Derivations of Equations for Pattern Recognition and Machine Learning

by

Soong Lee

August, 2024

This document is a collection of derivations of non-trivial equations and statements from PRML (Feb 2006). I did not include the equations that were assigned as exercises, since the solutions of them are available from the resources in the internet.

1. PRML Solutions to Exercises: Tutor's Edition
2. PRML Solutions to Exercises: Web Edition
3. Solution Manual for PRML by Zhengqi Gao

I used the same mathematical notation as in PRML except for $\bar{\mathbf{t}}$, which is a column vector of a list of observations in this document.

1. t: a binary-categorical target value.
2. $\mathbf{t}$: a vector for multi-categorical target values. $\{t_1, t_2, \cdots, t_M\}$, where M is the dimension of the feature space.
3. $\bar{\mathbf{t}}$: a vector for a list of observations of binary-categorical target values. $\{t_1, t_2, \cdots, t_N\}$, where N is the number of observations.

Reference:

The Matrix Cookbook (Nov 2012) by K. B. Peterson and M. S. Pederson (https://www2.imm.dtu.dk/pubdb/pubs/3274-full.html)

Chapter 1. Introduction

Eq 1.65: (PRML p.30)

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}\right\}$$

Proof :

From Eq (1.52),

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

Here, D = M + 1, $\mu = 0, \mathbf{\Sigma} = \alpha^{-1}I$

$\mathbf{w}$ is $(w_0, w_1, ..., w_M)$ vector.

$\Rightarrow$ M + 1 elements encluding 0th order term.

$$|\alpha^{-1}\mathbf{I}| = det\begin{bmatrix} \alpha^{-1} & & & \\ & \alpha^{-1} & & \\ & & \ddots & \\ & & & \alpha^{-1} \end{bmatrix} = (\alpha^{-1})^{M+1}$$

$$\begin{aligned} \therefore\ \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) &= \frac{1}{(2\pi)^{(M+1)/2}}\ \alpha^{(M+1)/2} \exp\left\{-\frac{1}{2}\mathbf{w}^T \cdot (\alpha^{-1})^{-1}\ \mathbf{w}\right\} \\ &= \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2}\ \exp\left\{-\frac{\alpha}{2}\mathbf{w}^T \cdot \mathbf{w}\right\} \end{aligned}$$

Eq 1.66: (PRML p.30)

$$p(\mathbf{w}|\mathbf{X}, \bar{\mathbf{t}}, \alpha, \beta)\ \propto\ p(\bar{\mathbf{t}}|\mathbf{X}, \mathbf{w}, \beta)\ p(\mathbf{w}|\alpha)$$

Proof :

Let's omit $\mathbf{X}$ and β for brevity.

We know that

$$p(R|E) = \frac{P(R \cap E)}{p(E)} \tag{1}$$

$$p(\mathbf{w}|\bar{\mathbf{t}}, \alpha) \;=\; \frac{p(\bar{\mathbf{t}}|\mathbf{w}, \alpha)\; p(\mathbf{w}, \alpha)}{p(\bar{\mathbf{t}}, \alpha)}$$

Using Eq (1),

$$p(\mathbf{w}, \alpha) = p(\alpha)p(\mathbf{w}|\alpha)$$

$$\begin{aligned} p(\bar{\mathbf{t}}, \alpha) &= p(\alpha)p(\bar{\mathbf{t}}|\alpha) \\ &= \frac{p(\bar{\mathbf{t}}|\mathbf{w}, \alpha)p(\alpha)p(\mathbf{w}|\alpha)}{p(\alpha)p(\bar{\mathbf{t}}|\alpha)} \\ &= \frac{p(\bar{\mathbf{t}}|\mathbf{w}, \alpha)p(\mathbf{w}|\alpha)}{p(\bar{\mathbf{t}}|\alpha)} \end{aligned}$$

From Eq (1.60), $p(\bar{\mathbf{t}}|\text{paramters})$ does not depend on α.

$$\Rightarrow \qquad \frac{p(\bar{\mathbf{t}}|\mathbf{w})p(\mathbf{w}|\alpha)}{p(\bar{\mathbf{t}})}$$

Eq 1.68: (PRML p.31)

$$p(t|x, D) = \int p(t|x, \mathbf{w})p(\mathbf{w}|D)d\mathbf{w}$$

Proof :

$(\mathbf{X}, \bar{\mathbf{t}})$: test data set

D $= [(x_1, t_1), (x_2, t_2), ..., (x_N, t_N)]$: training set

$$\begin{aligned} p(t|x, D) &= \frac{1}{p(x, D)}p(t, x, D) \\ &= \frac{1}{p(x, D)}\int p(t, x, D, \mathbf{w})d\mathbf{w} \qquad (\textit{sum rule}) \\ &= \frac{1}{p(x, D)}\int p(t|x, D, \mathbf{w})p(x, D, \mathbf{w})d\mathbf{w} \qquad (\textit{product rule}) \\ &= \frac{1}{p(x, D)}\int p(t|x, D, \mathbf{w})[p(\mathbf{w}|x, D)p(x, D)]d\mathbf{w} \\ &= \int p(t|x, D, \mathbf{w})p(\mathbf{w}|x, D)d\mathbf{w} \end{aligned}$$

Because $\mathbf{w}$ is determined by D, $p(t|x, D, \mathbf{w}) = p(t|x, \mathbf{w})$

and because $\mathbf{w} \perp x$, $p(\mathbf{w}|x, D) = p(\mathbf{w}|D)$.

$$\therefore \quad p(t|x, D) = \int p(t|x, \mathbf{w}) p(\mathbf{w}|D) d\mathbf{w}$$

Eq 1.80: (PRML p.41)

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

Proof :

	Positive$_{j=1}$	Negative$_{j=0}$
True$_{k=1}$	L_{11}	L_{10}
False$_{k=0}$	L_{01}	L_{00}

$$\int_{R_1} L_{11}\, p(\mathbf{x}, C_1) d\mathbf{x}$$
$$\int_{R_0} L_{10}\, p(\mathbf{x}, C_1) d\mathbf{x}$$
$$\int_{R_1} L_{01}\, p(\mathbf{x}, C_2) d\mathbf{x}$$
$$\int_{R_0} L_{00}\, p(\mathbf{x}, C_2) d\mathbf{x}$$

Sum of all these $= \mathbb{E}(L)$

Eq 1.88: (PRML p.46)

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

Proof :

Eq (1.88) is the result of a few steps beforehand.

To minimize $\mathbb{E}[L]$ in

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

we need to think of a functional E with y, $\mathrm{y_x}$, and x.

$$\begin{aligned} E(y, y_x, \mathbf{x}) &= \int_x \left[\int_t \{y(\mathbf{x}) - t\}^2 p(\mathbf{x}, t) dt \right] d\mathbf{x} \\ f(y, y_x, \mathbf{x}) &= \int_t \{y(\mathbf{x}) - t\}^2 p(\mathbf{x}, t) dt \end{aligned}$$

The Euler equation is

$$\frac{\partial f}{\partial y} - \frac{d}{dx} \frac{\partial f}{\partial y_x} = 0$$

Applying this Euler equation to the equation above,

$$\begin{aligned} \frac{\partial f}{\partial y} &= 2 \int_t \{y(\mathbf{x}) - t\} p(\mathbf{x}, t) dt \\ \frac{\partial f}{\partial y_x} &= 0 \quad (no\ y_x\ term) \end{aligned}$$

$$\therefore \quad 2 \int \{y(\mathbf{x}) - t\} p(\mathbf{x}, t) dt = 0$$

Eq 1.90: (PRML p.47)

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

Proof :

$$\begin{aligned} &\int \{y(\mathbf{x}) - \mathbb{E}_t[t|\mathbf{x}]\} \cdot \{\mathbb{E}_t[t|\mathbf{x}] - t\} \cdot p(\mathbf{x}, t) d\mathbf{t} \\ &= \{y(\mathbf{x}) - \mathbb{E}_t[t|\mathbf{x}]\} \cdot \left\{ \int \mathbb{E}_t[t|\mathbf{x}] p(\mathbf{x}, t) d\mathbf{t} - \int t\, p(\mathbf{x}, t) d\mathbf{t} \right\} \end{aligned}$$

where

$$\mathbb{E}_t[t|\mathbf{x}] = \int t\, p(\mathbf{x}, t) d\mathbf{t} \tag{1.37}$$

And $p(\mathbf{x}, t) = p(t|\mathbf{x}) \cdot p(\mathbf{x})$

$$\begin{aligned} \int \mathbb{E}_t[t|\mathbf{x}] \cdot p(\mathbf{x}, t) dt &= \mathbb{E}_t[t|\mathbf{x}] \int p(\mathbf{x}, t) dt \\ &= \mathbb{E}_t[t|\mathbf{x}] \cdot p(\mathbf{x}) \end{aligned} \tag{1}$$

$$\begin{aligned} \int t \cdot p(\mathbf{x}, t) dt &= \int t \cdot p(t|\mathbf{x}) \cdot p(\mathbf{x}) dt \\ &= p(\mathbf{x}) \int t \cdot p(t|\mathbf{x}) dt \\ &= p(\mathbf{x}) \cdot \mathbb{E}_t[t|\mathbf{x}] \end{aligned} \tag{2}$$

Eqs (1) and (2) are the same.

$$\therefore \quad \int \{y(\mathbf{x}) - \mathbb{E}_t[t|\mathbf{x}]\} \cdot \{\mathbb{E}_t[t|\mathbf{x}] - t\} \cdot p(\mathbf{x}, t) d\mathbf{t} = 0$$

Eq 1.97: (PRML p.51)

$$H = -\lim_{N \to \infty} \sum_i \left(\frac{n_i}{N}\right) \ln \left(\frac{n_i}{N}\right) = -\sum_i p_i \ln p_i$$

Proof :

Using Stirling's appx, (as $N \to \infty, n_i \to \infty$)

$$\ln N! \simeq N \ln N - N$$

$$\ln n! \simeq n_i \ln n_i - n_i$$

Eq (1.96) becomes

$$
\begin{aligned}
H &= \frac{1}{N}\ln N! - \frac{1}{N}\sum_i \ln n_i! \\
&\simeq \frac{1}{N}(N\ln N - N) - \frac{1}{N}\sum_i (n_i \ln n_i - n_i) \\
&= \ln N - 1 - \frac{1}{N}\sum_i n_i \ln n_i + \frac{1}{N}\sum_i n_i \\
&= \ln N - \frac{1}{N}\sum_i n_i \ln n_i \\
&= \left(\sum_i \frac{n_i}{N}\right) \cdot \ln N - \frac{1}{N}\sum_i n_i \ln n_i \\
&= -\sum_i \left(\frac{n_i}{N}\right) \cdot (\ln n_i - \ln N) \\
&= -\sum_i \left(\frac{n_i}{N}\right) \cdot \left(\ln \frac{n_i}{N}\right)
\end{aligned}
$$

PRML p.52:

"The corresponding value of the entropy is then H = lnM. "

Proof :

$$
\widetilde{\mathrm{H}} = -\sum_i p(x_i)\ln p(x_i) + \lambda\left(\sum_i p(x_i) - 1\right) \tag{1.99}
$$

Constraint is $\sum_i p(x_i) - 1 = 0$

The conditions to maximize $\widetilde{\mathrm{H}}$ will be,

$$
\begin{aligned}
\frac{\partial \widetilde{\mathrm{H}}}{\partial p(x_1)} &= -\ln p(x_1) - 1 + \lambda = 0 \\
\frac{\partial \widetilde{\mathrm{H}}}{\partial p(x_2)} &= -\ln p(x_2) - 1 + \lambda = 0 \\
&\vdots \\
\frac{\partial \widetilde{\mathrm{H}}}{\partial p(x_M)} &= -\ln p(x_M) - 1 + \lambda = 0
\end{aligned}
$$

$\Rightarrow \quad p(x_1) = p(x_2) = \ldots = p(x_M)$

Therefore, $p(x_i) = \frac{1}{M}$ to make H maximum.

$$H_{max} = -\sum_{i=1}^{M} \frac{1}{M} \ln\left(\frac{1}{M}\right)$$
$$= \ln M$$

Eq 1.108: (PRML p.54)

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

Proof :

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

Since $\lambda_1, \lambda_2, \lambda_3, \mu$, and σ^2 are given, to maximize J we need to maximize the integral.

$$\begin{aligned} K &= \int_{-\infty}^{\infty} \left[-p(x) \ln p(x) + \lambda_1 p(x) + \lambda_2 x p(x) + \lambda_3 (x-\mu)^2 p(x)\right] dx \\ &= \int_{-\infty}^{\infty} f(y, y_x, x)\, dx \end{aligned}$$

$f(y, y_x, x) = -p(x) \ln p(x) + \lambda_1 p(x) + \lambda_2 x p(x) + \lambda_3 (x-\mu)^2 p(x)$

Here y = p(x) and there is no y_x terms.

