\mathbf{x}_1,\ldots,\mathbf{x}_N} p(\mathbf{x}_1,\mathbf{z}_1)\ldots p(\mathbf{x}_N,\mathbf{z}_N) = p(\mathbf{z}_1)\ldots p(\mathbf{z}_N)$$

According to Bayes' Theorem, we have

$$\begin{aligned} p(\mathbf{Z}|\mathbf{X}) &= \frac{p(\mathbf{X}|\mathbf{Z})p(\mathbf{Z})}{p(\mathbf{X})} \\ &= \frac{\left[\prod_{n=1}^N p(\mathbf{x}_n|\mathbf{z}_n)\right]\left[\prod_{n=1}^N p(\mathbf{z}_n)\right]}{\prod_{n=1}^N p(\mathbf{x}_n)} \\ &= \prod_{n=1}^N \frac{p(\mathbf{x}_n|\mathbf{z}_n)p(\mathbf{z}_n)}{p(\mathbf{x}_n)} \\ &= \prod_{n=1}^N p(\mathbf{z}_n|\mathbf{x}_n) \end{aligned}$$

Just as required. The essence behind the problem is that in the directed graph, there are only links from $\mathbf{z}_n$ to $\mathbf{x}_n$. The deeper reason is that (i) the mixture model is given by Fig 9.4, and (ii) we assume the data $\{\mathbf{x}_n\}$ is i.i.d, and thus there is no link from $\mathbf{x}_m$ to $\mathbf{x}_n$.

Problem 9.6 Solution

By analogy to Eq (9.19), we calculate the derivative of Eq (9.14) with respect to $\mathbf{\Sigma}$:

$$\frac{\partial \ln p}{\partial \mathbf{\Sigma}} = \frac{\partial}{\partial \mathbf{\Sigma}}\{\sum_{n=1}^N \ln a_n\} = \sum_{n=1}^N \frac{1}{a_n}\frac{\partial a_n}{\partial \mathbf{\Sigma}} \qquad (*)$$

Where we have defined:

$$a_n = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\mathbf{\Sigma})$$

Recall that in Prob.2.34, we have proved:

$$\frac{\partial \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\mathbf{\Sigma})}{\partial \mathbf{\Sigma}} = -\frac{1}{2}\mathbf{\Sigma}^{-1} + \frac{1}{2}\mathbf{\Sigma}^{-1}\mathbf{S}_{nk}\mathbf{\Sigma}^{-1}$$

Where we have defined:

$$\mathbf{S}_{nk} = (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T$$

Therefore, we can obtain:

$$\begin{aligned}
\frac{\partial a_n}{\partial \boldsymbol{\Sigma}} &= \frac{\partial}{\partial \boldsymbol{\Sigma}}\Big\{\sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma})\Big\} \\
&= \sum_{k=1}^{K}\frac{\partial}{\partial \boldsymbol{\Sigma}}\Big\{\pi_k\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma})\Big\} \\
&= \sum_{k=1}^{K}\pi_k\frac{\partial}{\partial \boldsymbol{\Sigma}}\Big\{\exp\big[\ln\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma})\big]\Big\} \\
&= \sum_{k=1}^{K}\pi_k\cdot\exp\big[\ln\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma})\big]\cdot\frac{\partial}{\partial \boldsymbol{\Sigma}}\Big[\ln\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma})\Big] \\
&= \sum_{k=1}^{K}\pi_k\cdot\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma})\cdot(-\frac{1}{2}\boldsymbol{\Sigma}^{-1}+\frac{1}{2}\boldsymbol{\Sigma}^{-1}\mathbf{S}_{nk}\boldsymbol{\Sigma}^{-1})
\end{aligned}$$

Substitute the equation above into $(*)$, we can obtain:

$$\begin{aligned}
\frac{\partial \ln p}{\partial \boldsymbol{\Sigma}} &= \sum_{n=1}^{N}\frac{1}{a_n}\frac{\partial a_n}{\partial \boldsymbol{\Sigma}} \\
&= \sum_{n=1}^{N}\frac{\sum_{k=1}^{K}\pi_k\cdot\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma})\cdot(-\frac{1}{2}\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Sigma}^{-1}\mathbf{S}_{nk}\boldsymbol{\Sigma}^{-1})}{\sum_{j=1}^{K}\pi_j\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j,\boldsymbol{\Sigma})} \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\cdot(-\frac{1}{2}\boldsymbol{\Sigma}^{-1}+\frac{1}{2}\boldsymbol{\Sigma}^{-1}\mathbf{S}_{nk}\boldsymbol{\Sigma}^{-1}) \\
&= -\frac{1}{2}\Big\{\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\Big\}\boldsymbol{\Sigma}^{-1}+\frac{1}{2}\boldsymbol{\Sigma}^{-1}\Big\{\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\mathbf{S}_{nk}\Big\}\boldsymbol{\Sigma}^{-1}
\end{aligned}$$

If we set the derivative equal to 0, we can obtain:

$$\boldsymbol{\Sigma} = \frac{\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\mathbf{S}_{nk}}{\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})}$$

Problem 9.7 Solution

We begin by calculating the derivative of Eq (9.36) with respect to $\boldsymbol{\mu}_k$:

$$\begin{aligned}
\frac{\partial \ln p}{\partial \boldsymbol{\mu}_k} &= \frac{\partial}{\partial \boldsymbol{\mu}_k}\Big\{\sum_{n=1}^{N}\sum_{k=1}^{K}z_{nk}\left[\ln\pi_k+\ln\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right]\Big\} \\
&= \frac{\partial}{\partial \boldsymbol{\mu}_k}\Big\{\sum_{n=1}^{N}z_{nk}\left[\ln\pi_k+\ln\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right]\Big\} \\
&= \sum_{n=1}^{N}\frac{\partial}{\partial \boldsymbol{\mu}_k}\Big\{z_{nk}\ln\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\Big\} \\
&= \sum_{\mathbf{x}_n\in C_k}\frac{\partial}{\partial \boldsymbol{\mu}_k}\Big\{\ln\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\Big\}
\end{aligned}$$

Where we have used $\mathbf{x}_n \in C_k$ to represent the data point $\mathbf{x}_n$ which are assigned to the k-th cluster. Therefore, $\boldsymbol{\mu}_k$ is given by the mean of those $x_n \in C_k$ just as the case of a single Gaussian. It is exactly the same for the covariance. Next, we maximize Eq (9.36) with respect to π_k by enforcing a Lagrange multiplier:

$$L = \ln p + \lambda(\sum_{k=1}^{K} \pi_k - 1)$$

We calculate the derivative of L with respect to π_k and set it to 0:

$$\frac{\partial L}{\partial \pi_k} = \sum_{n=1}^{N} \frac{z_{nk}}{\pi_k} + \lambda = 0$$

We multiply both sides by π_k and sum over k making use of the constraint Eq (9.9), yielding $\lambda = -N$. Substituting it back into the expression, we can obtain:

$$\pi_k = \frac{1}{N} \sum_{n=1}^{N} z_{nk}$$

Just as required.

Problem 9.8 Solution

Since $\gamma(z_{nk})$ is fixed, the only dependency of Eq (9.40) on $\boldsymbol{\mu}_k$ occurs in the Gaussian, yielding:

$$\begin{aligned}
\frac{\partial \mathbb{E}_z[\ln p]}{\partial \boldsymbol{\mu}_k} &= \frac{\partial}{\partial \boldsymbol{\mu}_k} \Big\{ \sum_{n=1}^{N} \gamma(z_{nk}) \ln \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \Big\} \\
&= \sum_{n=1}^{N} \gamma(z_{nk}) \cdot \frac{\partial \ln \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\partial \boldsymbol{\mu}_k} \\
&= \sum_{n=1}^{N} \gamma(z_{nk}) \cdot \Big[-\boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k) \Big]
\end{aligned}$$

Setting the derivative equal to 0, we obtain exactly Eq (9.16), and consequently Eq (9.17) just as required. Note that there is a typo in Eq (9.16), $\boldsymbol{\Sigma}_k$ shoule be $\boldsymbol{\Sigma}_k^{-1}$.

Problem 9.9 Solution

We first calculate the derivative of Eq (9.40) with respect to $\boldsymbol{\Sigma}_k$:

$$\begin{aligned}
\frac{\partial \mathbb{E}_z}{\partial \boldsymbol{\Sigma}_k} &= \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \Big\{ \sum_{n=1}^{N} \gamma(z_{nk}) \ln \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \Big\} \\
&= \sum_{n=1}^{N} \gamma(z_{nk}) \frac{\partial \ln \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\partial \boldsymbol{\Sigma}_k} \\
&= \sum_{n=1}^{N} \gamma(z_{nk}) \cdot \Big[-\frac{1}{2} \boldsymbol{\Sigma}_k^{-1} + \frac{1}{2} \boldsymbol{\Sigma}_k^{-1} \mathbf{S}_{nk} \boldsymbol{\Sigma}_k^{-1} \Big]
\end{aligned}$$

As in Prob 9.6, we have defined:

$$\mathbf{S}_{nk} = (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T$$

Setting the derivative equal to 0 and rearranging it, we obtain:

$$\boldsymbol{\Sigma}_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})\,\mathbf{S}_{nk}}{\sum_{n=1}^{N} \gamma(z_{nk})} = \frac{\sum_{n=1}^{N} \gamma(z_{nk})\,\mathbf{S}_{nk}}{N_k}$$

Where N_k is given by Eq (9.18). So now we have obtained Eq (9.19) just as required. Next to maximize Eq (9.40) with respect to π_k, we still need to introduce Lagrange multiplier to enforce the summation of pi_k over k equal to 1, as in Prob 9.7:

$$L = \mathbb{E}_z + \lambda(\sum_{k=1}^{K} \pi_k - 1)$$

We calculate the derivative of L with respect to π_k and set it to 0:

$$\frac{\partial L}{\partial \pi_k} = \sum_{n=1}^{N} \frac{\gamma(z_{nk})}{\pi_k} + \lambda = 0$$

We multiply both sides by π_k and sum over k making use of the constraint Eq (9.9), yielding $\lambda = -N$ (you can see Eq (9.20)- Eq (9.22) for more details). Substituting it back into the expression, we can obtain:

$$\pi_k = \frac{1}{N} \sum_{n=1}^{N} \gamma(z_{nk}) = \frac{N_k}{N}$$

Just as Eq (9.22).