From the Euler equation

$$\frac{df}{dy} - \frac{d}{dx}\frac{\partial f}{\partial y_x} = 0$$

$$\frac{\partial}{\partial p}\left[-p \ln p + \lambda_1 p + \lambda_2 x p + \lambda_3 (x-\mu)^2 p\right] = -\ln p - 1 + \lambda_1 + \lambda_2 x + \lambda_3 (x-\mu)^2 = 0$$

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

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

Eq 1.112: (PRML p.55)

$$\mathrm{H}[\mathbf{x}, \mathbf{y}] = \mathrm{H}[\mathbf{y}|\mathbf{x}] + \mathrm{H}[\mathbf{x}]$$

Proof :

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

First term $= \mathrm{H}[\mathbf{y}|\mathbf{x}]$

$$\begin{aligned}
\text{Second term} &= -\iint p(\mathbf{y}, \mathbf{x}) \ln p(\mathbf{x}) d\mathbf{y} d\mathbf{x} \\
&= -\int_x \ln p(\mathbf{x}) \left[\int_y \ln p(\mathbf{y}, \mathbf{x}) d\mathbf{y} \right] d\mathbf{x} \\
&= -\iint p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x} \\
&= \mathrm{H}[\mathbf{x}]
\end{aligned}$$

$$\therefore \quad \mathrm{H}[\mathbf{x}, \mathbf{y}] = \mathrm{H}[\mathbf{y}|\mathbf{x}] + \mathrm{H}[\mathbf{x}]$$

Eq 1.118: (PRML p.56)

$$\mathrm{KL}(p||q) = -\int p(\mathbf{x}) \ln \left\{ \frac{q(\mathbf{x})}{p(\mathbf{y})} \right\} d\mathbf{x} \geq -\ln \int q(\mathbf{x}) d\mathbf{x} = 0$$

Proof : (*Ref: https://en.wikipedia.org/wiki/Jensen%27s_inequality*)

Let's define another random variable Y(X),

$$Y(X) = \frac{q(\mathbf{x})}{p(\mathbf{x})}$$

$$f(Y) = -\ln(y) \quad \leftarrow \quad f(y) \; is \; a \; convex \; function$$

The important point in this Y random variable is that its probability p(y) is still p(x), since Y is a function of X.

$$
\begin{aligned}
\mathbb{E}_x[f(y)] &\geq f(\mathbb{E}_x[y]) \\
\int p(x)f(y)dx &\geq f\left(\int p(x)ydx\right) \\
-\int p(x)\ln\frac{q(x)}{p(x)}dx &\geq -\ln\left(p(x)\frac{q(x)}{p(x)}dx\right) \\
&= -\ln\int q(x)dx \\
&= 0
\end{aligned}
$$

Eq 1.121: (PRML p.57)

$$
\mathrm{I}[\mathbf{x},\mathbf{y}] = \mathrm{H}[\mathbf{x}] - \mathrm{H}[\mathbf{x}|\mathbf{y}] = \mathrm{H}[\mathbf{y}] - \mathrm{H}[\mathbf{y}|\mathbf{x}]
$$

Proof : We know that

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

$$
\begin{aligned}
\text{First term 1} &= -\iint p(\mathbf{y},\mathbf{x})\,\ln p(\mathbf{x})d\mathbf{x}d\mathbf{y} \\
&= -\int \ln p(\mathbf{x})\left[\int \ln p(\mathbf{y},\mathbf{x})d\mathbf{y}\right]d\mathbf{x} \\
&= -\iint p(\mathbf{x})\ln p(\mathbf{x})d\mathbf{x} \\
&= +\mathrm{H}[\mathbf{x}] \\
\text{First term 2} &= -\int \ln p(\mathbf{y})\left[\int \ln p(\mathbf{x},\mathbf{y})d\mathbf{x}\right]d\mathbf{y} \\
&= +\mathrm{H}[\mathbf{y}] \\
\text{Second term} &= \iint p(\mathbf{x},\mathbf{y})\,\ln p(\mathbf{y}|\mathbf{x})d\mathbf{x}d\mathbf{y} + \iint p(\mathbf{x},\mathbf{y})\,\ln p(\mathbf{x})d\mathbf{x}d\mathbf{y} \\
&= -\mathrm{H}[\mathbf{y}|\mathbf{x}] - \mathrm{H}[\mathbf{x}]
\end{aligned}
$$

$$\begin{aligned}\therefore\quad \mathrm{I}[\mathbf{x},\mathbf{y}] &= \mathrm{H}[\mathbf{x}] + \mathrm{H}[\mathbf{y}] - \mathrm{H}[\mathbf{y}|\mathbf{x}] - \mathrm{H}[\mathbf{x}] \\ &= \mathrm{H}[\mathbf{y}] - \mathrm{H}[\mathbf{y}|\mathbf{x}]\end{aligned}$$

Chapter 2. Probability Distributions

Eq 2.19: (PRML p.73)

$$p(x=1|D) = \int_0^1 p(x=1|\mu)p(\mu|D)d\mu$$

Proof :

$$\begin{aligned}p(x|D) &= \frac{p(x,D)}{p(D)} \\ &= \frac{\int p(x,D,\mu)d\mu}{p(D)} \qquad \leftarrow \quad \textit{sum rule} \\ &= \frac{\int p(x|D,\mu)p(D,\mu)d\mu}{p(D)} \qquad \leftarrow \quad \textit{product rule} \\ &= \int p(x|D,\mu)p(\mu|D)d\mu\end{aligned}$$

where the integrand $p(x|D,\mu)$ must be a shorthand notation for $p(x|\mu)$.

Eq 2.20: (PRML p.73)

$$p(x=1|D) = \frac{m+a}{m+a+l+b}$$

Proof :

$$p(x=1|D) = \int_0^1 \mu p(\mu|D)d\mu$$

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

$$p(x=1|D) = \int_0^1 \mu\,\mathrm{Beta}(\mu|(m+a),(l+b))d\mu$$

Since

$$\int_0^1 \mu\,\mathrm{Beta}(\mu|a,b)d\mu = \frac{a}{a+b}$$

$$\therefore \;\; p(x=1|D) = \frac{m+a}{(m+a)+(l+b)} = \frac{m+a}{m+a+l+b}$$

Eq 2.23: (PRML p.74)

$$\mathbb{E}_D[E_\theta\boldsymbol{\theta}|D] \equiv \int \left\{ \int \boldsymbol{\theta} p(\boldsymbol{\theta}|D)d\boldsymbol{\theta} \right\} p(D)dD$$

Proof :

$$\begin{aligned} \int \left\{ \int \boldsymbol{\theta} p(\boldsymbol{\theta}|D)d\boldsymbol{\theta} \right\} p(D)dD &= \int \left\{ \int \boldsymbol{\theta} p(\boldsymbol{\theta}|D)p(D)dD \right\} d\boldsymbol{\theta} \\ (\text{Since } \int \boldsymbol{\theta} p(\boldsymbol{\theta}|D)p(D)dD = \boldsymbol{\theta} p(\boldsymbol{\theta})) & \\ &= \int \boldsymbol{\theta} p(\boldsymbol{\theta})d\boldsymbol{\theta} \end{aligned}$$

$$\therefore \;\; \mathbb{E}_\theta[\boldsymbol{\theta}] = \mathbb{E}_D[\mathbb{E}_\theta[\boldsymbol{\theta}|D]]$$

Eq 2.24: (PRML p.74)

$$\mathrm{var}_\theta[\boldsymbol{\theta}] = \mathbb{E}_D[\mathrm{var}_\theta[\boldsymbol{\theta}|D]] + \mathrm{var}_D[\mathbb{E}_\theta[\boldsymbol{\theta}|D]]$$

Proof :

1st term,

$$\mathbb{E}_D[\mathrm{var}_\theta[\boldsymbol{\theta}|D]] = \mathbb{E}_D[\mathbb{E}_\theta[(\boldsymbol{\theta} - \mathbb{E}_\theta[\boldsymbol{\theta}|D])^2|D]]$$

Let's calculate the inside therm on the right hand side equation above,

$$\begin{aligned} \mathbb{E}_\theta[(\boldsymbol{\theta} - \mathbb{E}_\theta[\boldsymbol{\theta}|D])^2|D] &= \mathbb{E}_\theta[\{\boldsymbol{\theta}^2 - 2\boldsymbol{\theta}\mathbb{E}_\theta[\boldsymbol{\theta}|D] + (\mathbb{E}_\theta[\boldsymbol{\theta}|D])^2\}|D] \\ &= \mathbb{E}_\theta[\boldsymbol{\theta}^2|D] - 2\mathbb{E}_\theta[\boldsymbol{\theta}|D]\mathbb{E}_\theta[\boldsymbol{\theta}|D] + (\mathbb{E}_\theta[\boldsymbol{\theta}|D])^2\mathbb{E}_\theta[1|D] \\ (\text{Since } \mathbb{E}_\theta[1|D] = \int 1\, p(\boldsymbol{\theta}|D)d\boldsymbol{\theta} = 1) & \\ &= \mathbb{E}_\theta[\boldsymbol{\theta}^2|D] - (\mathbb{E}_\theta[\boldsymbol{\theta}|D])^2 \qquad (1) \end{aligned}$$

2nd term,

$$\begin{aligned}
\mathrm{var}_D[\mathbb{E}_\theta[\boldsymbol{\theta}|D]] &= \mathbb{E}_D\left[(\mathbb{E}_\theta[\boldsymbol{\theta}|D] - \mathbb{E}_D[\mathbb{E}_\theta[\boldsymbol{\theta}|D]])^2\right] \\
&= \mathbb{E}_D[(\mathbb{E}_\theta[\theta|D])^2 - 2\mathbb{E}_\theta[\boldsymbol{\theta}|D]\cdot\mathbb{E}_D[\mathbb{E}_\theta[\boldsymbol{\theta}|D]] + (\mathbb{E}_D[\mathbb{E}_\theta[\boldsymbol{\theta}|D]])^2] \\
&= \mathbb{E}[(\mathbb{E}_\theta[\boldsymbol{\theta}|D])^2] - \mathbb{E}_D(2\mathbb{E}_\theta[\boldsymbol{\theta}|D])\cdot[\mathbb{E}_D[\mathbb{E}_\theta[\boldsymbol{\theta}|D]]) + (\mathbb{E}_D[\mathbb{E}_\theta[\boldsymbol{\theta}|D]])^2\mathbb{E}_D[1] \\
&\text{(Since } \ \mathbb{E}_D[1] = 1) \\
&= \mathbb{E}_D[(\mathbb{E}_\theta[\boldsymbol{\theta}|D])^2] - (\mathbb{E}_D[\mathbb{E}_\theta[\boldsymbol{\theta}|D]])^2 \qquad (2)
\end{aligned}$$