Problem 9.10 Solution

According to the property of PDF, we know that:

$$p(\mathbf{x}_b|\mathbf{x}_a) = \frac{p(\mathbf{x}_a, \mathbf{x}_b)}{p(\mathbf{x}_a)} = \frac{p(\mathbf{x})}{p(\mathbf{x}_a)} = \sum_{k=1}^{K} \frac{\pi_k}{p(\mathbf{x}_a)} \cdot p(\mathbf{x}|k)$$

Note that here $p(\mathbf{x}_a)$ can be viewed as a normalization constant used to guarantee that the integration of $p(\mathbf{x}_b|\mathbf{x}_a)$ equal to 1. Moreover, similarly, we can also obtain:

$$p(\mathbf{x}_a|\mathbf{x}_b) = \sum_{k=1}^{K} \frac{\pi_k}{p(\mathbf{x}_b)} \cdot p(\mathbf{x}|k)$$

Problem 9.11 Solution

According to the problem description, the expectation, i.e., Eq(9.40), can now be written as:

$$\mathbb{E}_z[\ln p] = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \left\{ \ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \epsilon \mathbf{I}) \right\}$$

In the M-step, we are required to maximize the expression above with respect to $\boldsymbol{\mu}_k$ and π_k. In Prob.9.8, we have already proved that $\boldsymbol{\mu}_k$ should be given by Eq (9.17):

$$\boldsymbol{\mu}_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\mathbf{x}_n \qquad (*)$$

Where N_k is given by Eq (9.18). Moreover, in this case, by analogy to Eq (9.16), $\gamma(z_{nk})$ is slightly different:

$$\gamma(z_{nk}) = \frac{\pi_k\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\epsilon\mathbf{I})}{\sum_j \pi_j\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j,\epsilon\mathbf{I})}$$

When $\epsilon \to 0$, we can obtain:

$$\sum_j \pi_j\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j,\epsilon\mathbf{I}) \approx \pi_m\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_m,\epsilon\mathbf{I}), \quad \text{where} \quad m = \text{argmin}_j||\mathbf{x}_n-\boldsymbol{\mu}_j||^2$$

To be more clear, the summation is dominated by the max of $\pi_j\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j,\epsilon\mathbf{I})$, and this term is further determined by the exponent, i.e., $-||\mathbf{x}_n-\boldsymbol{\mu}_j||^2$. Therefore, $\gamma(z_{nk})$ is given by exactly Eq (9.2), i.e., we have $\gamma(z_{nk}) = r_{nk}$. Combining with $(*)$, we can obtain exactly Eq (9.4). Next, according to Prob.9.9, π_k is given by Eq(9.22):

$$\pi_k = \frac{N_k}{N} = \frac{\sum_{n=}^{N}\gamma(z_{nk})}{N} = \frac{r_{nk}}{N}$$

In other words, π_k equals the fraction of the data points assigned to the k-th cluster.

Problem 9.12 Solution

First we calculate the mean $\boldsymbol{\mu}_k$:

$$\begin{aligned}
\boldsymbol{\mu}_k &= \int \mathbf{x}p(\mathbf{x})d\mathbf{x} \\
&= \int \mathbf{x}\sum_{k=1}^{K}\pi_k\, p(\mathbf{x}|k)d\mathbf{x} \\
&= \sum_{k=1}^{K}\pi_k\int \mathbf{x}\, p(\mathbf{x}|k)d\mathbf{x} \\
&= \sum_{k=1}^{K}\pi_k\,\boldsymbol{\mu}_k
\end{aligned}$$

Then we deal with the covariance matrix. For an arbitrary random variable $\mathbf{x}$, according to Eq (2.63) we have:

$$\begin{aligned}
\text{cov}[\mathbf{x}] &= \mathbb{E}[(\mathbf{x}-\mathbb{E}[\mathbf{x}])(\mathbf{x}-\mathbb{E}[\mathbf{x}])^T] \\
&= \mathbb{E}[\mathbf{x}\mathbf{x}^T]-\mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}]^T
\end{aligned}$$

Since $\mathbb{E}[\mathbf{x}]$ is already obtained, we only need to solve $\mathbb{E}[\mathbf{x}\mathbf{x}^T]$. First we only focus on the k-th component and rearrange the expression above, yielding:

$$\mathbb{E}_k[\mathbf{x}\mathbf{x}^T] = \text{cov}_k[\mathbf{x}] + \mathbb{E}_k[\mathbf{x}]\mathbb{E}_k[\mathbf{x}]^T = \boldsymbol{\Sigma}_k + \boldsymbol{\mu}_k\boldsymbol{\mu}_k^T$$

We further use Eq (2.62), yielding:

$$\begin{aligned}
\mathbb{E}[\mathbf{x}\mathbf{x}^T] &= \int \mathbf{x}\mathbf{x}^T \sum_{k=1}^{K} \pi_k \, p(\mathbf{x}|k) \, d\mathbf{x} \\
&= \sum_{k=1}^{K} \pi_k \int \mathbf{x}\mathbf{x}^T \, p(\mathbf{x}|k) \, d\mathbf{x} \\
&= \sum_{k=1}^{K} \pi_k \, \mathbb{E}_k[\mathbf{x}\mathbf{x}^T] \\
&= \sum_{k=1}^{K} \pi_k \, (\boldsymbol{\mu}_k\boldsymbol{\mu}_k^T + \boldsymbol{\Sigma}_k)
\end{aligned}$$

Therefore, we obtain Eq (9.50) just as required.

Problem 9.13 Solution

First, let's make this problem more clear. In a mixture of Bernoulli distribution, whose complete-data log likelihood is given by Eq (9.54) and whose model parameters are π_k and $\boldsymbol{\mu}_k$. If we want to obtain those parameters, we can adopt EM algorithm. In the E-step, we calculate $\gamma(z_{nk})$ as shown in Eq (9.56). In the M-step, we update π_k and $\boldsymbol{\mu}_k$ according to Eq (9.59) and Eq (9.60), where N_k and $\bar{\mathbf{x}_k}$ are defined in Eq (9.57) and Eq (9.58). Now let's back to this problem. The expectation of $\mathbf{x}$ is given by Eq (9.49):

$$\mathbb{E}[\mathbf{x}] = \sum_{k=1}^{K} \pi_k^{(opt)} \boldsymbol{\mu}_k^{(opt)}$$

Here $\pi_k^{(opt)}$ and $\boldsymbol{\mu}_k^{(opt)}$ are the parameters obtained when EM is converged.

Using Eq (9.58) and Eq(9.59), we can obtain:

$$\begin{aligned}
\mathbb{E}[\mathbf{x}] &= \sum_{k=1}^{K} \pi_k^{(opt)} \boldsymbol{\mu}_k^{(opt)} \\
&= \sum_{k=1}^{K} \pi_k^{(opt)} \frac{1}{N_K^{(opt)}} \sum_{n=1}^{N} \gamma(z_{nk})^{(opt)} \mathbf{x}_n \\
&= \sum_{k=1}^{K} \frac{N_k^{(opt)}}{N} \frac{1}{N_K^{(opt)}} \sum_{n=1}^{N} \gamma(z_{nk})^{(opt)} \mathbf{x}_n \\
&= \sum_{k=1}^{K} \frac{1}{N} \sum_{n=1}^{N} \gamma(z_{nk})^{(opt)} \mathbf{x}_n \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} \frac{\gamma(z_{nk})^{(opt)} \mathbf{x}_n}{N} \\
&= \sum_{n=1}^{N} \frac{\mathbf{x}_n}{N} \sum_{k=1}^{K} \gamma(z_{nk})^{(opt)} \\
&= \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n = \bar{\mathbf{x}}
\end{aligned}$$

If we set all $\boldsymbol{\mu}_k$ equal to $\widehat{\boldsymbol{\mu}}$ in initialization, in the first E-step, we can obtain:

$$\gamma(z_{nk})^{(1)} = \frac{\pi_k^{(0)} p(\mathbf{x}_n | \boldsymbol{\mu}_k = \widehat{\boldsymbol{\mu}})}{\sum_{j=1}^{K} \pi_j^{(0)} p(\mathbf{x}_n | \boldsymbol{\mu}_j = \widehat{\boldsymbol{\mu}})} = \frac{\pi_k^{(0)}}{\sum_{j=1}^{K} \pi_j^{(0)}} = \pi_k^{(0)}$$

Note that here $\widehat{\boldsymbol{\mu}}$ and $\pi_k^{(0)}$ are the initial values. In the subsequent M-step, according to Eq (9.57)-(9.60), we can obtain:

$$\boldsymbol{\mu}_k^{(1)} = \frac{1}{N_k^{(1)}} \sum_{n=1}^{N} \gamma(z_{nk})^{(1)} \mathbf{x}_n = \frac{\sum_{n=1}^{N} \gamma(z_{nk})^{(1)} \mathbf{x}_n}{\sum_{n=1}^{N} \gamma(z_{nk})^{(1)}} = \frac{\sum_{n=1}^{N} \pi_k^{(0)} \mathbf{x}_n}{\sum_{n=1}^{N} \pi_k^{(0)}} = \frac{\sum_{n=1}^{N} \mathbf{x}_n}{N}$$

And

$$\pi_k^{(1)} = \frac{N_k^{(1)}}{N} = \frac{\sum_{n=1}^{N} \gamma(z_{nk})^{(1)}}{N} = \frac{\sum_{n=1}^{N} \pi_k^{(0)}}{N} = \pi_k^{(0)}$$

In other words, in this case, after the first EM iteration, we find that the new $\boldsymbol{\mu}_k^{(1)}$ are all identical, which are all given by $\bar{\mathbf{x}}$. Moreover, the new $\pi_k^{(1)}$ are identical to their corresponding initial value $\pi_k^{(0)}$. Therefore, in the second EM iteration, we can similarly conclude that:

$$\boldsymbol{\mu}_k^{(2)} = \boldsymbol{\mu}_k^{(1)} = \bar{\mathbf{x}}, \quad \pi_k^{(2)} = \pi_k^{(1)} = \pi_k^{(0)}$$

In other words, the EM algorithm actually stops after the first iteration.