Putting them together,

$$\begin{aligned}
\mathbb{E}_D[\mathrm{var}_\theta[\boldsymbol{\theta}|D]] + \mathrm{var}_D[\mathbb{E}_\theta[\boldsymbol{\theta}|D]] &= \mathbb{E}_D[\mathbb{E}_\theta[\boldsymbol{\theta}^2|D] - (\mathbb{E}_\theta[\boldsymbol{\theta}|D])^2 + \mathbb{E}_D[(\mathbb{E}_\theta[\boldsymbol{\theta}|D])^2] \\
&\quad - (\mathbb{E}_D[\mathbb{E}_\theta[\boldsymbol{\theta}|D]])^2 \\
&= \mathbb{E}_D[\mathbb{E}_\theta[\boldsymbol{\theta}^2|D]] - (\mathbb{E}_D[\mathbb{E}_\theta[\boldsymbol{\theta}|D]])^2
\end{aligned}$$

1st term,

$$\begin{aligned}
\mathbb{E}_D[\mathbb{E}_\theta[\boldsymbol{\theta}^2|D]] &= \int_D \mathbb{E}_\theta[\boldsymbol{\theta}^2|D]\cdot p(D)dD \\
&= \int_D \left\{\int_\theta \boldsymbol{\theta}^2 p(\boldsymbol{\theta}|D)d\boldsymbol{\theta}\right\} p(D)dD \\
&= \int_\theta \boldsymbol{\theta}^2 \left\{\int_D p(\boldsymbol{\theta}|D)p(D)dD\right\} d\boldsymbol{\theta} \\
&(\textit{the inner integral becomes } p(\boldsymbol{\theta})) \\
&= \int \boldsymbol{\theta}^2 p(\boldsymbol{\theta})d\boldsymbol{\theta} \\
&= \mathbb{E}_\theta[\boldsymbol{\theta}^2]
\end{aligned}$$

2nd term,

$$\begin{aligned}
\mathbb{E}_D[\mathbb{E}_\theta[\boldsymbol{\theta}|D]] &= \int_D \mathbb{E}_\theta[\boldsymbol{\theta}|D] \cdot p(D)dD \\
&= \int_D \left\{ \int_\theta p(\boldsymbol{\theta}|D)d\boldsymbol{\theta} \right\} p(D)dD \\
&= \int_\theta \boldsymbol{\theta} \left\{ \int_D p(\boldsymbol{\theta}|D) \cdot p(D)dD \right\} d\boldsymbol{\theta} \\
&= \int \boldsymbol{\theta} p(\boldsymbol{\theta}) d\boldsymbol{\theta} \\
&= \mathbb{E}_\theta[\boldsymbol{\theta}]
\end{aligned}$$

$$\begin{aligned}
\therefore \quad \mathbb{E}_D[\mathrm{var}_\theta[\boldsymbol{\theta}|D]] + \mathrm{var}_D[\mathbb{E}_\theta[\boldsymbol{\theta}|D]] &= \mathbb{E}_\theta[\boldsymbol{\theta}^2] - (\mathbb{E}_\theta[\boldsymbol{\theta}])^2 \\
&= \mathrm{var}_\theta[\boldsymbol{\theta}]
\end{aligned}$$

Eq 2.56: (PRML p.82)

$$p(\mathbf{y}) = p(\mathbf{x})|\mathbf{J}| = \prod_{j=1}^{D} \frac{1}{(2\pi\lambda_j)^{1/2}} \exp\left\{ -\frac{y_j^2}{2\lambda_j} \right\}$$

Proof :

From Eqs (2.43), (2.44), and (2.50),

$$\begin{aligned}
p(\mathbf{y}) &= \frac{1}{(2\pi)^{D/2}} \ \frac{1}{\prod_{j=1}^{D} \lambda_j^{1/2}} \ \exp\left\{ -\frac{1}{2} \sum_{i=1}^{D} \frac{y_i^2}{\lambda_i} \right\} \\
&= \left[\prod_{j=1}^{D} \frac{1}{(2\pi\lambda_j)^{1/2}} \right] \cdot \left[\prod_{i=1}^{D} \exp\left\{ -\frac{y_j^2}{2\lambda_j} \right\} \right] \\
&= \prod_{j=1}^{D} \frac{1}{(2\pi\lambda_j)^{1/2}} \ \exp\left\{ -\frac{y_j^2}{2\lambda_j} \right\}
\end{aligned}$$

Eq 2.60: (PRML p.83)

$$\mathbf{z} = \sum_{j=1}^{D} y_j \mathbf{u}_j$$

Proof :

Let $\mathbf{z} = \sum_{j=1}^{D} c_j \mathbf{u}_j$

Multiplying $\mathbf{u}_k^T$ from the left,

$$\mathbf{u}_k^T \mathbf{z} = \mathbf{u}_k^T \sum_{j=1}^{D} c_j \mathbf{u}_j = \sum_{j=1}^{D} c_j \mathbf{u}_k^T \mathbf{u}_j = \sum_{j=1}^{D} c_j \ \mathrm{I}_{kj} = c_k$$

$\Rightarrow \qquad c_k = \mathbf{u}_k^T \mathbf{z}$

From Eq (2.51), c_k is actually y_k.

$$y_i = \mathbf{u}_i^T (\mathbf{x} - \mathbf{u}) \tag{2.51}$$

$$\therefore \quad \mathbf{z} = \sum_{j=1}^{D} y_j \mathbf{u}_j \quad (\text{where} \ \ y_j = \mathbf{u}_j^T \mathbf{z})$$

Eq :2.61 (PRML p.83)

$$\frac{1}{(2\pi)^{D/2}} \frac{1}{|\mathbf{\Sigma}|^{1/2}} \int \exp\left\{-\frac{1}{2}\mathbf{z}^T \Sigma^{-1} z\right\} \mathbf{z}\mathbf{z}^T d\mathbf{z}$$
$$= \frac{1}{(2\pi)^{D/2}} \frac{1}{|\mathbf{\Sigma}|^{1/2}} \sum_{i=1}^{D} \sum_{j=1}^{D} \mathbf{u}_j \mathbf{u}_j^T \int \exp\left\{-\sum_{k=1}^{D} \frac{y_k^2}{2\lambda_k}\right\} y_i y_j d\mathbf{y}$$
$$= \sum_{i=1}^{D} \mathbf{u}_i \mathbf{u}_i^T \lambda_i = \mathbf{\Sigma}$$

Proof :

Using $\mathbf{z} = \sum_{j=1}^{D} y_j \mathbf{u}_j$ and $\mathbf{\Sigma}^{-1} = \sum_{i=1}^{D} \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^T$

$$\mathbf{z}^T \mathbf{\Sigma}^{-1} \mathbf{z} = \left(\sum_{j=1}^{D} y_j \mathbf{u}_j^T\right) \left(\sum_{i=1}^{D} \frac{1}{\lambda_i} \mathbf{u}_i T \mathbf{u}_i^T\right) \left(\sum_{k=1}^{D} y_k \mathbf{u}_k\right)$$
$$= \sum_{i,j,k} y_j \frac{1}{\lambda_i} y_k \cdot \mathbf{u}_j^T \mathbf{u}_i \mathbf{u}_i^T \mathbf{u}_k$$

$$(\text{Since} \ \ \mathbf{u}_j^T \mathbf{u}_i = I_{ji} \ \ \text{and} \ \ \mathbf{u}_i^T \mathbf{u}_k = I_{ik})$$

$$= \sum_{i=1}^{D} \frac{y_i^2}{2\lambda_i}$$

$\mathbf{z} \cdot \mathbf{z}^T = \sum_{i,j} \mathbf{u}_i \mathbf{u}_j^T y_i y_j = \sum_{i=1}^D y_i^2 \mathbf{u}_i \mathbf{u}_i^T$

$d\mathbf{z} = dz_1 \cdot dz_2 \cdots dz_D$

Since $y_i = \mathbf{u_i}^T \mathbf{z}$, and z runs $-\infty \to \infty$,

$d\mathbf{z} = d\mathbf{y} = dy_1 dy_2 dy_3 \cdots dy_D$

(This is not clear to me, but I can buy that, since $\mathbf{u}_i$ is a normalized vector; $\mathbf{u}_i \mathbf{u}_i^T = 1$)

Putting these together,

$$\begin{aligned} \int \exp\left\{-\frac{1}{2}\mathbf{z}^T \mathbf{\Sigma}^{-1} \mathbf{z}\right\} \mathbf{z}\mathbf{z}^T d\mathbf{z} &= \sum_{i=1}^{D} \int \exp\left\{-\sum_{k=1}^{D} \frac{y_k^2}{2\lambda_k}\right\} y_i^2 \mathbf{u}_i \mathbf{u}_i^T d\mathbf{y} \\ &= \sum_{i=1}^{D} \int \prod_{k=1}^{D} \exp\left(-\frac{y_k^2}{2\lambda_k}\right) y_i^2 \mathbf{u}_i \mathbf{u}_i^T d\mathbf{y} \qquad (1) \end{aligned}$$

We know that,

$$\int_{-\infty}^{\infty} \mathrm{e}^{-y^2/2\lambda} dy = \sqrt{2\pi\lambda}$$

$$\int_{-\infty}^{\infty} \mathrm{e}^{-y^2/2\lambda}\, y^2 dy = \lambda \sqrt{2\pi\lambda}$$

For i = 1, eq (1) becomes

$$\lambda_1 \sqrt{2\pi\lambda_1} (2\pi)^{(D-1)/2} [\lambda_2 \cdots \lambda_D]^{1/2} \mathbf{u}_1 \mathbf{u}_1^T = \lambda_1 (2\pi)^{D/2} \prod_{i=1}^{D} \lambda_i^{1/2}\ \mathbf{u}_1 \mathbf{u}_1^T$$

For i = 2, eq (1) becomes

$$\lambda_2 (2\pi)^{D/2} \prod_{i=1}^{D} \lambda_i^{1/2}\ \mathbf{u}_2 \mathbf{u}_2^T$$

$$\vdots$$

Summing all up,

$$\sum_{i=1}^{D} \int \prod_{k=1}^{D} \exp\left(-\frac{y_k^2}{2\lambda_k}\right) y_i^2 d\mathbf{y} \mathbf{u}_i \mathbf{u}_i^T = (2\pi)^{D/2} \prod_{i=1}^{D} \lambda_i^{1/2} \sum_{j=1}^{D} \lambda_j \mathbf{u}_j \mathbf{u}_j^T$$

Therefore,

$$
\begin{aligned}
&\frac{1}{(2\pi)^{D/2}} \frac{1}{|\mathbf{\Sigma}|^{1/2}} \int \exp\left\{-\frac{1}{2}\mathbf{z}^T\Sigma^{-1}\mathbf{z}\right\}\mathbf{z}\mathbf{z}^T d\mathbf{z} \\
&= \frac{1}{(2\pi)^{D/2}} \cdot \frac{1}{\prod_{j=1}^{D}\lambda_j^{1/2}} \cdot (2\pi)^{D/2} \cdot \prod_{i=1}^{D}\lambda_i^{1/2} \sum_{j=1}^{D}\lambda_j \mathbf{u}_j\mathbf{u}_j^T \\
&= \sum_{j=1}^{D}\lambda_j \mathbf{u}_j\mathbf{u}_j^T \\
&= \mathbf{\Sigma}
\end{aligned}
$$

Eq 2.84: (PRML p.88)

$$
-\frac{1}{2}\mathbf{x}_b^T\mathbf{\Lambda}_{bb}\mathbf{b} + \mathbf{x}_b^T\mathbf{m} = -\frac{1}{2}(\mathbf{x}_b - \mathbf{\Lambda}_{bb}^{-1}\mathbf{m})^T\ \mathbf{\Lambda}_{bb}\ (\mathbf{x}_b - \mathbf{\Lambda}_{bb}^{-1}\mathbf{m}) + \frac{1}{2}\mathbf{m}^T\mathbf{\Lambda}_{bb}^{-1}\mathbf{m}
$$

Proof :