Problem 9.14 Solution

Let's follow the hint.

$$\begin{aligned} p(\mathbf{x},\mathbf{z}|\boldsymbol{\mu},\pi) &= p(\mathbf{x}|\mathbf{z},\boldsymbol{\mu})\cdot p(\mathbf{z}|\pi) \\ &= \prod_{k=1}^{K} p(\mathbf{x}|\boldsymbol{\mu}_k)^{z_k}\cdot\prod_{k=1}^{K}\pi_k^{z_k} \\ &= \prod_{k=1}^{K}\left[\pi_k p(\mathbf{x}|\boldsymbol{\mu}_k)\right]^{z_k} \end{aligned}$$

Then we marginalize over $\mathbf{z}$, yielding:

$$p(\mathbf{x}|\boldsymbol{\mu}) = \sum_{\mathbf{z}} p(\mathbf{x},\mathbf{z}|\boldsymbol{\mu},\pi) = \sum_{\mathbf{z}}\prod_{k=1}^{K}\left[\pi_k p(\mathbf{x}|\boldsymbol{\mu}_k)\right]^{z_k}$$

The summation over $\mathbf{z}$ is made up of K terms and the k-th term corresponds to $z_k = 1$ and other z_j, where $j \neq k$, equals 0. Therefore, the k-th term will simply reduce to $\pi_k p(\mathbf{x}|\boldsymbol{\mu}_k)$. Hence, performing the summation over $\mathbf{z}$ will finally give Eq (9.47) just as required. To be more clear, we summarize the aforementioned statement:

$$\begin{aligned} p(\mathbf{x}|\boldsymbol{\mu}) &= \sum_{\mathbf{z}}\prod_{k=1}^{K}\left[\pi_k p(\mathbf{x}|\boldsymbol{\mu}_k)\right]^{z_k} \\ &= \prod_{k=1}^{K}\left[\pi_k p(\mathbf{x}|\boldsymbol{\mu}_k)\right]^{z_k}\Big|_{z_1=1} + \ldots + \prod_{k=1}^{K}\left[\pi_k p(\mathbf{x}|\boldsymbol{\mu}_k)\right]^{z_k}\Big|_{z_K=1} \\ &= \pi_1 p(\mathbf{x}|\boldsymbol{\mu}_1) + \ldots + \pi_K p(\mathbf{x}|\boldsymbol{\mu}_K) \\ &= \sum_{k=1}^{K}\pi_k p(\mathbf{x}|\boldsymbol{\mu}_k) \end{aligned}$$

Problem 9.15 Solution

Noticing that π_k doesn't depend on any μ_{ki}, we can omit the first term in the open brace when calculating the derivative of Eq (9.55) with respect to μ_{ki}:

$$\begin{aligned} \frac{\partial\mathbb{E}_z[\ln p]}{\partial\mu_{ki}} &= \frac{\partial}{\partial\mu_{ki}}\sum_{n=1}^{N}\sum_{k=1}^{K}\left\{\gamma(z_{nk})\sum_{i=1}^{D}\left[x_{ni}\ln\mu_{ki}+(1-x_{ni})\ln(1-\mu_{ki})\right]\right\} \\ &= \frac{\partial}{\partial\mu_{ki}}\sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{i=1}^{D}\left\{\gamma(z_{nk})\left[x_{ni}\ln\mu_{ki}+(1-x_{ni})\ln(1-\mu_{ki})\right]\right\} \\ &= \sum_{n=1}^{N}\frac{\partial}{\partial\mu_{ki}}\left\{\gamma(z_{nk})\left[x_{ni}\ln\mu_{ki}+(1-x_{ni})\ln(1-\mu_{ki})\right]\right\} \\ &= \sum_{n=1}^{N}\gamma(z_{nk})\left(\frac{x_{ni}}{\mu_{ki}}-\frac{1-x_{ni}}{1-\mu_{ki}}\right) \\ &= \sum_{n=1}^{N}\gamma(z_{nk})\frac{x_{ni}-\mu_{ki}}{\mu_{ki}(1-\mu_{ki})} \end{aligned}$$

Setting the derivative equal to 0, we can obtain:

$$\mu_{ki} = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) x_{ni}}{\sum_{n=1}^{N} \gamma(z_{nk})} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) x_{ni}$$

Where N_k is defined as Eq (9.57). If we group all the μ_{ki} as a column vector, i.e., $\boldsymbol{\mu}_k = [\mu_{k1}, \mu_{k2}, ..., \mu_{kD}]^T$, we will obtain Eq (9.59) just as required.

Problem 9.16 Solution

We follow the hint beginning by introducing a Lagrange multiplier:

$$L = \mathbb{E}_z[\ln p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\mu}, \boldsymbol{\pi})] + \lambda (\sum_{k=1}^{K} \pi_k - 1)$$

We calculate the derivative of L with respect to π_k and then set it equal to 0:

$$\frac{\partial L}{\partial \pi_k} = \sum_{n=1}^{N} \frac{\gamma(z_{nk})}{\pi_k} + \lambda = 0 \qquad (*)$$

Here $\mathbb{E}_z[\ln p]$ is given by Eq (9.55). We first multiply both sides of the expression by π_k and then adopt summation with respect to k, which gives:

$$\sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) + \sum_{k=1}^{K} \lambda \pi_k = 0$$

Noticing that $\sum_{k=1}^{K} \pi_k$ equals 1, we can obtain:

$$\lambda = -\sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk})$$

Finally, substituting it back into $(*)$ and rearranging it, we can obtain:

$$\pi_k = -\frac{\sum_{k=1}^{K} \gamma(z_{nk})}{\lambda} = \frac{\sum_{k=1}^{K} \gamma(z_{nk})}{\sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk})} = \frac{N_k}{N}$$

Where N_k is defined by Eq (9.57) and N is the summation of N_k over k, and also equal to the number of data points.

Problem 9.17 Solution

The incomplete-data log likelihood is given by Eq (9.51), and $p(\mathbf{x}_n | \boldsymbol{\mu}_k)$ lies in the interval [0, 1], which can be easily verified by its definition, i.e., Eq (9.44). Therefore, we can obtain:

$$\ln p(\mathbf{X} | \boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \ln \{ \sum_{k=1}^{K} \pi_k p(\mathbf{x}_n | \boldsymbol{\mu}_k) \} \leq \sum_{n=1}^{N} \ln \{ \sum_{k=1}^{K} \pi_k \times 1 \} \leq \sum_{n=1}^{N} \ln 1 = 0$$

Where we have used the fact that the logarithm is monotonic increasing, and that the summation of π_k over k equals 1. Moreover, if we want to achieve the equality, we need $p(\mathbf{x}_n|\boldsymbol{\mu}_k)$ equal to 1 for all $n = 1,2,...,N$. However, this is hardly possible.

To illustrate this, suppose that $p(\mathbf{x}_n|\boldsymbol{\mu}_k)$ equals 1 for all data points. Without loss of generality, consider two data points $\mathbf{x}_1 = [x_{11}, x_{12}, ..., x_{1D}]^T$ and $\mathbf{x}_2 = [x_{21}, x_{22}, ..., x_{2D}]^T$, whose i-th entries are different. We further assume $x_{1i} = 1$ and $x_{2i} = 0$ since x_i is a binary variable. According to Eq (9.44), if we want $p(\mathbf{x}_1|\boldsymbol{\mu}_k) = 1$, we must have $\mu_i = 1$ (otherwise it muse be less than 1). However, this will lead $p(\mathbf{x}_2|\boldsymbol{\mu}_k)$ equal to 0 since there is a term $1 - \mu_i = 0$ in the product shown in Eq (9.44).

Therefore, when the data set is pathological, we will achieve this singularity point by adopting EM. Note that in the main text, the author states that the condition should be pathological initialization. This is also true. For instance, in the extreme case, when the data set is not pathological, if we initialize one π_k equal to 1 and others all 0, and some of μ_i to 1 and others 0, we may also achieve the singularity.

Problem 9.18 Solution

In Prob.9.4, we have proved that if we want to maximize the posterior by EM, the only modification is that in the M-step, we need to maximize $Q'(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) + \ln p(\boldsymbol{\theta})$. Here $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$ has already been given by $\mathbb{E}_z[\ln p]$, i.e., Eq (9.55). Therefore, we derive for $\ln p(\boldsymbol{\theta})$. Note that $\ln p(\boldsymbol{\theta})$ is made up of two parts:(i) the prior for $\boldsymbol{\mu}_k$ and (ii) the prior for $\boldsymbol{\pi}$, we begin by dealing with the first part. Here we assume the Beta prior for μ_{ki}, where k is fixed, is the same, i.e.,:

$$p(\mu_{ki}|a_k, b_k) = \frac{\Gamma(a_k + b_k)}{\Gamma(a_k)\Gamma(b_k)} \mu_{ki}^{a_k - 1} (1 - \mu_{ki})^{b_k - 1}, \quad i = 1, 2, ..., D$$

Therefore, the contribution of this Beta prior to $\ln p(\boldsymbol{\theta})$ should be given by:

$$\sum_{k=1}^{K} \sum_{i=1}^{D} (a_i - 1) \ln \mu_{ki} + (b_i - 1) \ln(1 - \mu_{ki})$$

One thing worthy mentioned is that since we will maximize $Q'(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$ with respect to $\boldsymbol{\pi}, \boldsymbol{\mu}_k$, we can omit the terms which do not depend on $\boldsymbol{\pi}, \boldsymbol{\mu}_k$, such as $\Gamma(a_k + b_k)/\Gamma(a_k)\Gamma(b_k)$. Then we deal with the second part. According to Eq (2.38), we can obtain:

$$p(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)...\Gamma(\alpha_K)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}$$

Therefore, the contribution of the Dirichlet prior to $\ln p(\boldsymbol{\theta})$ should be given by:

$$\sum_{k=1}^{K} (\alpha_k - 1) \ln \pi_k$$

Therefore, now $Q^{'}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$ can be written as:

$$Q^{'}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \mathbb{E}_z[\ln p] + \sum_{k=1}^{K}\sum_{i=1}^{D}\Big[(a_i - 1)\ln \mu_{ki} + (b_i - 1)\ln(1-\mu_{ki})\Big] + \sum_{k=1}^{K}(\alpha_k - 1)\ln \pi_k$$

Similarly, we calculate the derivative of $Q^{'}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$ with respect to μ_{ki}. This can be simplified by reusing the deduction in Prob.9.15:

$$\begin{aligned}
\frac{\partial Q^{'}}{\partial \mu_{ki}} &= \frac{\partial \mathbb{E}_z[\ln p]}{\partial \mu_{ki}} + \frac{a_i - 1}{\mu_{ki}} - \frac{b_i - 1}{1-\mu_{ki}} \\
&= \sum_{n=1}^{N}\gamma(z_{nk})(\frac{x_{ni}}{\mu_{ki}} - \frac{1-x_{ni}}{1-\mu_{ki}}) + \frac{a_i - 1}{\mu_{ki}} - \frac{b_i - 1}{1-\mu_{ki}} \\
&= \frac{\sum_{n=1}^{N} x_{ni}\cdot\gamma(z_{nk}) + a_i - 1}{\mu_{ki}} - \frac{\sum_{n=1}^{N}(1-x_{ni})\gamma(z_{nk}) + b_i - 1}{1-\mu_{ki}} \\
&= \frac{N_k \bar{x}_{ki} + a_i - 1}{\mu_{ki}} - \frac{N_k - N_k\bar{x}_{ki} + b_i - 1}{1-\mu_{ki}}
\end{aligned}$$

Note that here $\bar{x}_{ki}$ is defined as the i-th entry of $\bar{x}_k$ defined in Eq (9.58). To be more clear, we have used Eq (9.57) and Eq (9.58) in the last step:

$$\sum_{n=1}^{N} x_{ni}\cdot\gamma(z_{nk}) = N_k \cdot \Big[\frac{1}{N_k}\sum_{n=1}^{N} x_{ni}\cdot\gamma(z_{nk})\Big] = N_k \cdot \bar{x}_{ki}$$

Setting the derivative equal to 0 and rearranging it, we can obtain:

$$\mu_{ki} = \frac{N_k\bar{x}_{ki} + a_i - 1}{N_k + a_i - 1 + b_i - 1}$$

Next we maximize $Q^{'}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$ with respect to $\boldsymbol{\pi}$. By analogy to Prob.9.16, we introduce Lagrange multiplier:

$$L \propto \mathbb{E}_z + \sum_{k=1}^{K}(\alpha_k - 1)\ln \pi_k + \lambda(\sum_{k=1}^{K}\pi_k - 1)$$

Note that the second term on the right hand side of $Q^{'}$ in its definition has been omitted, since that term can be viewed as a constant with regard to $\boldsymbol{\pi}$. We then calculate the derivative of L with respect to π_k by taking advantage of Prob.9.16:

$$\frac{\partial L}{\partial \pi_k} = \sum_{n=1}^{N}\frac{\gamma(z_{nk})}{\pi_k} + \frac{\alpha_k - 1}{\pi_k} + \lambda = 0$$

Similarly, We first multiply both sides of the expression by π_k and then adopt summation with respect to k, which gives:

$$\sum_{k=1}^{K}\sum_{n=1}^{N}\gamma(z_{nk}) + \sum_{k=1}^{K}(\alpha_k - 1) + \sum_{k=1}^{K}\lambda\pi_k = 0$$

Noticing that $\sum_{k=1}^{K} \pi_k$ equals 1, we can obtain:

$$\lambda = -\sum_{k=1}^{K} N_k - \sum_{k=1}^{K} (\alpha_k - 1) = -N - \alpha_0 + K$$

Here we have used Eq (2.39). Substituting it back into the derivative, we can obtain:

$$\pi_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) + \alpha_k - 1}{-\lambda} = \frac{N_k + \alpha_k - 1}{N + \alpha_0 - K}$$

It is not difficult to show that if N is large, the update formula for $\boldsymbol{\pi}$ and $\boldsymbol{\mu}$ in this case (MAP), will reduce to the results given in the main text (MLE).

Problem 9.19 Solution

We first introduce a latent variable $\mathbf{z} = [z_1, z_2, ..., z_K]^T$, only one of which equals 1 and others all 0. The conditional distribution of $\mathbf{x}$ is given by:

$$p(\mathbf{x}|\mathbf{z}, \boldsymbol{\mu}) = \prod_{k=1}^{K} p(\mathbf{x}|\boldsymbol{\mu}_k)^{z_k}$$

The distribution of the latent variable is given by:

$$p(\mathbf{z}|\boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{z_k}$$

If we follow the same procedure as in Prob.9.14, we can show that Eq (9.84) holds. In other words, the introduction of the latent variable is valid. Therefore, according to Bayes' Theorem, we can obtain:

$$p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\pi}) = \prod_{n=1}^{N} p(\mathbf{z}_n|\boldsymbol{\pi}) p(\mathbf{x}_n|\mathbf{z}_n, \boldsymbol{\mu}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \Big[\pi_k p(\mathbf{x}|\boldsymbol{\mu})\Big]^{z_{nk}}$$

We further use Eq (9.85), which gives:

$$\begin{aligned} \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\pi}) &= \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \ln \Big[\pi_k \prod_{d=1}^{D} \prod_{j=1}^{M} \mu_{kij}^{x_{nij}}\Big] \\ &= \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \Big[\ln \pi_k + \sum_{d=1}^{D} \sum_{j=1}^{M} x_{nij} \ln \mu_{kij}\Big] \end{aligned}$$

Similarly, in the E-step, the responsibilities are evaluated using Bayes' theorem, which gives:

$$\gamma(z_{nk}) = \mathbb{E}[z_{nk}] = \frac{\pi_k p(\mathbf{x}_n|\boldsymbol{\mu}_k)}{\sum_{j=1}^{K} \pi_j p(\mathbf{x}_n|\boldsymbol{\mu}_j)}$$

Next, in the M-step, we are required to maximize $\mathbb{E}_z[\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\mu},\boldsymbol{\pi})]$ with respect to $\boldsymbol{\pi}$ and $\boldsymbol{\mu}_k$, where $\mathbb{E}_z[\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\mu},\boldsymbol{\pi})]$ is given by:

$$\mathbb{E}_z[\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\mu},\boldsymbol{\pi})] = \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\Big[\ln\pi_k + \sum_{i=1}^{D}\sum_{j=1}^{M}x_{nij}\ln\mu_{kij}\Big]$$

Notice that there exists two constraints: (i) the summation of π_k over k equals 1, and (ii) the summation of μ_{kij} over j equals 1 for any k and i, we need to introduce Lagrange multiplier:

$$L = \mathbb{E}_z[\ln p] + \lambda(\sum_{k=1}^{K}\pi_k - 1) + \sum_{k=1}^{K}\sum_{i=1}^{D}\eta_{ki}(\sum_{j=1}^{M}\mu_{kij} - 1)$$

First we maximize L with respect to π_k. This is actually identical to the case in the main text. To be more clear, we calculate the derivative of L with respect to π_k:

$$\frac{\partial L}{\partial \pi_k} = \sum_{n=1}^{N}\frac{\gamma(z_{nk})}{\pi_k} + \lambda$$

As in Prob.9.16, we can obtain:

$$\pi_k = \frac{N_k}{N}$$

Where N_k is defined as:

$$N_k = \sum_{n=1}^{N}\gamma(z_{nk})$$

N is the summation of N_k over k, and also equals the number of data points. Then we calculate the derivative of L with respect to μ_{kij}:

$$\frac{\partial L}{\partial \mu_{kij}} = \sum_{n=1}^{N}\frac{\gamma(z_{nk})x_{nij}}{\mu_{kij}} + \eta_{ki}$$

We set it to 0 and multiply both sides by μ_{kij}, which gives:

$$\sum_{n=1}^{N}\gamma(z_{nk})x_{nij} + \eta_{ki}\mu_{kij} = 0$$

By analogy to deriving π_k, an intuitive idea is to perform summation for the above expression over j and hence we can use the constraint $\sum_j \mu_{kij} = 1$.

$$\eta_{ki} = -\sum_{j=1}^{M}\sum_{n=1}^{N}\gamma(z_{nk})x_{nij} = -\sum_{n=1}^{N}\gamma(z_{nk})\Big[\sum_{j=1}^{M}x_{nij}\Big] = -\sum_{n=1}^{N}\gamma(z_{nk}) = -N_k$$

Where we have used the fact that $\sum_j x_{nij} = 1$. Substituting back into the derivative, we can obtain:

$$\mu_{kij} = -\frac{\sum_{n=1}^{N} \gamma(z_{nk}) x_{nij}}{\eta_{ki}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) x_{nij}$$

Problem 9.20 Solution

We first calculate the derivative of Eq (9.62) with respect to α and set it to 0:

$$\frac{\partial E[\ln p]}{\partial \alpha} = \frac{M}{2} \frac{1}{2\pi} \frac{2\pi}{\alpha} - \frac{\mathbb{E}[\mathbf{w}^T \mathbf{w}]}{2} = 0$$

We rearrange the equation above, which gives:

$$\alpha = \frac{M}{\mathbb{E}[\mathbf{w}^T \mathbf{w}]} \tag{$*$}$$

Therefore, we now need to calculate the expectation $\mathbb{E}[\mathbf{w}^T \mathbf{w}]$. Notice that the posterior has already been given by Eq (3.49):

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{m}_N, \mathbf{S}_N)$$

To calculate $\mathbb{E}[\mathbf{w}^T \mathbf{w}]$, here we write down an property for a Gaussian random variable: if $\mathbf{x} \sim \mathcal{N}(\mathbf{m}, \mathbf{\Sigma})$, we have:

$$\mathbb{E}[\mathbf{x}^T \mathbf{A} \mathbf{x}] = \mathrm{Tr}[\mathbf{A}\mathbf{\Sigma}] + \mathbf{m}^T \mathbf{A} \mathbf{m}$$

This property has been shown in Eq(378) in 'the Matrix Cookbook'. Utilizing this property, we can obtain:

$$\mathbb{E}[\mathbf{w}^T \mathbf{w}] = \mathrm{Tr}[\mathbf{S}_N] + \mathbf{m}_N^T \mathbf{m}_N$$

Substituting it back into $(*)$, we obtain what is required.

Problem 9.21 Solution

We calculate the derivative of Eq (9.62) with respect to β and set it equal to 0:

$$\frac{\partial \ln p}{\partial \beta} = \frac{N}{2} \frac{1}{2\pi} \frac{2\pi}{\beta} - \frac{1}{2} \sum_{n=1}^{N} \mathbb{E}[(t_n - \mathbf{w}^T \boldsymbol{\phi}_n)^2] = 0$$

Rearranging it, we obtain:

$$\beta = \frac{N}{\sum_{n=1}^{N} \mathbb{E}[(t_n - \mathbf{w}^T \boldsymbol{\phi}_n)^2]}$$

Therefore, we are required to calculate the expectation. To be more clear, this expectation is with respect to the posterior defined by Eq (3.49):

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{m}_N, \mathbf{S}_N)$$

We expand the expectation:

$$\begin{aligned}
\mathbb{E}[(t_n - \mathbf{w}^T\boldsymbol{\phi}_n)^2] &= \mathbb{E}[t_n^2 - 2t_n \cdot \mathbf{w}^T\boldsymbol{\phi}_n + \mathbf{w}^T\boldsymbol{\phi}_n\boldsymbol{\phi}_n^T\mathbf{w}] \\
&= \mathbb{E}[t_n^2] - \mathbb{E}[2t_n \cdot \mathbf{w}^T\boldsymbol{\phi}_n] + \mathbb{E}[\mathbf{w}^T(\boldsymbol{\phi}_n\boldsymbol{\phi}_n^T)\mathbf{w}] \\
&= t_n^2 - 2t_n \cdot \mathbb{E}[\boldsymbol{\phi}_n^T\mathbf{w}] + \mathrm{Tr}[\boldsymbol{\phi}_n\boldsymbol{\phi}_n^T\mathbf{S}_N] + \mathbf{m}_N^T\boldsymbol{\phi}_n\boldsymbol{\phi}_n^T\mathbf{m}_N \\
&= t_n^2 - 2t_n\boldsymbol{\phi}_n^T \cdot \mathbb{E}[\mathbf{w}] + \mathrm{Tr}[\boldsymbol{\phi}_n\boldsymbol{\phi}_n^T\mathbf{S}_N] + \mathbf{m}_N^T\boldsymbol{\phi}_n\boldsymbol{\phi}_n^T\mathbf{m}_N \\
&= t_n^2 - 2t_n\boldsymbol{\phi}_n^T\mathbf{m}_N + \mathrm{Tr}[\boldsymbol{\phi}_n\boldsymbol{\phi}_n^T\mathbf{S}_N] + \mathbf{m}_N^T\boldsymbol{\phi}_n\boldsymbol{\phi}_n^T\mathbf{m}_N \\
&= (t_n - \mathbf{m}_N^T\boldsymbol{\phi}_N)^2 + \mathrm{Tr}[\boldsymbol{\phi}_n\boldsymbol{\phi}_n^T\mathbf{S}_N]
\end{aligned}$$

Substituting it back into the derivative, we can obtain:

$$\begin{aligned}
\frac{1}{\beta} &= \frac{1}{N}\sum_{n=1}^{N}\left\{(t_n - \mathbf{m}_N^T\boldsymbol{\phi}_N)^2 + \mathrm{Tr}[\boldsymbol{\phi}_n\boldsymbol{\phi}_n^T\mathbf{S}_N]\right\} \\
&= \frac{1}{N}\left\{||\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}_N||^2 + \mathrm{Tr}[\boldsymbol{\Phi}^T\boldsymbol{\Phi}\mathbf{S}_N]\right\}
\end{aligned}$$

Note that in the last step, we have performed vectorization. Here the j-th row of $\boldsymbol{\Phi}$ is given by $\boldsymbol{\phi}_j$, identical to the definition given in Chapter 3.