Let's prove it backward,

$$
\begin{aligned}
&-\frac{1}{2}(\mathbf{x}_b - \mathbf{\Lambda}_{bb}^{-1}\mathbf{m})^T\ \mathbf{\Lambda}_{bb}\ (\mathbf{x}_b - \mathbf{\Lambda}_{bb}^{-1}\mathbf{m}) + \frac{1}{2}\mathbf{m}^T\mathbf{\Lambda}_{bb}^{-1}\mathbf{m} \\
&= -\frac{1}{2}[(\mathbf{x}_b^T - \mathbf{m}^T\mathbf{\Lambda}_{bb}^{-1})(\mathbf{\Lambda}_{bb}\mathbf{x}_b - \mathbf{m})] + \frac{1}{2}\mathbf{m}^T\mathbf{\Lambda}_{bb}^{-1}\mathbf{m} \\
&= -\frac{1}{2}[\mathbf{x}_b^T\mathbf{\Lambda}_{bb}\mathbf{x}_b - \mathbf{x}_b^T\mathbf{m} - \mathbf{m}^T\mathbf{x}_b + \mathbf{m}^T\mathbf{\Lambda}_{bb}^{-1}\mathbf{m}] + \frac{1}{2}\mathbf{m}^T\mathbf{\Lambda}_{bb}^{-1}\mathbf{m} \\
&= -\frac{1}{2}\mathbf{x}_b^T\mathbf{\Lambda}_{bb}\mathbf{x}_b + \mathbf{x}_b^T\mathbf{m}
\end{aligned}
$$

Eq 2.87: (PRML p.89)

$$
\begin{aligned}
&\frac{1}{2}[\mathbf{\Lambda}_{bb}\boldsymbol{\mu}_b - \mathbf{\Lambda}_{ba}(\mathbf{x}_a - \boldsymbol{\mu}_a)]^T\mathbf{\Lambda}_{bb}^{-1}[\mathbf{\Lambda}_{bb}\boldsymbol{\mu}_b - \mathbf{\Lambda}_{ba}(\mathbf{x}_a - \boldsymbol{\mu}_a)] \\
&- \frac{1}{2}\mathbf{x}_a^T\mathbf{\Lambda}_{aa}\mathbf{x}_a + x_a^T(\mathbf{\Lambda}_{aa}\boldsymbol{\mu}_a + \mathbf{\Lambda}_{ab}\boldsymbol{\mu}_b) + const \\
&= -\frac{1}{2}\mathbf{x}_a^T(\mathbf{\Lambda}_{aa} - \mathbf{\Lambda}_{ab}\mathbf{\Lambda}_{bb}^{-1}\mathbf{\Lambda}_{ba})\mathbf{x}_a \\
&+ \mathbf{x}_a^T(\mathbf{\Lambda}_{aa} - \mathbf{\Lambda}_{ab}\mathbf{\Lambda}_{bb}^{-1}\mathbf{\Lambda}_{ba})\boldsymbol{\mu}_a + const
\end{aligned}
$$

Proof :

$$
\begin{aligned}
\text{1st term} &= \frac{1}{2}[\boldsymbol{\mu}_b^T\boldsymbol{\Lambda}_{bb} - (\mathbf{x}_a^T - \boldsymbol{\mu}_a^T)\boldsymbol{\Lambda}_{ab}] \cdot [\boldsymbol{\mu}_b - \Lambda_{bb}^{-1}\boldsymbol{\Lambda}_{ba}(\mathbf{x}_a - \boldsymbol{\mu}_a)] \\
&= \frac{1}{2}[\boldsymbol{\mu}_b^T\boldsymbol{\Lambda}_{bb}\boldsymbol{\mu}_b - \boldsymbol{\mu}_b^T\boldsymbol{\Lambda}_{ba}(\mathbf{x}_a - \boldsymbol{\mu}_a) - (\mathbf{x}_a^T - \boldsymbol{\mu}_a^T)\boldsymbol{\Lambda}_{ab}\boldsymbol{\mu}_b + (\mathbf{x}_a^T - \boldsymbol{\mu}_a^T)\boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}(\mathbf{x}_a - \boldsymbol{\mu}_a)] \\
&= \frac{1}{2}[\boldsymbol{\mu}_b^T\boldsymbol{\Lambda}_{bb}\boldsymbol{\mu}_b + \boldsymbol{\mu}_a^T\boldsymbol{\Lambda}_{ab}\boldsymbol{\mu}_b + \boldsymbol{\mu}_a^T\boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}\boldsymbol{\mu}_a + \boldsymbol{\mu}_b^T\Lambda_{ba}\boldsymbol{\mu}_a] \\
&\quad + \frac{1}{2}[-\boldsymbol{\mu}_b^T\boldsymbol{\Lambda}_{ba}\mathbf{x}_a - \mathbf{x}_a^T\boldsymbol{\Lambda}_{ab}\boldsymbol{\mu}_b - \mathbf{x}_a^T\boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}\boldsymbol{\mu}_a - \boldsymbol{\mu}_a^T\boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}\mathbf{x}_a] \\
&\quad - \frac{1}{2}\mathbf{x}_a^T\boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}\mathbf{x}_a
\end{aligned}
$$

(The first term is constant. And using $\boldsymbol{\mu}_b^T\boldsymbol{\Lambda}_{ba}\mathbf{x}_a = \mathbf{x}_a^T\boldsymbol{\Lambda}_{ab}\boldsymbol{\mu}_b,$)

$$
\begin{aligned}
&= -\mathbf{x}_a^T\boldsymbol{\Lambda}_{ab}\boldsymbol{\mu}_b - \mathbf{x}_a^T\boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}\boldsymbol{\mu}_a - \boldsymbol{\mu}_a^T\boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}\mathbf{x}_a \\
&\quad - \frac{1}{2}\mathbf{x}_a^T\boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}\mathbf{x}_a + const
\end{aligned}
$$

The third term above becomes,

$$
\begin{aligned}
\boldsymbol{\mu}_a^T[\boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}]\mathbf{x}_a &= \mathbf{x}_a^T[\boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}]^T\boldsymbol{\mu}_a \\
&= \mathbf{x}_a^T[\boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}]\boldsymbol{\mu}_a
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\text{1st term } &= -\mathbf{x}_a^T\boldsymbol{\Lambda}_{ab}\boldsymbol{\mu}_b - \mathbf{x}_a^T\boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}\boldsymbol{\mu}_a \\
&\quad - \frac{1}{2}\mathbf{x}_a^T\boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}\mathbf{x}_a + const
\end{aligned}
$$

Combining the first, second, and the third terms in LHS of eq 2.87,

$$
\begin{aligned}
&[-\mathbf{x}_a^T\boldsymbol{\Lambda}_{ab}\boldsymbol{\mu}_b - \mathbf{x}_a^T\boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}\boldsymbol{\mu}_a - \frac{1}{2}\mathbf{x}_a^T\boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}\mathbf{x}_a + const] \\
&\quad - \frac{1}{2}\mathbf{x}_a^T\boldsymbol{\Lambda}_{aa}\mathbf{x}_a + \mathbf{x}_a^T(\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a + \boldsymbol{\Lambda}_{ab}\boldsymbol{\mu}_b) + const \\
&\quad = -\frac{1}{2}\mathbf{x}_a^T(\boldsymbol{\Lambda}_{aa} - \boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba})\mathbf{x}_a + \mathbf{x}_a^T(\boldsymbol{\Lambda}_{aa} - \boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba})\boldsymbol{\mu}_a + const
\end{aligned}
$$

Eq 2.111: (PRML p.92)

$$
\mathbb{E}[\mathbf{x}|\mathbf{y}] = (\boldsymbol{\Lambda} + \mathbf{A}^T\mathbf{L}\mathbf{A})^{-1}\left\{\mathbf{A}^T\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\right\}
$$

Proof :

$$\mathbf{R} = \begin{bmatrix} \mathbf{\Lambda} + \mathbf{A}^T\mathbf{L}\mathbf{A} & -\mathbf{A}^T\mathbf{L} \\ -\mathbf{L}\mathbf{A} & \mathbf{L} \end{bmatrix}$$

where $\mathbf{R}$ is precision $(= \mathbf{\Sigma}^{-1})$.

Compared to eqs (2.73) and (2.75),

$$\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} \Longleftrightarrow \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}$$

$$\mathbf{x}_a \Longleftrightarrow \mathbf{x}$$

$$\mathbf{x}_b \Longleftrightarrow \mathbf{y}$$

$$\mathbf{\Sigma}_{a|b} \Longleftrightarrow \mathbf{\Sigma}_{x|y} = \mathbf{R}_{xx}^{-1}(\mathbf{\Lambda} + \mathbf{A}^T\mathbf{L}\mathbf{A})^{-1}$$

$$\mathbf{\Lambda}_{aa} \Longleftrightarrow \mathbf{R}_{xx} = [\mathbf{\Lambda} + \mathbf{A}^T\mathbf{L}\mathbf{A}]$$

$$\mathbf{\Lambda}_{ab} \Longleftrightarrow \mathbf{R}_{xy} - \mathbf{A}^T\mathbf{L}$$

$$\boldsymbol{\mu}_b \Longleftrightarrow \boldsymbol{\mu}_y = \mathbf{A}\boldsymbol{\mu} + b$$

$$\begin{aligned}
\mathbb{E}[\mathbf{x}|\mathbf{y}] &= \mathbf{\Sigma}_{x|y}\{\mathbf{R}_{xx}\boldsymbol{\mu}_x - \mathbf{R}_{xy}(\mathbf{y} - \boldsymbol{\mu}_y)\} \\
&= (\mathbf{\Lambda} + \mathbf{A}^T\mathbf{L}\mathbf{A})^{-1}\{(\mathbf{\Lambda} + \mathbf{A}^T\mathbf{L}\mathbf{A})\mu - (-\mathbf{A}^T\mathbf{L})(\mathbf{y} - \mathbf{A}\boldsymbol{\mu} - \mathbf{b})\} \\
&= (\mathbf{\Lambda} + \mathbf{A}^T\mathbf{L}\mathbf{A})^{-1}\{\mathbf{A}^T\mathbf{L}(\mathbf{y} - \mathbf{b}) + \mathbf{\Lambda}\boldsymbol{\mu}\}
\end{aligned}$$

Eq 2.118: (PRML p.93)

$$\ln p(\mathbf{X}|\boldsymbol{\mu}, \mathbf{\Sigma}) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\mathbf{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu})$$

Proof :

Since $\mathbf{x}_n$ is drawn independently,

$$
\begin{aligned}
p(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma}) &= p(\mathbf{x}_1|\boldsymbol{\mu},\boldsymbol{\Sigma})\cdot p(\mathbf{x}_2|\boldsymbol{\mu},\boldsymbol{\Sigma})\cdots p(\mathbf{x}_N|\boldsymbol{\mu},\boldsymbol{\Sigma}) \\
&= \prod_{n=1}^{N} p(\mathbf{x}_n|\boldsymbol{\mu},\Sigma) \\
&= \prod \frac{1}{(2\pi)^{D/2}}\cdot\frac{1}{(\boldsymbol{\Sigma})^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}_n-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu})\right\} \\
&= \frac{1}{(2\pi)^{ND/2}}\cdot\frac{1}{(\boldsymbol{\Sigma})^{N/2}}\exp\left\{-\frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu})\right\}
\end{aligned}
$$

Eq 2.120: (PRML p.93)

$$
\frac{\partial}{\partial\boldsymbol{\mu}}\ln p(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sum_{n=1}^{N}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu})
$$

Proof :

$$
\begin{aligned}
\frac{\partial}{\partial\boldsymbol{\mu}}\ln p(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma}) &= -\frac{1}{2}\sum_{n=1}^{N}\frac{\partial}{\partial\boldsymbol{\mu}}\left\{(\mathbf{x}_n^T-\boldsymbol{\mu}^T)\boldsymbol{\Sigma}^{-1}\cdot(\mathbf{x}_n-\boldsymbol{\mu})\right\} \\
&= -\frac{1}{2}\sum_{n=1}^{N}\frac{\partial}{\partial\boldsymbol{\mu}}\left\{\mathbf{x}_n^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_n-\mathbf{x}_n^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_n+\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right\}
\end{aligned}
$$

First term:

$$
\frac{\partial}{\partial\boldsymbol{\mu}}\mathbf{x}_n^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_n = 0
$$

Second and third terms:

Using eq (C.19),

$$
\frac{\partial}{\partial\boldsymbol{\mu}}(\mathbf{x}_n^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}) = \frac{\partial}{\partial\boldsymbol{\mu}}\left\{(\boldsymbol{\Sigma}^{-1}\mathbf{x}_n)^T\boldsymbol{\mu}\right\}
$$

(Above, used $(\boldsymbol{\Sigma}^{-1})^T = \boldsymbol{\Sigma}^{-1}$)

$$
\begin{aligned}
&= \frac{\partial}{\partial\boldsymbol{\mu}}\left\{\boldsymbol{\mu}^T(\boldsymbol{\Sigma}^{-1}\mathbf{x}_n)\right\} \\
&= (\boldsymbol{\Sigma}^{-1}\mathbf{x}_n)^T \\
&= \mathbf{x}_n^T\boldsymbol{\Sigma}^{-1} \\
&= \boldsymbol{\Sigma}^{-1}\mathbf{x}_n
\end{aligned}
$$

Fourth term,

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\mu}}(\boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}) &= \frac{\partial}{\partial \boldsymbol{\mu}} \left\{ (\boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}) \boldsymbol{\mu}) \right\} \\ &= (\boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}) + \left\{ \frac{\partial}{\partial \boldsymbol{\mu}} (\boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}) \right\} \mu \\ &= 2\, \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}\end{aligned}$$

Putting all together,

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\mu}} \ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= -\frac{1}{2} \sum_{n=1}^{N} \left\{ -2\boldsymbol{\Sigma}^{-1} \mathbf{x}_n + 2\boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \right\} \\ &= \sum_{n=1}^{N} \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu})\end{aligned}$$

Eqs 2.123 & 2.124: (PRML p.94)

$$\mathbb{E}[\boldsymbol{\mu}_{ML}] = \boldsymbol{\mu}$$

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

Proof :

$$\begin{aligned}\mathbb{E}[\boldsymbol{\mu}_{ML}] &= \mathbb{E}\left[\frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n \right] \\ &= \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}[\mathbf{x}_n] \\ &= \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{\mu} \\ &= \boldsymbol{\mu}\end{aligned}$$

We are going to use Eq. (2.291),

$$\begin{aligned}\mathbb{E}[\mathbf{x}_n \mathbf{x}_m^T] &= \boldsymbol{\mu}\boldsymbol{\mu}^T + I_{mn} \boldsymbol{\Sigma} \\ \mathbb{E}[\boldsymbol{\Sigma}_{ML}] &= \mathbb{E}\left[\frac{1}{N} \sum_{n=1}^{N} (\mathbf{x}_n - \boldsymbol{\mu}_{ML})(\mathbf{x}_n - \boldsymbol{\mu}_{ML})^T \right] \\ &= \mathbb{E}\left[\frac{1}{N} \sum_{n=1}^{N} (\mathbf{x}_n \mathbf{x}_n^T - \mathbf{x}_n \boldsymbol{\mu}_{ML} - \boldsymbol{\mu}_{ML} \mathbf{x}_n + \boldsymbol{\mu}_{ML} \boldsymbol{\mu}_{ML}^T \right] \\ &= \frac{1}{N} \sum_{n=1}^{N} \left\{ \mathbb{E}[\mathbf{x}\mathbf{x}^T] - \mathbb{E}[\mathbf{x}\boldsymbol{\mu}_{ML}^T] - \mathbb{E}[\boldsymbol{\mu}_{ML} \mathbf{x}^T] + \mathbb{E}[\boldsymbol{\mu}_{ML} \boldsymbol{\mu}_{ML}^T] \right\}\end{aligned}$$

1st term:

$$\mathbb{E}[\mathbf{x}\mathbf{x}^T] = \boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\Sigma}$$

2nd term:

$$\begin{aligned}
\mathbb{E}[\mathbf{x}_m \boldsymbol{\mu}_{ML}^T] &= \mathbb{E}\left[\mathbf{x}_m \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n^T\right] \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[\mathbf{x}_m\mathbf{x}_n^T] \\
&= \frac{1}{N}\sum_{n=1}^{N}[\boldsymbol{\mu}\boldsymbol{\mu}^T + I_{mn}\boldsymbol{\Sigma}] \\
&= \boldsymbol{\mu}\boldsymbol{\mu}^T + \frac{1}{N}\boldsymbol{\Sigma}
\end{aligned}$$

3rd term:

$$\begin{aligned}
\mathbb{E}[\boldsymbol{\mu}_{ML}\mathbf{x}_m^T &= \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \cdot \mathbf{x}_m^T\right] \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[\mathbf{x}_n \cdot \mathbf{x}_m^T] \\
&= \boldsymbol{\mu}\boldsymbol{\mu}^T + \frac{1}{N}\boldsymbol{\Sigma}
\end{aligned}$$

4th term:

$$\begin{aligned}
\mathbb{E}[\boldsymbol{\mu}_{ML} \cdot \boldsymbol{\mu}_{ML}^T] &= \mathbb{E}\left[\frac{1}{N^2}\sum_{n=1}^{N}\sum_{m=1}^{N}\mathbf{x}_n \cdot \mathbf{x}_m^T\right] \\
&= \boldsymbol{\mu}\boldsymbol{\mu}^T + \frac{1}{N^2}\sum_{m,n} I_{mn}\boldsymbol{\Sigma} \\
&= \boldsymbol{\mu}\boldsymbol{\mu}^T + \frac{1}{N^2}\sum_{n=1}^{N}\boldsymbol{\Sigma} \\
&= \boldsymbol{\mu}\boldsymbol{\mu}^T + \frac{1}{N}\boldsymbol{\Sigma}
\end{aligned}$$

Putting all these together,

$$\begin{aligned}
\mathbb{E}[\boldsymbol{\Sigma}_{ML}] &= (\boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\Sigma}) - 2(\boldsymbol{\mu}\boldsymbol{\mu}^T + \frac{1}{N}\boldsymbol{\Sigma}) + (\boldsymbol{\mu}\boldsymbol{\mu}^T + \frac{1}{N}\boldsymbol{\Sigma}) \\
&= \boldsymbol{\Sigma} - \frac{1}{N}\boldsymbol{\Sigma} \\
&= \frac{N-1}{N}\boldsymbol{\Sigma}
\end{aligned}$$

Eq 2.136: (PRML p.97)

$$z = -\frac{\partial}{\partial \mu_{ML}} \ln[(x|\mu_{ML}, \sigma^2) = -\frac{1}{\sigma^2}(x - \mu_{ML})$$

Proof :

$$p(x|\mu_{ML}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{}{2\sigma^2}(x - \mu_{ML})^2\right\} \qquad \text{(Eq. 2.42)}$$

$$\ln p(x|\mu_{ML}, \sigma^2) = -\frac{1}{2}\ln(2\pi\sigma^2_+ \left\{-\frac{1}{2\sigma^2}(x - \mu_{ML})^2\right\}$$

$$\begin{aligned} \frac{\partial}{\partial \mu_{ML}} \ln p(x|\mu_{ML}, \sigma^2) &= 0 - \frac{1}{2\sigma^2}\ 2(x - \mu_{ML})\ (-1) \\ &= \frac{1}{\sigma^2}(x - \mu_{ML}) \end{aligned}$$

Eq 2.141 & 2.142: (PRML p.98)

$$\begin{aligned} p(\mu|\overline{\mathbf{x}}) &= \mathcal{N}(\mu|\mu_N, \sigma_N^2) \\ \text{where } \mu_N &= \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\ \mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\ \mu_{ML} \\ \frac{1}{\sigma_N^2} &= \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2} \end{aligned}$$

Proof :

$$\begin{aligned} p(\overline{\mathbf{x}}|\mu) &= \prod_{n=1}^{N} p(x_n|\mu) \\ &= \frac{1}{(2\pi\sigma^2)^{N/2}}\ \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2\right\} \end{aligned}$$

$$p(\mu) = \frac{1}{(2\pi\sigma_0^2)^{1/2}}\ \exp\left\{-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right\}$$

$$\begin{aligned} p(\overline{\mathbf{x}}|\mu) \cdot p(\mu) &= \frac{1}{(2\pi\sigma^2)^{N/2}}\ \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2\right\} \frac{1}{(2\pi\sigma_0^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right\} \\ &= \frac{1}{(2\pi\sigma^2)^{N/2}}\ \frac{1}{(2\pi\sigma_0^2)^{1/2}}\ \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2 - \frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right\} \end{aligned}$$

Inside { },

$$\frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2+\frac{1}{\sigma_0^2}(\mu-\mu_0)^2=\left(\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2-2\left(\frac{1}{\sigma^2}\sum_{n=1}^{N}x_n+\frac{\mu_0}{\sigma_0^2}\right)\mu+\cdots$$
$$=A(\mu-B)^2+C$$

$$\begin{aligned}
\text{where } A &= \frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}\\
B &= \frac{\frac{1}{\sigma^2}N\mu_{ML}+\frac{\mu_0}{\sigma_0^2}}{\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}}\\
(\text{where } \mu_{ML} &= \frac{1}{N}\sum_{n=1}^{N}x_n)\\
&= \frac{\sigma_0^2N\mu_{ML}+\sigma^2\mu_0}{\sigma_0^2N+\sigma^2}
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mu_N &= \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\mu_{ML}+\frac{\sigma^2}{N\sigma_0+\sigma^2}\mu_0\\
\frac{1}{\sigma_N^2} &= \frac{N}{\sigma^2}+\frac{1}{\sigma_0^2} \qquad (\Leftarrow A)
\end{aligned}$$

Eq 2.150 & 2.151: (PRML p.100)

$$\begin{aligned}
a_N &= \frac{N}{2}+a_0\\
b_N &= b_0+\frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^2
\end{aligned}$$

Proof :

$$
\begin{aligned}
p(\overline{\mathbf{x}}|\lambda) &= \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \lambda^{-1}) \\
&\propto \lambda^{N/2} \text{exp} \left\{ -\frac{\lambda}{2} \sum_{n=1}^{N} (x_n - \mu)^2 \right\} \\
p(\lambda) = Gam(\lambda|a_0, b_0) &= \frac{1}{\Gamma(a_0)} b_0^{a_0} \lambda^{a_0 - 1} \text{exp}(-b_0\lambda) \\
p(\lambda|\overline{\mathbf{x}} &= p(\overline{\mathbf{x}}|\lambda) \cdot p(\lambda) \\
&\propto \lambda^{N/2} \text{exp} \left\{ -\frac{\lambda}{2} \sum_{n=1}^{N} (x_n - \mu)^2 \right\} \cdot \frac{1}{\Gamma(a_0)} b_0^{a_0} \lambda^{a_0 - 1} \text{exp}(-b_0\lambda) \\
&= \frac{1}{\Gamma(a_0)} \cdot b_0^{a_0} \lambda^{N/2 + a_0 - 1} \cdot \text{exp} \left\{ -\frac{\lambda}{2} \sum_{n=1}^{N} (x_n - \mu)^2 - b_0\lambda \right\}
\end{aligned}
$$

Inside { },

Comparing $-\lambda \left[\frac{1}{2} \sum_{n=1}^{N} (x_n - \mu)^2 + b_0 \right]$ with $-b_0\lambda$,

$\Rightarrow \qquad b_N = \frac{1}{2} \sum_{n=1}^{N} (x_n - \mu)^2 + b_0$

From $\lambda^{N/2 + a_0 - 1} \Longleftrightarrow \lambda^{a_0 - 1}$

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

Eq 2.158: (PRML p.103)

$$
p(x|\mu, a, b) = \frac{b^a}{\Gamma(a)} \left(\frac{1}{2\pi} \right)^{1/2} \left[b + \frac{(x-\mu)^2}{2} \right]^{-a-1/2} \Gamma(a + \frac{1}{2})
$$

Proof :

$$
\int_0^\infty \frac{b^a e^{-b\tau} \tau^{a-1}}{\Gamma(a)} \left(\frac{\tau}{2\pi} \right)^{1/2} \text{exp} \left\{ -\frac{\tau}{2} (x-\mu)^2 \right\} d\tau
$$