Problem 9.22 Solution

First let's expand the complete-data log likelihood using Eq (7.79), Eq (7.80) and Eq (7.76).

$$\begin{aligned}
\ln p(\mathbf{t}|\mathbf{X},\mathbf{w},\beta)p(\mathbf{w}|\boldsymbol{\alpha}) &= \ln p(\mathbf{t}|\mathbf{X},\mathbf{w},\beta) + \ln p(\mathbf{w}|\boldsymbol{\alpha}) \\
&= \sum_{n=1}^{N}\ln p(t_n|x_n,\mathbf{w},\beta^{-1}) + \sum_{i=1}^{M}\ln\mathcal{N}(w_i|0,\alpha_i^{-1}) \\
&= \sum_{n=1}^{N}\ln\mathcal{N}(t_n|\mathbf{w}^T\boldsymbol{\phi}_n,\beta^{-1}) + \sum_{i=1}^{M}\ln\mathcal{N}(w_i|0,\alpha_i^{-1}) \\
&= \frac{N}{2}\ln\frac{\beta}{2\pi} - \frac{\beta}{2}\sum_{n=1}^{N}(t_n - \mathbf{w}^T\boldsymbol{\phi}_n)^2 + \frac{1}{2}\sum_{i=1}^{M}\ln\frac{\alpha_i}{2\pi} - \sum_{i=1}^{M}\frac{\alpha_i}{2}w_i^2
\end{aligned}$$

Therefore, the expectation of the complete-data log likelihood with respect to the posterior of $\mathbf{w}$ equals:

$$\mathbb{E}_{\mathbf{w}}[\ln p] = \frac{N}{2}\ln\frac{\beta}{2\pi} - \frac{\beta}{2}\sum_{n=1}^{N}\mathbb{E}_{\mathbf{w}}[(t_n - \mathbf{w}^T\boldsymbol{\phi}_n)^2] + \frac{1}{2}\sum_{i=1}^{M}\ln\frac{\alpha_i}{2\pi} - \sum_{i=1}^{M}\frac{\alpha_i}{2}\mathbb{E}_{\mathbf{w}}[w_i^2]$$

We calculate the derivative of $\mathbb{E}_{\mathbf{w}}[\ln p]$ with respect to α_i and set it to 0:

$$\frac{\partial\mathbb{E}_{\mathbf{w}}[\ln p]}{\partial\alpha_i} = \frac{1}{2}\frac{1}{2\pi}\frac{2\pi}{\alpha_i} - \frac{1}{2}\mathbb{E}_{\mathbf{w}}[w_i^2] = 0$$

Rearranging it, we can obtain:

$$\alpha_i = \frac{1}{\mathbb{E}_{\mathbf{w}}[w_i^2]} = \frac{1}{\mathbb{E}_{\mathbf{w}}[\mathbf{w}\mathbf{w}^T]_{(i,i)}}$$

Here the subscript (i,i) represents the entry on the i-th row and i-th column of the matrix $\mathbb{E}_{\mathbf{w}}[\mathbf{w}\mathbf{w}^T]$. So now, we are required to calculate the expectation. To be more clear, this expectation is with respect to the posterior defined by Eq (7.81):

$$p(\mathbf{w}|\mathbf{t},\mathbf{X},\boldsymbol{\alpha},\beta) = \mathcal{N}(\mathbf{m},\boldsymbol{\Sigma})$$

Here we use Eq (377) described in 'the Matrix Cookbook'. We restate it here: if $\mathbf{w} \sim \mathcal{N}(\mathbf{m},\boldsymbol{\Sigma})$, we have:

$$\mathbb{E}[\mathbf{w}\mathbf{w}^T] = \boldsymbol{\Sigma} + \mathbf{m}\mathbf{m}^T$$

According to this equation, we can obtain:

$$\alpha_i = \frac{1}{\mathbb{E}_{\mathbf{w}}[\mathbf{w}\mathbf{w}^T]_{(i,i)}} = \frac{1}{(\boldsymbol{\Sigma} + \mathbf{m}\mathbf{m}^T)_{(i,i)}} = \frac{1}{\Sigma_{ii} + m_i^2}$$

Now We calculate the derivative of $\mathbb{E}_{\mathbf{w}}[\ln p]$ with respect to β and set it to 0:

$$\frac{\partial \mathbb{E}_{\mathbf{w}}[\ln p]}{\partial \beta} = \frac{N}{2}\frac{1}{2\pi}\frac{2\pi}{\beta} - \frac{1}{2}\sum_{n=1}^{N}\mathbb{E}_{\mathbf{w}}[(t_n - \mathbf{w}^T\boldsymbol{\phi}_n)^2] = 0$$

Rearranging it, we obtain:

$$\beta^{(new)} = \frac{N}{\sum_{n=1}^{N}\mathbb{E}_{\mathbf{w}}[(t_n - \mathbf{w}^T\boldsymbol{\phi}_n)^2]}$$

Therefore, we are required to calculate the expectation. By analogy to the deduction in Prob.9.21, we can obtain:

$$\begin{aligned}\frac{1}{\beta^{(new)}} &= \frac{1}{N}\sum_{n=1}^{N}\left\{(t_n - \mathbf{m}^T\boldsymbol{\phi}_N)^2 + \mathrm{Tr}[\boldsymbol{\phi}_n\boldsymbol{\phi}_n^T\boldsymbol{\Sigma}]\right\} \\ &= \frac{1}{N}\left\{||\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}||^2 + \mathrm{Tr}[\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{\Sigma}]\right\}\end{aligned}$$

To make it consistent with Eq (9.68), let's first prove a statement:

$$(\beta^{-1}\mathbf{A} + \boldsymbol{\Phi}^T\boldsymbol{\Phi})\boldsymbol{\Sigma} = \beta^{-1}\mathbf{I}$$

This can be easily shown by substituting $\boldsymbol{\Sigma}$, i.e., Eq(7.83), back into the expression:

$$(\beta^{-1}\mathbf{A} + \boldsymbol{\Phi}^T\boldsymbol{\Phi})\,\boldsymbol{\Sigma} = (\beta^{-1}\mathbf{A} + \boldsymbol{\Phi}^T\boldsymbol{\Phi})\,(\mathbf{A} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1} = \beta^{-1}\mathbf{I}$$

Now we start from this statement and rearrange it, which gives:

$$\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{\Sigma} = \beta^{-1}\mathbf{I} - \beta^{-1}\mathbf{A}\boldsymbol{\Sigma} = \beta^{-1}(\mathbf{I} - \mathbf{A}\boldsymbol{\Sigma})$$

Substituting back into the expression for $\beta^{(new)}$:

$$\begin{aligned}
\frac{1}{\beta^{(new)}} &= \frac{1}{N}\left\{||\mathbf{t}-\mathbf{\Phi m}||^2 + \text{Tr}[\mathbf{\Phi}^T\mathbf{\Phi\Sigma}]\right\} \\
&= \frac{1}{N}\left\{||\mathbf{t}-\mathbf{\Phi m}||^2 + \text{Tr}[\beta^{-1}(\mathbf{I}-\mathbf{A\Sigma})]\right\} \\
&= \frac{1}{N}\left\{||\mathbf{t}-\mathbf{\Phi m}||^2 + \beta^{-1}\text{Tr}[\mathbf{I}-\mathbf{A\Sigma}]\right\} \\
&= \frac{1}{N}\left\{||\mathbf{t}-\mathbf{\Phi m}||^2 + \beta^{-1}\sum_i(1-\alpha_i\Sigma_{ii})\right\} \\
&= \frac{||\mathbf{t}-\mathbf{\Phi m}||^2 + \beta^{-1}\sum_i\gamma_i}{N}
\end{aligned}$$

Here we have defined $\gamma_i = 1-\alpha_i\Sigma_{ii}$ as in Eq (7.89). Note that there is a typo in Eq (9.68), $\mathbf{m}_N$ should be $\mathbf{m}$.

Problem 9.23 Solution

Some clarifications must be made here, Eq (7.87)-(7.88) only gives the same stationary points, i.e., the same $\alpha^\star$ and $\beta^\star$, as those given by Eq (9.67)-(9.68). However, the hyper-parameters estimated at some specific iteration may not be the same by those two different methods.

When convergence is reached, Eq (7.87) can be written as:

$$\alpha^\star = \frac{1-\alpha^\star\Sigma_{ii}}{m_i^2}$$

Rearranging it, we can obtain:

$$\alpha^\star = \frac{1}{m_i^2+\Sigma_{ii}}$$

This is identical to Eq (9.67). When convergence is reached, Eq (9.68) can be written as:

$$(\beta^\star)^{-1} = \frac{||\mathbf{t}-\mathbf{\Phi m}||^2 + (\beta^\star)^{-1}\sum_i\gamma_i}{N}$$

Rearranging it, we can obtain:

$$(\beta^\star)^{-1} = \frac{||\mathbf{t}-\mathbf{\Phi m}||^2}{N-\sum_i\gamma_i}$$

This is identical to Eq (7.88).

Problem 9.24 Solution

We substitute Eq (9.71) and Eq (9.72) into Eq (9.70):

$$\begin{aligned} L(q,\boldsymbol{\theta}) + \mathrm{KL}(q||p) &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \Big\{ \ln \frac{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} - \ln \frac{p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})}{q(\mathbf{Z})} \Big\} \\ &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \Big\{ \ln \frac{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})}{p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})} \Big\} \\ &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln p(\mathbf{X}|\boldsymbol{\theta}) \\ &= \ln p(\mathbf{X}|\boldsymbol{\theta}) \end{aligned}$$

Note that in the last step, we have used the fact that $\ln p(\mathbf{X}|\boldsymbol{\theta})$ doesn't depend on $\mathbf{Z}$, and that the summation of $q(\mathbf{Z})$ over $\mathbf{Z}$ equal to 1 because $q(\mathbf{Z})$ is a PDF.