(Let $z = [b + \frac{1}{2}(x-\mu)^2]\tau$ and then $dz = [b + \frac{1}{2}(x-\mu)^2]d\tau$)

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

$$
\begin{aligned}
\Gamma(t) &= \int_0^\infty y^{t-1}e^{-y}dy \\
(z \Longleftrightarrow y,\quad & t \Longleftrightarrow a+\frac{1}{2}\ ) \\
\int_0^\infty e^{-z}e^{a-1/2}dz &= \int_0^\infty e^{-z}e^{(a+1/2)-1}dz \\
&= \Gamma(a+\frac{1}{2})
\end{aligned}
$$

$$
\therefore\quad p(x|\mu,a,b) = \frac{b^a}{\Gamma(a)}\left(\frac{1}{2\pi}\right)^{1/2}\left[b+\frac{(x-\mu)^2}{2}\right]^{-a-1/2}\ \Gamma(a+\frac{1}{2})
$$

Eq 2.160: (PRML p.104)

$$
St(x|\mu,\lambda,\nu) = \int_0^\infty \mathcal{N}(x|\mu,(\eta\lambda)^{-1})Gam(\eta|\nu/2,\nu/2)d\eta
$$

Proof :

$\nu = 2a,\ \lambda = \frac{a}{b},\ \eta = \frac{\tau b}{a}$

$\Rightarrow\quad a = \frac{\nu}{2},\ b = \frac{a}{\lambda} = \frac{\nu}{2\lambda},\ \tau = \eta\lambda$

Substituting these into Eq (2.158),

$$
St(x|\mu,\lambda,\nu) = \int_0^\infty \mathcal{N}(x|\mu,(\eta\lambda)^{-1})\cdot Gam(\eta\lambda|\frac{\nu}{2},\frac{\nu}{2\lambda})\lambda d\eta
$$

Let's calculate $Gam(\eta\lambda|\frac{\nu}{2},\frac{\nu}{2\lambda})$,

$$
\begin{aligned}
\text{Since } Gam(\lambda|a,b) &= \frac{1}{\Gamma(a)}b^a\lambda^{a-1}\exp(-b\lambda), \\
Gam(\eta\lambda|\frac{\nu}{2},\frac{\nu}{2\lambda}) &= \frac{1}{\Gamma(\frac{\nu}{2})}\cdot\left(\frac{\nu}{2\lambda}\right)^{\nu/2}(\eta\lambda)^{\nu/2-1}\cdot\exp(-\frac{\nu}{2\lambda}\lambda\eta) \\
&= \frac{1}{\Gamma(\frac{\nu}{2})}\cdot\left(\frac{\nu}{2}\right)^{\nu/2}\cdot\eta^{\nu/2-1}\cdot\frac{1}{\lambda^{\nu/2}}\cdot\lambda^{\nu/2-1}\cdot\exp(-\frac{\nu}{2}\eta) \\
&= \frac{1}{\Gamma(\frac{\nu}{2})}\cdot\left(\frac{\nu}{2}\right)^{\nu/2}\cdot\eta^{\nu/2-1}\cdot\frac{1}{\lambda}\exp(-\frac{\nu}{2}\eta) \\
&= \frac{1}{\lambda}Gam(\eta|\frac{\nu}{2},\frac{\nu}{2})
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
St(x|\mu,\lambda,\nu) &= \int_0^\infty \mathcal{N}(x|\mu,(\eta\lambda)^{-1})\cdot\frac{1}{\lambda}Gam(\eta|\frac{\nu}{2},\frac{\nu}{2})\cdot\lambda d\eta \\
&= \int_0^\infty \mathcal{N}(x|\mu,(\eta\lambda)^{-1})\ Gam(\eta|\frac{\nu}{2},\frac{\nu}{2}))d\eta
\end{aligned}
$$

Eq 2.213: (PRML p.115)

$$\mu_k = \frac{\exp(\eta_k)}{1 + \sum_j \exp(\eta_j)}$$

Proof :

$$\ln\left(\frac{\mu_k}{1 - \sum_{j=1}^{M-1} \mu_j}\right) = \eta_k \tag{2.212}$$

$$\frac{\mu_k}{1 - \sum_{j=1}^{M-1} \mu_j} = \exp(\eta_k) \tag{1}$$

$$\sum_{k=1}^{M} \mu_k = \sum_{k=1}^{M} \exp(\eta_k) \cdot \left[1 - \sum_{j=1}^{M-1} \mu_j\right]$$

Since LHS = 1,

$$\Rightarrow \quad 1 - \sum_{j=1}^{M-1} \mu_j = \frac{1}{\sum_{k=1}^{M} \exp(\eta_k)} \tag{2}$$

Substituting (2) into (1),

$$\begin{aligned} \mu_k &= \exp(\eta_k) \left[1 - \sum_{j=1}^{M-1} \mu_j\right] \\ &= \frac{\exp(\eta_k)}{\sum_{k=1}^{M} \exp(\eta_k)} \end{aligned} \tag{3}$$

When k = M in Eq (1),

$$\exp(\eta_M) = \frac{\mu_M}{1 - \sum_{j=1}^{M-1} \mu_j} = \frac{\mu_M}{\mu_M} = 1$$

Therefore Eq (3) becomes,

$$\begin{aligned} \mu_k &= \frac{\exp(\eta_k)}{\sum_{k=1}^{M} \exp(\eta_k)} = \frac{\exp(\eta_k)}{\exp(\eta_M) + \sum_{j=1}^{M-1} \exp(\eta_j)} \\ &= \frac{\exp(\eta_k)}{1 + \sum_j \exp(\eta_j)} \end{aligned}$$

Eq 2.214: (PRML p.115)

$$p(\mathbf{x}|\boldsymbol{\eta}) = \left[1 + \sum_{k=1}^{M-1} \exp(\boldsymbol{\eta}_k)\right]^{-1} \exp(\boldsymbol{\eta}^T \mathbf{x})$$

Proof :

$$\begin{aligned}
p(\mathbf{x}|\boldsymbol{\eta}) &= \exp\left\{\sum_{k=1}^{M} \mathbf{x}_k \ln \boldsymbol{\mu}_k\right\} \\
&= \exp\left\{\sum_{k=1} \mathbf{x}_k \ln\left(\frac{\boldsymbol{\mu}_k}{1 - \sum_{j=1}^{M-1} \boldsymbol{\mu}_j}\right) + \ln\left(1 - \sum_{k=1}^{M-1} \boldsymbol{\mu}_k\right)\right\} \\
&= \exp\left\{\sum_{k=1}^{M-1} \mathbf{x}_k \boldsymbol{\eta}_k\right\} \cdot \left\{1 - \sum_{k=1}^{M-1} \boldsymbol{\mu}_k\right\}
\end{aligned}$$

From Eq (2.212),

$$1 - \sum_{j=1}^{M-1} \boldsymbol{\mu}_j = \frac{\boldsymbol{\mu}_k}{\exp(\boldsymbol{\eta}_k)}$$

And using Eq (2.213),

$$\boldsymbol{\mu}_k = \frac{\exp(\boldsymbol{\eta}_k)}{1 + \sum_{j=1}^{M-1} \exp(\boldsymbol{\eta}_j)}$$

$$\begin{aligned}
1 - \sum_{j=1}^{M-1} \boldsymbol{\mu}_j &= \frac{\left(\frac{\exp(\boldsymbol{\eta}_k)}{1+\sum_{j=1}^{M-1} \exp(\boldsymbol{\eta}_j)}\right)}{\exp(\boldsymbol{\eta}_k)} \\
&= \left\{1 + \sum_{j=1}^{M-1} \exp(\boldsymbol{\eta}_j)\right\}^{-1}
\end{aligned}$$

Therefore,

$$\begin{aligned}
p(\mathbf{x}|\boldsymbol{\eta}) &= \exp\left\{\sum_{k=1}^{M-1} \mathbf{x}_k \boldsymbol{\eta}_k\right\} \left[1 + \sum_{k=1}^{M-1} \exp(\boldsymbol{\eta}_k)\right]^{-1} \\
&= \exp(\boldsymbol{\eta}^T \mathbf{x}) \left[1 + \sum_{k=1}^{M-1} \exp(\boldsymbol{\eta}_k)\right]^{-1}
\end{aligned}$$

Eq 2.231: (PRML p.118)

$$p_\eta(\eta) = p_\lambda(\lambda) \left|\frac{d\lambda}{d\eta}\right| = p_\lambda(\eta^2) 2\eta \ \propto \ \eta$$

Proof :

$\lambda = \eta^2$ ($\Longleftrightarrow \quad x = g(y)$ in PRML p.18)

From Eq (1.27),

$$\begin{aligned}
p_\eta(\eta) &= p\lambda(\lambda)\left|\frac{d\lambda}{d\eta}\right| \\
&= p\lambda(\eta^2)\left|\frac{d(\eta^2)}{d\eta}\right| \\
&= p\lambda(\eta^2)\cdot 2\eta
\end{aligned}$$

If $p_\lambda(\lambda) = \lambda^2 + 1$, for example,

Since $\lambda = \eta^2$,

$$p_\lambda(\eta^2) = \eta^4 + 1$$

From this, $p_\eta(\eta)$ is

$$p_\eta(\eta) = \eta^4 + 1$$

So p_λ and p_η are different functions!

Going back to our $p_\eta(\eta) = p_\lambda(\eta^2)\ 2\eta$ equation,

Since $p_\lambda(\eta^2)$ is still constant,

$$\therefore \;\; p_\eta(\eta) \;\propto\; \eta$$

Chapter 3. Linear Models for Regression

Eq 3.15: (PRML p.142)

$$\mathbf{w}_{ML} = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\bar{\mathbf{t}}$$

Proof :

$$\begin{aligned}
E_D(\mathbf{w}) &= \frac{1}{2}\sum_n\{t_n - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n)\}^2 \\
&= \frac{1}{2}\sum_n\{t_n - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n)\}^T\{t_n - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n)\} \qquad (3.12)
\end{aligned}$$

Utilizing Matrix Cookbook Eq (84),

$$\frac{\partial}{\partial\mathbf{s}}(\mathbf{x} - \mathbf{A}\mathbf{s})^T(\mathbf{x} - \mathbf{A}\mathbf{s}) = -2\mathbf{A}^T(\mathbf{x} - \mathbf{A}\mathbf{s})$$

$$\begin{aligned}\frac{\partial E_D}{\partial \mathbf{w}^T} &= -\sum_n 2\{t_n - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n)\}\,\boldsymbol{\phi}(\mathbf{x}_n)^T \\ &= -\sum_n 2t_n\boldsymbol{\phi}(\mathbf{x}_n)^T + \sum_n 2\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n)\boldsymbol{\phi}(\mathbf{x}_n)^T\end{aligned}$$

$$\boldsymbol{\Phi} = \begin{bmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & & & \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{bmatrix}$$

$$\boldsymbol{\phi}(\mathbf{x}_n) = \begin{bmatrix} \phi_0(\mathbf{x}_n) \\ \phi_1(\mathbf{x}_n) \\ \vdots \\ \phi_{M-1}(\mathbf{x}_n) \end{bmatrix} \quad \Rightarrow \quad \boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{\phi}(\mathbf{x}_1)^T \\ \boldsymbol{\phi}(\mathbf{x}_2)^T \\ \vdots \\ \boldsymbol{\phi}(\mathbf{x}_N)^T \end{bmatrix}$$

Setting $\dfrac{\partial E_D}{\partial \mathbf{w}^T} = 0$, we have

$$0 = \sum_n t_n\boldsymbol{\phi}(\mathbf{x}_n)^T - \mathbf{w}^T\sum_n \boldsymbol{\phi}(\mathbf{x}_n)\boldsymbol{\phi}(\mathbf{x}_n)^T$$

The first term:

$$\begin{aligned}\sum_n t_n\boldsymbol{\phi}(\mathbf{x}_n)^T &= t_1\boldsymbol{\phi}(\mathbf{x}_1)^T + t_2\boldsymbol{\phi}(\mathbf{x}_2)^T + \cdots + t_N\boldsymbol{\phi}(\mathbf{x}_N) \\ &= \bar{\mathbf{t}}^T\boldsymbol{\Phi}\end{aligned}$$

The second term:

$$\begin{aligned}\sum_n \boldsymbol{\phi}(\mathbf{x}_n)\boldsymbol{\phi}(\mathbf{x}_n)^T &= \begin{bmatrix} \boldsymbol{\phi}(\mathbf{x}_1) & \boldsymbol{\phi}(\mathbf{x}_2) & \cdots & \boldsymbol{\phi}(\mathbf{x}_N) \end{bmatrix} \begin{bmatrix} \boldsymbol{\phi}(\mathbf{x}_1)^T \\ \boldsymbol{\phi}(\mathbf{x}_2)^T \\ \vdots \\ \boldsymbol{\phi}(\mathbf{x}_N)^T \end{bmatrix} \\ &= \Phi^T\Phi\end{aligned}$$

$$\Rightarrow \qquad 0 = \bar{\mathbf{t}}^T\boldsymbol{\Phi} - \mathbf{w}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}$$

Taking transpose,

$$0 = \boldsymbol{\Phi}^T\bar{\mathbf{t}} - \boldsymbol{\Phi}^T\boldsymbol{\Phi}\mathbf{w}$$

$$\therefore \quad \mathbf{w} = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\overline{\mathbf{t}}$$

Eq 3.33: (PRML p.146)

$$\ln p(\mathbf{T}|\mathbf{X},\mathbf{W},\beta) = \frac{NK}{2}\ln\left(\frac{\beta}{2\pi}\right) - \frac{\beta}{2}\sum_{n=1}^{N}\|t_n - \mathbf{W}^T\boldsymbol{\phi}(\mathbf{x}_n)\|^2$$

Proof :

Difference between Eq (3.32) and Eq (3.8) is $\mathbf{t}$ and t.

$\mathbf{t}$: K target variables

t: single target varialbe

When there are N observations: $\mathbf{t} \to \mathbf{T}, \;\; t \to \overline{\mathbf{t}}$

Eq (3.11):

$$\ln p(\overline{\mathbf{t}}|\mathbf{w},\beta) = \sum_{n=1}^{N}\ln\mathcal{N}(t_n|\mathbf{w}^T\phi(\mathbf{x}_n),\beta^{-1})$$

Eq (3.33):

$$\begin{aligned}
\ln p(\mathbf{T}|\mathbf{X},\mathbf{W},\beta) &= \sum_{n=1}^{N}\ln\mathcal{N}(\mathbf{t}_n|\mathbf{W}^T\boldsymbol{\phi}(\mathbf{x}_n),\beta^{-1}\mathbf{I}) \\
&= \sum_{n=1}^{N}\ln\left[\frac{1}{(2\pi\beta^{-1})^{K/2}}\exp\left\{-\frac{1}{2\beta^{-1}}\|\mathbf{t}_n - \mathbf{W}^T\boldsymbol{\phi}(\mathbf{x}_n)\|^2\right\}\right] \\
&= \sum_{n=1}^{N}\frac{K}{2}\ln\left(\frac{\beta}{2\pi}\right) - \sum_{n=1}^{N}\frac{\beta}{2}\|t_n - \mathbf{W}^T\boldsymbol{\phi}(\mathbf{x}_n)\|^2 \\
&= \frac{NK}{2}\ln\left(\frac{\beta}{2\pi}\right) - \frac{\beta}{2}\sum_{n=1}^{N}\|t_n - \mathbf{W}^T\boldsymbol{\phi}(\mathbf{x}_n)\|^2
\end{aligned}$$

Eq 3.40: (PRML p.149)

$$\mathbb{E}_D[\{y(\mathbf{x};D) - h(\mathbf{x})\}^2] = \{\mathbb{E}_D[y(\mathbf{x};D)] - h(\mathbf{x})\}^2 + \mathbb{E}_D[\{y(\mathbf{x};D) - \mathbb{E}_D[y(\mathbf{x};D)]\}^2]$$

Proof :

To derive Eq (3.40) from Eq (3.39), all we have to do is to prove the final term in the expectation of Eq (3.39) vanishes.

$$
\begin{aligned}
&\mathbb{E}_D\{\text{The final term in Eq (3.39)}\} \\
&= \mathbb{E}_D\{\{y(\mathbf{x};D) - \mathbb{E}_D[y(\mathbf{x};D)]\} \cdot \{\mathbb{E}_D[y(\mathbf{x};D)] - h(\mathbf{x})\}\} \\
&= \mathbb{E}_D\{y(\mathbf{x};D) \cdot \mathbb{E}_D[y(\mathbf{x};D)] - y(x;D) \cdot h(\mathbf{x}) - \mathbb{E}_D[y(\mathbf{x};D) \cdot \mathbb{E}_D[y(\mathbf{x};D)] \\
&\quad + \mathbb{E}_D[y(\mathbf{x};D)] \cdot h(\mathbf{x})\} \\
&= \mathbb{E}_D[y(\mathbf{x};D)] \cdot \mathbb{E}_D[y(\mathbf{x};D)] - \mathbb{E}_D[y(\mathbf{x};D)] \cdot h(\mathbf{x}) - \mathbb{E}_D[y(\mathbf{x};D)] \cdot \mathbb{E}_D[y(\mathbf{x};D)] \\
&\quad + \mathbb{E}_D[y(\mathbf{x};D)] \cdot h(\mathbf{x}) \\
&= 0
\end{aligned}
$$

Eq 3.49: (PRML p.153)

$$
\begin{aligned}
& p(\mathbf{w}|\overline{\mathbf{t}}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N) \\
\text{where} \quad & \mathbf{m}_N = \mathbf{S}_N(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\boldsymbol{\Phi}^T\overline{\mathbf{t}}) \\
& \mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}
\end{aligned}
$$

Proof :

First off, let's show the following relationship,

$$
\prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^T\phi(\mathbf{x}_n), \beta^{-1}) = \mathcal{N}(\overline{\mathbf{t}}|\mathbf{w}^T\Phi, \beta^{-1}) \tag{1}
$$

$$
\text{where} \quad \overline{\mathbf{t}} = (t_1, t_2, \cdots, t_N)
$$

$$
\begin{aligned}
\boldsymbol{\Phi} &= \begin{bmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & & & \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{\phi}(\mathbf{x}_1) \\ \boldsymbol{\phi}(\mathbf{x}_2) \\ \vdots \\ \boldsymbol{\phi}(\mathbf{x}_N) \end{bmatrix}
\end{aligned}
$$

where $\boldsymbol{\phi}(\mathbf{x}_n) = [\phi_0(\mathbf{x}_n), \phi_1(\mathbf{x}_n), \cdots, \phi_{M-1}(\mathbf{x}_n)]$

$$\begin{aligned}
\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}) &= \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{x}-\boldsymbol{\mu})^2\right\} \\
\prod_{n=1}^{N}(\mathbf{x}_n|\boldsymbol{\mu}) &= \frac{1}{(2\pi\sigma^2)^{N/2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu})^2\right\}
\end{aligned}$$

Thus,

$$\prod_{n=1}^{N}\mathcal{N}(t_n|\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}) = \frac{1}{(2\pi\beta^{-1})^{N/2}} \exp\left\{-\frac{1}{2\beta^{-2}}\sum_{n=1}^{N}(t_n - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n))^2\right\} \tag{2}$$

$$\begin{aligned}
\mathcal{N}(\bar{\mathbf{t}}|\mathbf{w}^T\boldsymbol{\Phi}, \beta^{-1}) &= \frac{1}{(2\pi\beta^{-1})^{N/2}} \exp\left\{-\frac{1}{2\beta^{-2}}(\bar{\mathbf{t}} - \mathbf{w}^T\boldsymbol{\Phi})^2\right\} \\
&= \frac{1}{(2\pi\beta^{-1})^{N/2}} \exp\left\{-\frac{1}{2\beta^{-2}}[(t_1 - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_1))^2 + (t_2 - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_2))^2\right. \\
&\quad \left. + \cdots + (t_N - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_N))^2]\right\} \\
&= \frac{1}{(2\pi\beta^{-1})^{N/2}} \exp\left\{-\frac{1}{2\beta^{-2}}\sum_{n=1}^{N}(t_n - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n))^2\right\}
\end{aligned} \tag{3}$$

Since Eq (2) = Eq (3), we have proved Eq (1). Now let's prove Eq (3.49).

Eq (3.10) : $p(\bar{\mathbf{t}}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N}\mathcal{N}(t_n|\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1})$

Eq (3.48) : $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0)$

Using Eq (2.116) (and 2.113 $\sim$ 2.115)

$$\begin{aligned}
&p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \\
&p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{X} + b, \mathbf{L}^{-1}) \\
&p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + b, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Gamma}^{-1}\mathbf{A}^T) \\
&p(\mathbf{x}|\mathbf{y}) = (\mathbf{x}|\boldsymbol{\Sigma}\{\mathbf{A}^T\mathbf{L}(\mathbf{y} - b) + \boldsymbol{\Gamma}\boldsymbol{\mu}\}, \boldsymbol{\Sigma}) \qquad \leftarrow \text{Eq (2.116)} \\
&\text{where } \boldsymbol{\Sigma} = (\boldsymbol{\Gamma} + \mathbf{A}^T\mathbf{L}\mathbf{A})^{-1}
\end{aligned} \tag{4}$$

Comparing with Eqs (3.10) and (3.48),

$$\begin{aligned}
\mathbf{x} &\Longleftrightarrow \mathbf{w} \\
\mathbf{y} &\Longleftrightarrow \bar{\mathbf{t}} \\
\mathbf{A} &\Longleftrightarrow \boldsymbol{\Phi} \\
\boldsymbol{\Gamma} &\Longleftrightarrow \mathbf{S_0}^{-1} \\
\mathbf{L} &\Longleftrightarrow \beta \\
\mathbf{b} &\Longleftrightarrow 0 \\
\boldsymbol{\mu} &\Longleftrightarrow \mathbf{m}_0
\end{aligned}$$

$$\begin{aligned}
p(\bar{\mathbf{t}}|\mathbf{w},\beta) &= \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n),\beta^{-1}) \\
&= \mathcal{N}(\bar{\mathbf{t}}|\mathbf{w}^T\boldsymbol{\Phi},\beta^{-1}) \\
p(\mathbf{x}|\mathbf{y}) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}\{(\mathbf{A}^T\mathbf{L}(\mathbf{y}-b)+\boldsymbol{\Gamma}\boldsymbol{\mu}\},\boldsymbol{\Sigma}) \\
p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{m}_0,\mathbf{S}_0) \\
p(\mathbf{x}) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\Lambda^{-1})
\end{aligned}$$

Therefore, since $p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}\{(\mathbf{A}^T\mathbf{L}(\mathbf{y}-b)+\Lambda\boldsymbol{\mu}\},\boldsymbol{\Sigma})$
by substituting the parameters,

$$p(\mathbf{w}|\bar{\mathbf{t}},\beta) = \mathcal{N}(\mathbf{w}|\boldsymbol{\Sigma}\{\boldsymbol{\Phi}^T\beta\bar{\mathbf{t}}+\mathbf{S}_0^{-1}\mathbf{m}_0\},\boldsymbol{\Sigma})$$

By identifying $\boldsymbol{\Sigma}\{\boldsymbol{\Phi}^T\beta t+\mathbf{S}_0^{-1}\mathbf{m}_0\}$ as $\mathbf{m}_N$, and $\boldsymbol{\Sigma}$ as $\mathbf{S}_N$, we finally have,

$$\begin{aligned}
& p(\mathbf{w}|\bar{\mathbf{t}}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N,\mathbf{S}_N) \\
\text{where}\quad & \mathbf{m}_N = \mathbf{S}_N(\mathbf{S}_0^{-1}\mathbf{m}_0+\beta\boldsymbol{\Phi}^T\bar{\mathbf{t}}) \\
& \mathbf{S}_N^{-1} = \mathbf{S}_0^{-1}+\beta\boldsymbol{\Phi}^T\boldsymbol{\Phi} \qquad \text{(from Eq (4))}
\end{aligned}$$