Problem 9.25 Solution

We calculate the derivative of Eq (9.71) with respect to $\boldsymbol{\theta}$, given $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(\mathrm{old})})$:

$$\begin{aligned} \frac{\partial L(q,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \frac{\partial}{\partial \boldsymbol{\theta}} \Big\{ \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(\mathrm{old})}) \ln \frac{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})}{p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(\mathrm{old})})} \Big\} \\ &= \frac{\partial}{\partial \boldsymbol{\theta}} \Big\{ \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(\mathrm{old})}) \ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta}) - \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(\mathrm{old})}) \ln p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(\mathrm{old})}) \Big\} \\ &= \frac{\partial}{\partial \boldsymbol{\theta}} \Big\{ \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(\mathrm{old})}) \ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta}) \Big\} \\ &= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(\mathrm{old})}) \frac{\partial \ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\ &= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(\mathrm{old})}) \frac{1}{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})} \frac{\partial p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\ &= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(\mathrm{old})}) \frac{1}{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})} \frac{\partial p(\mathbf{X}|\boldsymbol{\theta}) \cdot p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\ &= \sum_{\mathbf{Z}} \frac{p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(\mathrm{old})})}{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})} \Big[p(\mathbf{X}|\boldsymbol{\theta}) \frac{\partial p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}) \frac{\partial p(\mathbf{X}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \Big] \end{aligned}$$

We evaluate this derivative at $\boldsymbol{\theta} = \boldsymbol{\theta}^{\text{old}}$:

$$\begin{aligned}
\frac{\partial L(q,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}^{\text{old}}} &= \Big\{ \sum_{\mathbf{Z}} \frac{p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(\text{old})})}{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})} \Big[p(\mathbf{X}|\boldsymbol{\theta}) \frac{\partial p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}) \frac{\partial p(\mathbf{X}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \Big] \Big\} \Big|_{\boldsymbol{\theta}^{\text{old}}} \\
&= \sum_{\mathbf{Z}} \frac{p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(\text{old})})}{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta}^{(\text{old})})} \Big[p(\mathbf{X}|\boldsymbol{\theta}^{(\text{old})}) \frac{\partial p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}^{(\text{old})}} + p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(\text{old})}) \frac{\partial p(\mathbf{X}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}^{(\text{old})}} \Big] \\
&= \sum_{\mathbf{Z}} \frac{1}{p(\mathbf{X}|\boldsymbol{\theta}^{(\text{old})})} \Big[p(\mathbf{X}|\boldsymbol{\theta}^{(\text{old})}) \frac{\partial p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}^{(\text{old})}} + p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(\text{old})}) \frac{\partial p(\mathbf{X}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}^{(\text{old})}} \Big] \\
&= \sum_{\mathbf{Z}} \frac{\partial p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}^{(\text{old})}} + \sum_{\mathbf{Z}} \frac{p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(\text{old})})}{p(\mathbf{X}|\boldsymbol{\theta}^{(\text{old})})} \cdot \frac{\partial p(\mathbf{X}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}^{(\text{old})}} \\
&= \sum_{\mathbf{Z}} \frac{\partial p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}^{(\text{old})}} + \frac{1}{p(\mathbf{X}|\boldsymbol{\theta}^{(\text{old})})} \cdot \frac{\partial p(\mathbf{X}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}^{(\text{old})}} \\
&= \sum_{\mathbf{Z}} \frac{\partial p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}^{(\text{old})}} + \frac{\partial \ln p(\mathbf{X}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}^{(\text{old})}} \\
&= \Big\{ \frac{\partial}{\partial \boldsymbol{\theta}} \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}) \Big\} \Big|_{\boldsymbol{\theta}^{(\text{old})}} + \frac{\partial \ln p(\mathbf{X}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}^{(\text{old})}} \\
&= \frac{\partial 1}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}^{(\text{old})}} + \frac{\partial \ln p(\mathbf{X}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}^{(\text{old})}} \\
&= \frac{\partial \ln p(\mathbf{X}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}^{(\text{old})}}
\end{aligned}$$

This problem can be much easier to prove if we view it from the perspective of KL divergence. Note that when $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(\text{old})})$, the KL divergence vanishes, and that in general KL divergence is less or equal to zero. Therefore, we must have:

$$\frac{\partial KL(q||p)}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}^{(\text{old})}} = 0$$

Otherwise, there exists a point $\boldsymbol{\theta}$ in the neighborhood near $\boldsymbol{\theta}^{(\text{old})}$ which leads the KL divergence less than 0. Then using Eq (9.70), it is trivial to prove.

Problem 9.26 Solution

From Eq (9.18), we have:

$$N_k^{\text{old}} = \sum_n \gamma^{\text{old}}(z_{nk})$$

If now we just re-evaluate the responsibilities for one data point $\mathbf{x}_m$, we can obtain:

$$\begin{aligned}
N_k^{\text{new}} &= \sum_{n \neq m} \gamma^{\text{old}}(z_{nk}) + \gamma^{\text{new}}(z_{mk}) \\
&= \sum_n \gamma^{\text{old}}(z_{nk}) + \gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk}) \\
&= N_k^{\text{old}} + \gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk})
\end{aligned}$$

Similarly, according to Eq (9.17), we can obtain:

$$\begin{aligned}
\boldsymbol{\mu}_k^{\text{new}} &= \frac{1}{N_k^{\text{new}}} \sum_{n \neq m} \gamma^{\text{old}}(z_{nk}) \mathbf{x}_n + \frac{\gamma^{\text{new}}(z_{mk}) \mathbf{x}_m}{N_k^{\text{new}}} \\
&= \frac{1}{N_k^{\text{new}}} \sum_{n} \gamma^{\text{old}}(z_{nk}) \mathbf{x}_n + \frac{\gamma^{\text{new}}(z_{mk}) \mathbf{x}_m}{N_k^{\text{new}}} - \frac{\gamma^{\text{old}}(z_{mk}) \mathbf{x}_m}{N_k^{\text{new}}} \\
&= \frac{N_k^{\text{old}}}{N_k^{\text{new}}} \frac{1}{N_k^{\text{old}}} \sum_{n} \gamma^{\text{old}}(z_{nk}) \mathbf{x}_n + \frac{\gamma^{\text{new}}(z_{mk}) \mathbf{x}_m}{N_k^{\text{new}}} - \frac{\gamma^{\text{old}}(z_{mk}) \mathbf{x}_m}{N_k^{\text{new}}} \\
&= \frac{N_k^{\text{old}}}{N_k^{\text{new}}} \boldsymbol{\mu}_k^{\text{old}} + \left[\gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk}) \right] \frac{\mathbf{x}_m}{N_k^{\text{new}}} \\
&= \boldsymbol{\mu}_k^{\text{old}} - \frac{N_k^{\text{new}} - N_k^{\text{old}}}{N_k^{\text{new}}} \boldsymbol{\mu}_k^{\text{old}} + \left[\gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk}) \right] \frac{\mathbf{x}_m}{N_k^{\text{new}}} \\
&= \boldsymbol{\mu}_k^{\text{old}} - \frac{\gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk})}{N_k^{\text{new}}} \boldsymbol{\mu}_k^{\text{old}} + \left[\gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk}) \right] \frac{\mathbf{x}_m}{N_k^{\text{new}}} \\
&= \boldsymbol{\mu}_k^{\text{old}} + \frac{\gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk})}{N_k^{\text{new}}} \cdot \left(\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{old}} \right)
\end{aligned}$$

Just as required.

Problem 9.27 Solution

By analogy to the previous problem, we use Eq (9.24)-Eq(9.27), beginning by first deriving an update formula for mixing coefficients π_k:

$$\begin{aligned}
\pi_k^{\text{new}} &= \frac{N_k^{\text{new}}}{N} = \frac{1}{N} \left\{ N_k^{\text{old}} + \gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk}) \right\} \\
&= \pi_k^{\text{old}} + \frac{\gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk})}{N}
\end{aligned}$$

Here we have used the conclusion (the update formula for N_k^{new}) in the previous problem. Next we deal with the covariance matrix $\boldsymbol{\Sigma}$. By analogy to

the previous problem, we can obtain:

$$
\begin{aligned}
\boldsymbol{\Sigma}_k^{new} &= \frac{1}{N_k^{\text{new}}} \sum_{n \neq m} \gamma^{\text{old}}(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}})(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}})^T \\
&\quad + \frac{1}{N_k^{\text{new}}} \gamma^{\text{new}}(z_{mk})(\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{new}})(\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{new}})^T \\
&\approx \frac{1}{N_k^{\text{new}}} \sum_{n \neq m} \gamma^{\text{old}}(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{old}})(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{old}})^T \\
&\quad + \frac{1}{N_k^{\text{new}}} \gamma^{\text{new}}(z_{mk})(\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{new}})(\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{new}})^T \\
&= \frac{1}{N_k^{\text{new}}} \sum_{n} \gamma^{\text{old}}(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{old}})(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{old}})^T \\
&\quad + \frac{1}{N_k^{\text{new}}} \gamma^{\text{new}}(z_{mk})(\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{new}})(\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{new}})^T \\
&\quad - \frac{1}{N_k^{\text{new}}} \gamma^{\text{old}}(z_{mk})(\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{old}})(\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{old}})^T \\
&= \frac{1}{N_k^{\text{new}}} N_k^{\text{old}} \boldsymbol{\Sigma}_k^{\text{old}} + \frac{1}{N_k^{\text{new}}} \gamma^{\text{new}}(z_{mk})(\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{new}})(\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{new}})^T \\
&\quad - \frac{1}{N_k^{\text{new}}} \gamma^{\text{old}}(z_{mk})(\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{old}})(\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{old}})^T \\
&= \left\{1 + \frac{N_k^{\text{old}} - N_k^{\text{new}}}{N_k^{\text{new}}}\right\} \boldsymbol{\Sigma}_k^{\text{old}} \\
&\quad + \frac{1}{N_k^{\text{new}}} \gamma^{\text{new}}(z_{mk})(\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{new}})(\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{new}})^T \\
&\quad - \frac{1}{N_k^{\text{new}}} \gamma^{\text{old}}(z_{mk})(\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{old}})(\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{old}})^T \\
&= \left\{1 + \frac{\gamma^{\text{old}}(z_{mk}) - \gamma^{\text{new}}(z_{mk})}{N_k^{\text{new}}}\right\} \boldsymbol{\Sigma}_k^{\text{old}} \\
&\quad + \frac{\gamma^{\text{new}}(z_{mk})}{N_k^{\text{new}}} (\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{new}})(\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{new}})^T \\
&\quad - \frac{\gamma^{\text{old}}(z_{mk})}{N_k^{\text{new}}} (\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{old}})(\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{old}})^T \\
&= \boldsymbol{\Sigma}_k^{\text{old}} \\
&\quad + \frac{\gamma^{\text{new}}(z_{mk})}{N_k^{\text{new}}} \left\{(\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{new}})(\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{new}})^T - \boldsymbol{\Sigma}^{\text{old}}\right\} \\
&\quad - \frac{\gamma^{\text{old}}(z_{mk})}{N_k^{\text{new}}} \left\{(\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{old}})(\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{old}})^T - \boldsymbol{\Sigma}_k^{\text{old}}\right\}
\end{aligned}
$$

One important thing worthy mentioned is that in the second step, there is an approximate equal sign. Note that in the previous problem, we have

shown that if we only recompute the data point $\mathbf{x}_m$, all the center $\boldsymbol{\mu}_k$ will also change from $\boldsymbol{\mu}_k^{\text{old}}$ to $\boldsymbol{\mu}_k^{\text{new}}$, and the update formula is given by Eq (9.78). However, for the convenience of computing, we have made an approximation here. Other approximation methods can also be applied here. For instance, you can replace $\boldsymbol{\mu}_k^{\text{new}}$ with $\boldsymbol{\mu}_k^{\text{old}}$ whenever it occurs.

The complete solution should be given by substituting Eq (9.78) into the right side of the first equal sign and then rearranging it, in order to construct a relation between $\boldsymbol{\Sigma}_k^{\text{new}}$ and $\boldsymbol{\Sigma}_k^{\text{old}}$. However, this is too complicated.

0.10 Variational Inference

Problem 10.1 Solution

This problem is very similar to Prob.9.24. We substitute Eq (10.3) and Eq (10.4) into Eq (10.2):

$$\begin{aligned} L(q)+\mathrm{KL}(q||p) &= \int_{\mathbf{Z}} q(\mathbf{Z})\Big\{\ln\frac{p(\mathbf{X},\mathbf{Z})}{q(\mathbf{Z})}-\ln\frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})}\Big\}d\mathbf{Z} \\ &= \int_{\mathbf{Z}} q(\mathbf{Z})\Big\{\ln\frac{p(\mathbf{X},\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X})}\Big\}d\mathbf{Z} \\ &= \int_{\mathbf{Z}} q(\mathbf{Z})\ln p(\mathbf{X})d\mathbf{Z} \\ &= \ln p(\mathbf{X}) \end{aligned}$$

Note that in the last step, we have used the fact that $\ln p(\mathbf{X})$ doesn't depend on $\mathbf{Z}$, and that the integration of $q(\mathbf{Z})$ over $\mathbf{Z}$ equal to 1 because $q(\mathbf{Z})$ is a PDF.

Problem 10.2 Solution

To be more clear, we are required to solve:

$$\begin{cases} m_1 = \mu_1 - \Lambda_{11}^{-1}\Lambda_{12}(m_2-\mu_2) \\ m_2 = \mu_2 - \Lambda_{22}^{-1}\Lambda_{21}(m_1-\mu_1) \end{cases}$$

To obtain the equation above, we need to substitute $\mathbb{E}[z_i] = m_i$, where $i = 1, 2$, into Eq (10.13) and Eq (10.14). Here the unknown parameters are m_1 and m_2. It is trivial to notice that $m_i = \mu_i$ is a solution for the equation above.

Let's solve this equation from another perspective. Firstly, if any (or both) of Λ_{11}^{-1} and Λ_{22}^{-1} equals 0, we can obtain $m_i = \mu_i$ directly from Eq (10.13)-(10.14). When none of Λ_{11}^{-1} and Λ_{22}^{-1} equals 0, we substitute m_1, i.e., the first

line, into the second line:

$$\begin{aligned} m_2 &= \mu_2 - \Lambda_{22}^{-1}\Lambda_{21}(m_1 - \mu_1) \\ &= \mu_2 - \Lambda_{22}^{-1}\Lambda_{21}\left[\mu_1 - \Lambda_{11}^{-1}\Lambda_{12}(m_2 - \mu_2) - \mu_1\right] \\ &= \mu_2 - \Lambda_{22}^{-1}\Lambda_{21}\mu_1 + \Lambda_{22}^{-1}\Lambda_{21}\Lambda_{11}^{-1}\Lambda_{12}(m_2 - \mu_2) + \Lambda_{22}^{-1}\Lambda_{21}\mu_1 \\ &= (1 - \Lambda_{22}^{-1}\Lambda_{21}\Lambda_{11}^{-1}\Lambda_{12})\,\mu_2 + \Lambda_{22}^{-1}\Lambda_{21}\Lambda_{11}^{-1}\Lambda_{12}\,m_2 \end{aligned}$$

We rearrange the expression above, yielding:

$$(1 - \Lambda_{22}^{-1}\Lambda_{21}\Lambda_{11}^{-1}\Lambda_{12})(m_2 - \mu_2) = 0$$

The first term at the left hand side will equal 0 only when the distribution is singular, i.e., the determinant of the precision matrix $\mathbf{\Lambda}$ (i.e., $\Lambda_{11}\Lambda_{22} - \Lambda_{12}\Lambda_{21}$) is 0. Therefore, if the distribution is nonsingular, we must have $m_2 = \mu_2$. Substituting it back into the first line, we obtain $m_1 = \mu_1$.

Problem 10.3 Solution Let's start from the definition of KL divergence given in Eq (10.16).

$$\begin{aligned} KL(p||q) &= -\int p(\mathbf{Z})\Big[\sum_{i=1}^{M}\ln q_i(\mathbf{Z}_i)\Big]\,d\mathbf{Z} + \text{const} \\ &= -\int p(\mathbf{Z})\Big[\ln q_j(\mathbf{Z}_j) + \sum_{i\neq j}\ln q_i(\mathbf{Z}_i)\Big]\,d\mathbf{Z} + \text{const} \\ &= -\int p(\mathbf{Z})\ln q_j(\mathbf{Z}_j)\,d\mathbf{Z} + \text{const} \\ &= -\int\Big[\int p(\mathbf{Z})\prod_{i\neq j}d\mathbf{Z}_i\Big]\ln q_j(\mathbf{Z}_j)\,d\mathbf{Z}_j + \text{const} \\ &= -\int P(\mathbf{Z}_j)\ln q_j(\mathbf{Z}_j)\,d\mathbf{Z}_j + \text{const} \end{aligned}$$

Note that in the third step, since all the factors $q_i(\mathbf{Z}_i)$, where $i \neq j$, are fixed, they can be absorbed into the 'Const' variable. In the last step, we have denoted the marginal distribution:

$$p(\mathbf{Z}_j) = \int p(\mathbf{Z})\prod_{i\neq j}d\mathbf{Z}_i$$

We introduce the Lagrange multiplier to enforce $q_j(\mathbf{Z}_j)$ integrate to 1.

$$L = -\int P(\mathbf{Z}_j)\ln q_j(\mathbf{Z}_j)\,d\mathbf{Z}_j + \lambda\Big(\int q_j(\mathbf{Z}_j)\,d\mathbf{Z}_j - 1\Big)$$

Using the functional derivative (for more details, you can refer to Appendix D or Prob.1.34), we calculate the functional derivative of L with respect to $q_j(\mathbf{Z}_j)$ and set it to 0:

$$-\frac{p(\mathbf{Z}_j)}{q_j(\mathbf{Z}_j)} + \lambda = 0$$

Rearranging it, we can obtain:

$$\lambda q_j(\mathbf{Z}_j) = p(\mathbf{Z}_j)$$

Integrating both sides with respect to $\mathbf{Z}_j$, we see that $\lambda = 1$. Substituting it back into the derivative, we can obtain the optimal $q_j(\mathbf{Z}_j)$:

$$q_j^\star(\mathbf{Z}_j) = p(\mathbf{Z}_j)$$

Notice that actually we should also enforce $q_j(\mathbf{Z}_j) > 0$ in the Lagrange multiplier, however as we can see that when we only enforce $q_j(\mathbf{Z}_j)$ integrate to 1 and obtain the final close expression, $q_j(\mathbf{Z}_j)$ is definitely larger than 0 at all $\mathbf{Z}_j$ because $p(\mathbf{Z}_j)$ is a PDF. Therefore, there is no need to introduce this inequality constraint in the Lagrange multiplier.

Problem 10.4 Solution

We begin by writing down the KL divergence.

$$\begin{aligned}
\mathrm{KL}(p||q) &= -\int p(\mathbf{x}) \ln\left\{\frac{q(\mathbf{x})}{p(\mathbf{x})}\right\} d\mathbf{x} \\
&= -\int p(\mathbf{x}) \ln q(\mathbf{x}) d\mathbf{x} + \text{const} \\
&= -\int p(\mathbf{x}) \Big[-\frac{D}{2} \ln 2\pi - \frac{1}{2} \ln |\boldsymbol{\Sigma}| - \frac{1}{2} (\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu}) \Big] d\mathbf{x} + \text{const} \\
&= \int p(\mathbf{x}) \Big[\frac{1}{2} \ln |\boldsymbol{\Sigma}| + \frac{1}{2} (\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu}) \Big] d\mathbf{x} + \text{const} \\
&= \frac{1}{2} \ln |\boldsymbol{\Sigma}| + \int p(\mathbf{x}) \Big[\frac{1}{2} (\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu}) \Big] d\mathbf{x} + \text{const} \\
&= \frac{1}{2} \ln |\boldsymbol{\Sigma}| + \int p(\mathbf{x}) \frac{1}{2} \Big[\mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x} - 2\boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \mathbf{x} + \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \Big] d\mathbf{x} + \text{const} \\
&= \frac{1}{2} \ln |\boldsymbol{\Sigma}| + \frac{1}{2} \int p(\mathbf{x}) \mathrm{Tr}[\boldsymbol{\Sigma}^{-1} (\mathbf{x}\mathbf{x}^T)] \, d\mathbf{x} - \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \mathbb{E}[\mathbf{x}] + \frac{1}{2} \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} + \text{const} \\
&= \frac{1}{2} \ln |\boldsymbol{\Sigma}| + \frac{1}{2} \mathrm{Tr}[\boldsymbol{\Sigma}^{-1} \mathbb{E}(\mathbf{x}\mathbf{x}^T)] - \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \mathbb{E}[\mathbf{x}] + \frac{1}{2} \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} + \text{const}
\end{aligned}$$

Here D is the dimension of $\mathbf{x}$. We first calculate the derivative of $\mathrm{KL}(p||q)$ with respect to $\boldsymbol{\mu}$ and set it to 0:

$$\frac{\partial \mathrm{KL}}{\partial \boldsymbol{\mu}} = -\boldsymbol{\Sigma}^{-1} \mathbb{E}[x] + \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} = 0$$

Therefore, we can obtain $\boldsymbol{\mu} = \mathbb{E}[\mathbf{x}]$. When $\boldsymbol{\mu} = \mathbb{E}[\mathbf{x}]$ is satisfied, KL divergence reduces to:

$$\mathrm{KL}(p||q) = \frac{1}{2} \ln |\boldsymbol{\Sigma}| + \frac{1}{2} \mathrm{Tr}[\boldsymbol{\Sigma}^{-1} \mathbb{E}(\mathbf{x}\mathbf{x}^T)] - \frac{1}{2} \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} + \text{const}$$

Then we calculate the derivative of $\mathrm{KL}(p||q)$ with respect to $\boldsymbol{\Sigma}$ and set it to 0:

$$\frac{\partial \mathrm{KL}}{\partial \boldsymbol{\Sigma}} = \frac{1}{2}\boldsymbol{\Sigma}^{-1} - \frac{1}{2}\boldsymbol{\Sigma}^{-1}\mathbb{E}[\mathbf{x}\mathbf{x}^T]\boldsymbol{\Sigma}^{-1} + \frac{1}{2}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1} = 0$$

Note that here we have used and Eq (61) and Eq (124) in 'MatrixCook-Book', and that $\boldsymbol{\Sigma}$, $\mathbb{E}[\mathbf{x}\mathbf{x}^T]$ are both symmetric. We rewrite those equations here for your reference:

$$\frac{\partial \mathbf{a}^T\mathbf{X}^{-1}\mathbf{b}}{\partial \mathbf{X}} = -\mathbf{X}^{-T}\mathbf{a}\mathbf{b}^T\mathbf{X}^{-T} \quad \text{and} \quad \frac{\partial \mathrm{Tr}(\mathbf{A}\mathbf{X}^{-1}\mathbf{B})}{\partial \mathbf{X}} = -\mathbf{X}^{-T}\mathbf{A}^T\mathbf{B}^T\mathbf{X}^{-T}$$

Rearranging the derivative, we can obtain:

$$\boldsymbol{\Sigma} = \mathbb{E}[\mathbf{x}\mathbf{x}^T] - \boldsymbol{\mu}\boldsymbol{\mu}^T = \mathbb{E}[\mathbf{x}\mathbf{x}^T] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}]^T = \mathrm{cov}[\mathbf{x}]$$

Problem 10.5 Solution

We introduce a property of Dirac function:

$$\int \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_0) f(\boldsymbol{\theta})\, d\boldsymbol{\theta} = f(\boldsymbol{\theta}_0)$$

We first calculate the optimal $q(\mathbf{z}, \boldsymbol{\theta})$ by fixing $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$. This is achieved by minimizing the KL divergence given in Eq (10.4):

$$\begin{aligned}
\mathrm{KL}(q||p) &= -\int\int q(\mathbf{Z}) \ln\left\{\frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})}\right\} d\mathbf{Z} \\
&= -\int\int q_{\mathbf{z}}(\mathbf{z}) q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \ln\left\{\frac{p(\mathbf{z}, \boldsymbol{\theta}|\mathbf{X})}{q_{\mathbf{z}}(\mathbf{z}) q_{\boldsymbol{\theta}}(\boldsymbol{\theta})}\right\} d\mathbf{z}\, d\boldsymbol{\theta} \\
&= -\int\int q_{\mathbf{z}}(\mathbf{z}) q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \ln\left\{\frac{p(\mathbf{z}, \boldsymbol{\theta}|\mathbf{X})}{q_{\mathbf{z}}(\mathbf{z})}\right\} d\mathbf{z}\, d\boldsymbol{\theta} + \int q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \ln q_{\boldsymbol{\theta}}(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\
&= -\int\int q_{\mathbf{z}}(\mathbf{z}) q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \ln\left\{\frac{p(\mathbf{z}, \boldsymbol{\theta}|\mathbf{X})}{q_{\mathbf{z}}(\mathbf{z})}\right\} d\mathbf{z}\, d\boldsymbol{\theta} + \mathrm{const} \\
&= -\int q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \left\{\int q_{\mathbf{z}}(\mathbf{z}) \ln\left\{\frac{p(\mathbf{z}, \boldsymbol{\theta}|\mathbf{X})}{q_{\mathbf{z}}(\mathbf{z})}\right\} d\mathbf{z}\right\} d\boldsymbol{\theta} + \mathrm{const} \\
&= -\int q_{\mathbf{z}}(\mathbf{z}) \ln\left\{\frac{p(\mathbf{z}, \boldsymbol{\theta}_0|\mathbf{X})}{q_{\mathbf{z}}(\mathbf{z})}\right\} d\mathbf{z} + \mathrm{const} \\
&= -\int q_{\mathbf{z}}(\mathbf{z}) \ln\left\{\frac{p(\mathbf{z}|\boldsymbol{\theta}_0, \mathbf{X}) p(\boldsymbol{\theta}_0|\mathbf{X})}{q_{\mathbf{z}}(\mathbf{z})}\right\} d\mathbf{z} + \mathrm{const} \\
&= -\int q_{\mathbf{z}}(\mathbf{z}) \ln\left\{\frac{p(\mathbf{z}|\boldsymbol{\theta}_0, \mathbf{X})}{q_{\mathbf{z}}(\mathbf{z})}\right\} d\mathbf{z} + \mathrm{const}
\end{aligned}$$

Here the 'Const' denotes the terms independent of $q_{\mathbf{z}}(\mathbf{z})$. Note that we will show at the end of this problem, here 'Const' actually is $-\infty$ due to the existence of the entropy of Dirac function:

$$\int q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \ln q_{\boldsymbol{\theta}}(\boldsymbol{\theta})\, d\boldsymbol{\theta}$$

Now it is clear that when $q_{\mathbf{z}}(\mathbf{z})$ equals $p(\mathbf{z}|\boldsymbol{\theta}_0,\mathbf{X})$, the KL divergence is minimized. This corresponds to the E-step. Next, we calculate the optimal $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$, i.e., $\boldsymbol{\theta}_0$, by maximizing $L(q)$ given in Eq (10.3), but fixing $q_{\theta}(\boldsymbol{\theta})$:

$$\begin{aligned}
L(q) &= \int\int q(\mathbf{Z})\ln\Big\{\frac{p(\mathbf{X},\mathbf{Z})}{q(\mathbf{Z})}\Big\}d\mathbf{Z} \\
&= \int\int q_{\mathbf{z}}(\mathbf{z})q_{\boldsymbol{\theta}}(\boldsymbol{\theta})\ln\Big\{\frac{p(\mathbf{X},\mathbf{z},\boldsymbol{\theta})}{q_{\mathbf{z}}(\mathbf{z})q_{\boldsymbol{\theta}}(\boldsymbol{\theta})}\Big\}d\mathbf{z}\,d\boldsymbol{\theta} \\
&= \int\int q_{\mathbf{z}}(\mathbf{z})q_{\boldsymbol{\theta}}(\boldsymbol{\theta})\ln\Big\{\frac{p(\mathbf{X},\mathbf{z},\boldsymbol{\theta})}{q_{\mathbf{z}}(\mathbf{z})}\Big\}d\mathbf{z}\,d\boldsymbol{\theta} - \int q_{\boldsymbol{\theta}}(\boldsymbol{\theta})\ln q_{\boldsymbol{\theta}}(\boldsymbol{\theta})d\boldsymbol{\theta} \\
&= \int\int q_{\mathbf{z}}(\mathbf{z})q_{\boldsymbol{\theta}}(\boldsymbol{\theta})\ln\Big\{p(\mathbf{X},\mathbf{z},\boldsymbol{\theta})\Big\}d\mathbf{z}\,d\boldsymbol{\theta} - \int q_{\boldsymbol{\theta}}(\boldsymbol{\theta})\ln q_{\boldsymbol{\theta}}(\boldsymbol{\theta})d\boldsymbol{\theta} + \text{const} \\
&= \int q_{\boldsymbol{\theta}}(\boldsymbol{\theta})\,\mathbb{E}_{q_{\mathbf{z}}}[\ln p(\mathbf{X},\mathbf{z},\boldsymbol{\theta})]\,d\boldsymbol{\theta} - \int q_{\boldsymbol{\theta}}(\boldsymbol{\theta})\ln q_{\boldsymbol{\theta}}(\boldsymbol{\theta})d\boldsymbol{\theta} + \text{const} \\
&= \mathbb{E}_{q_{\mathbf{z}}(\mathbf{z})}[\ln p(\mathbf{X},\mathbf{z},\boldsymbol{\theta}_0)] - \int q_{\boldsymbol{\theta}}(\boldsymbol{\theta})\ln q_{\boldsymbol{\theta}}(\boldsymbol{\theta})d\boldsymbol{\theta} + \text{const}
\end{aligned}$$

The second term is actually the entropy of a Dirac function, which is $-\infty$ and independent of the value of $\boldsymbol{\theta}_0$. Not strictly speaking, we only need to maximize the first term. This is exactly the M-step.

One important thing needs to be clarified here. You may ask no matter how we set $\boldsymbol{\theta}_0$, $L(q)$ will always be $-\infty$. Actually, this is an intrinsic problem as long as we use a point estimate $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$. This will even occur when we derive the optimal $q_{\mathbf{z}}(\mathbf{z})$ by minimizing the KL divergence at the first step. Therefore, the 'Maximizing' and 'Minimizing' is a general meaning in this problem where we neglect the $-\infty$ term.

Problem 10.6 Solution

Let's use the hint by first enforcing $\alpha \to 1$.

$$\begin{aligned}
D_\alpha(p||q) &= \frac{4}{1-\alpha^2}\Big(1-\int p^{(1+\alpha)/2}q^{(1-\alpha)/2}\,dx\Big) \\
&= \frac{4}{1-\alpha^2}\Big\{1-\int \frac{p}{p^{(1-\alpha)/2}}\,\big[1+\frac{1-\alpha}{2}\ln q+O(\frac{1-\alpha}{2})^2\,\big]\,dx\Big\} \\
&= \frac{4}{1-\alpha^2}\Big\{1-\int p\cdot\frac{1+\frac{1-\alpha}{2}\ln q+O(\frac{1-\alpha}{2})^2}{1+\frac{1-\alpha}{2}\ln p+O(\frac{1-\alpha}{2})^2}\,dx\Big\} \\
&\approx \frac{4}{1-\alpha^2}\Big\{1-\int p\cdot\frac{1+\frac{1-\alpha}{2}\ln q}{1+\frac{1-\alpha}{2}\ln p}\,dx\Big\} \\
&= \frac{4}{1-\alpha^2}\Big\{-\int p\cdot\big[\frac{1+\frac{1-\alpha}{2}\ln q}{1+\frac{1-\alpha}{2}\ln p}-1\big]\,dx\Big\} \\
&= \frac{4}{1-\alpha^2}\Big\{-\int p\cdot\frac{\frac{1-\alpha}{2}\ln q-\frac{1-\alpha}{2}\ln p}{1+\frac{1-\alpha}{2}\ln p}\,dx\Big\} \\
&= \frac{2}{1+\alpha}\Big\{-\int p\cdot\frac{\ln q-\ln p}{1+\frac{1-\alpha}{2}\ln p}\,dx\Big\} \\
&\approx -\int p\cdot(\ln q-\ln p)\,dx=-\int p\cdot\ln\frac{q}{p}\,dx
\end{aligned}$$

Here p and q is short for $p(x)$ and $q(x)$. It is similar when $\alpha \to -1$. One important thing worthy mentioning is that if we directly approximate $p^{(1+\alpha)/2}$ by p instead of $p/p^{(1-\alpha)/2}$ in the first step, we won't get the desired result.

Problem 10.7 Solution