Eq 3.55: (PRML p.153)

$$\ln p(\mathbf{w}|\bar{\mathbf{t}}) = -\frac{\beta}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n)\}^2 - \frac{\alpha}{2}\mathbf{w}^T\mathbf{w} + const$$

Proof :

Prior: $p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$ (3.52)

Likelihood: $p(\bar{\mathbf{t}}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N}\mathcal{N}(t_n|\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1})$

$$\begin{aligned}
\text{Posterior } &= \text{Prior x Likelihood} \\
&= \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) \cdot \prod_{n=1}^{N}\mathcal{N}(t_n|\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}) \\
&= \frac{1}{(2\pi\alpha^{-1})^{1/2}}\exp\left\{-\frac{1}{2\alpha^{-1}}\mathbf{w}^T\mathbf{w}\right\} \cdot \prod_{n=1}^{N}\frac{1}{(2\pi\beta^{-1})^{1/2}}\exp\left\{-\frac{1}{2\beta^{-1}}[t_n - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n)]^2\right\}
\end{aligned}$$

Taking log of the above equantion,

$$\therefore \ \ln p(\mathbf{w}|\bar{\mathbf{t}}) = -\frac{\beta}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n)\}^2 - \frac{\alpha}{2}\mathbf{w}^T\mathbf{w} + const$$

Eq 3.57: (PRML p.156)

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

Proof :

$$p(t|\bar{\mathbf{t}}, \alpha, \beta) = \int p(t, \mathbf{w}|\bar{\mathbf{t}}, \alpha, \beta)d\mathbf{w} \quad \leftarrow \text{sum rule}$$

The integrand is calculated to be

$$\begin{aligned}
p(t, \mathbf{w}|\bar{\mathbf{t}}, \alpha, \beta) &= \frac{p(t, \mathbf{w}, \bar{\mathbf{t}}, \alpha, \beta)}{p(\bar{\mathbf{t}}, \alpha, \beta)} \\
&= \frac{p(t|\mathbf{w}, \bar{\mathbf{t}}, \alpha, \beta) \cdot p(\mathbf{w}, \bar{\mathbf{t}}, \alpha, \beta)}{p(\bar{\mathbf{t}}, \alpha, \beta)} \\
&= \frac{p(t|\mathbf{w}, \bar{\mathbf{t}}, \alpha, \beta) \cdot p(\mathbf{w}|\bar{\mathbf{t}}, \alpha, \beta) \cdot p(\bar{\mathbf{t}}, \alpha, \beta)}{p(\bar{\mathbf{t}}, \alpha, \beta)} \\
&= p(t|\mathbf{w}, \bar{\mathbf{t}}, \alpha, \beta) \cdot p(\mathbf{w}|\bar{\mathbf{t}}, \alpha, \beta)
\end{aligned}$$

$$\therefore \quad p(t|\bar{\mathbf{t}}, \alpha, \beta) = \int p(t|\mathbf{w}, \bar{\mathbf{t}}, \alpha, \beta) \cdot p(\mathbf{w}|\bar{\mathbf{t}}, \alpha, \beta) d\mathbf{w}$$

Eq 3.63: (PRML p.160)

$$\begin{aligned}\operatorname{cov}[y(\mathbf{x}), y(\mathbf{x}')] &= \operatorname{cov}[\boldsymbol{\phi}(\mathbf{x})^T \mathbf{w}, \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}')] \\ &= \boldsymbol{\phi}(\mathbf{x})^T \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}')\end{aligned}$$

Proof :

$$\operatorname{cov}[\boldsymbol{\phi}(\mathbf{x})^T \mathbf{w}, \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}') = \mathbb{E}_w[\boldsymbol{\phi}(\mathbf{x})^T \mathbf{w} \cdot \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}')] - \mathbb{E}[\boldsymbol{\phi}(\mathbf{x})^T \mathbf{w}] \cdot \mathbb{E}[\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}')] \tag{1}$$

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

and Eq (3.49) $p(\mathbf{w}|\bar{\mathbf{t}}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$,

$$\begin{aligned}\mathbb{E}_w[\boldsymbol{\phi}(\mathbf{x})^T \mathbf{w} \cdot \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}')] &= \int_{-\infty}^{\infty} \boldsymbol{\phi}(\mathbf{x})^T \mathbf{w}^2 \boldsymbol{\phi}(\mathbf{x}') \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N) d\mathbf{w} \\ &= \boldsymbol{\phi}(\mathbf{x})^T \boldsymbol{\phi}(\mathbf{x}') \int_{-\infty}^{\infty} \mathbf{w}^2 \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N) d\mathbf{w} \\ &= \boldsymbol{\phi}(\mathbf{x})^T \boldsymbol{\phi}(\mathbf{x}') \cdot (\mathbf{m}_N^2 + \mathbf{S}_N) \\ &\text{(since } \mathbf{m}_N^2 = 0) \\ &= \boldsymbol{\phi}(\mathbf{x})^T \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}')\end{aligned}$$

Since we assume $\mathbf{m}_N = 0$, the second term in Eq (1) is 0. (Since $\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}$)

$$\therefore \quad \operatorname{cov}[y(\mathbf{x}), y(\mathbf{x}')] = \boldsymbol{\phi}(\mathbf{x})^T \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}')$$

Eq 3.74: (PRML p.165)

$$p(t|\bar{\mathbf{t}}) = \iiint p(t|\mathbf{w}, \beta)\ p(\mathbf{w}|\bar{\mathbf{t}}, \alpha, \beta)\ p(\alpha, \beta|\bar{\mathbf{t}}) d\mathbf{w} d\alpha d\beta$$

Proof :

$$
\begin{aligned}
p(t|\bar{\mathbf{t}}) &= \sum_{\mathbf{w},\alpha,\beta} p(t, \mathbf{w}, \alpha, \beta|\bar{\mathbf{t}}) \qquad : \text{ sum rule} \\
&= \sum p(t, \mathbf{w}, \alpha, \beta, \bar{\mathbf{t}}) \cdot \frac{1}{p(\bar{\mathbf{t}})} \\
&= \sum p(t|\mathbf{w}, \alpha, \beta, \bar{\mathbf{t}}) \cdot p(\mathbf{w}, \alpha, \beta, \bar{\mathbf{t}}) \cdot \frac{1}{p(\bar{\mathbf{t}})} \\
&= \sum p(t|\mathbf{w}, \alpha, \beta, \bar{\mathbf{t}}) \cdot p(\mathbf{w}|\alpha, \beta, \bar{\mathbf{t}}) \cdot p(\alpha, \beta, \bar{\mathbf{t}}) \cdot \frac{1}{p(\bar{\mathbf{t}})} \\
&= \sum p(t|\mathbf{w}, \alpha, \beta, \bar{\mathbf{t}}) \cdot p(\mathbf{w}|\alpha, \beta, \bar{\mathbf{t}}) \cdot p(\alpha, \beta|\bar{\mathbf{t}})
\end{aligned}
$$

From Eq (3.52),

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

$\rightarrow$ $\mathbf{w}$ depends on α, so when $\mathbf{w}$ is conditioned, α does not need to be conditioned. $\mathbf{w}$ also depends on $\bar{\mathbf{t}}$ (training targets), since $\mathbf{w}$ will be determined from the training data.

$\Rightarrow$ So α and $\bar{\mathbf{t}}$ will be dropped from $p(t|\mathbf{w}, \alpha, \beta, \bar{\mathbf{t}})$

$$\therefore \; p(t|\bar{\mathbf{t}}) = \iiint p(t|\mathbf{w}, \beta) \; p(\mathbf{w}|\bar{\mathbf{t}}, \alpha, \beta) \; p(\alpha, \beta|\bar{\mathbf{t}}) d\mathbf{w} d\alpha d\beta$$

Eq 3.77: (PRML p.166)

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

Proof :

$$
p(\bar{\mathbf{t}}|\alpha, \beta) = \int p(\bar{\mathbf{t}}, \mathbf{w}|\alpha, \beta) d\mathbf{w} \qquad \text{(sum rule)}
$$

$$
\begin{aligned}
p(\bar{\mathbf{t}}, \mathbf{w}|\alpha, \beta) &= \frac{p(\bar{\mathbf{t}}, \mathbf{w}, \alpha, \beta)}{p(\alpha, \beta)} \\
&= \frac{p(\bar{\mathbf{t}}|\mathbf{w}, \alpha, \beta) \cdot p(\mathbf{w}, \alpha, \beta)}{p(\alpha, \beta)} \\
&= \frac{p(\bar{\mathbf{t}}|\mathbf{w}, \alpha, \beta) \cdot p(\mathbf{w}|\alpha, \beta) \cdot p(\alpha, \beta)}{p(\alpha, \beta)} \\
&= p(\bar{\mathbf{t}}|\mathbf{w}, \alpha, \beta) \cdot p(\mathbf{w}|\alpha, \beta)
\end{aligned}
$$

$\mathbf{w}$ inclues α's information and does not have anything to do with β.

$$\therefore \quad p(\bar{\mathbf{t}}|\alpha,\beta) = \int p(\bar{\mathbf{t}}|\mathbf{w},\beta)\; p(\mathbf{w}|\alpha) d\mathbf{w}$$

Eq 3.78: (PRML p.166)

$$p(\bar{\mathbf{t}}|\alpha,\beta) = \left(\frac{\beta}{2\pi}\right)^{N/2} \left(\frac{\alpha}{2\pi}\right)^{M/2} \int \exp\{-E(\mathbf{w})\} d\mathbf{w}$$

Proof :

We will need two previous equations to derive Eq 3.78.

Eq 3.10: $\quad p(\bar{\mathbf{t}}|\mathbf{X},\mathbf{w},\beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n),\beta^{-1})$

Eq 3.52: $\quad p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0},\alpha^{-1}\mathbf{I})$

$$\begin{aligned}
\prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n),\beta^{-1}) &= \prod_{n=1}^{N} \left(\frac{\beta}{2\pi}\right)^{1/2} \exp\left[-\frac{\beta}{2}(t_n - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n))^2\right] \\
&= \left(\frac{\beta}{2\pi}\right)^{N/2} \prod_{n=1}^{N} \exp\left[-\frac{\beta}{2}(t_n - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n))^2\right]
\end{aligned}$$

$$\mathcal{N}(\mathbf{w}|\mathbf{0},\alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{M/2} \prod_{m=1}^{M} \exp\left(-\frac{\alpha}{2} w_m^T w_m\right)$$

Using the above two equations, we can derive $p(\bar{\mathbf{t}}|\alpha,\beta)$.

$$p(\bar{\mathbf{t}}|\alpha,\beta) = \int p(\bar{\mathbf{t}}|\mathbf{w},\beta)\; p(\mathbf{w}|\alpha) d\mathbf{w} \qquad \text{(sum rule)}$$

$$\begin{aligned}
p(\bar{\mathbf{t}}|\mathbf{w},\beta)\; p(\mathbf{w}|\alpha) &= \left(\frac{\beta}{2\pi}\right)^{N/2} \prod_{n=1}^{N} \exp\left[-\frac{\beta}{2}(t_n - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n))^2\right] \cdot \left(\frac{\alpha}{2\pi}\right)^{M/2} \prod_{m=1}^{M} \exp\left(-\frac{\alpha}{2} w_m^T w_m\right) \\
&= \left(\frac{\beta}{2\pi}\right)^{N/2} \left(\frac{\alpha}{2\pi}\right)^{M/2} \exp\left\{\sum_{n=1}^{N}\left(-\frac{\beta}{2}\right)(t_n - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n))^2 + \sum_{m=1}^{M}\left(-\frac{\alpha}{2} w_m^T w_m\right)\right\}
\end{aligned}$$

We can identify

$$\sum_{n=1}^{N}\left(-\frac{\beta}{2}\right)(t_n - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n))^2 = -\frac{\beta}{2}\|\bar{\mathbf{t}} - \boldsymbol{\Phi}\mathbf{w}\|^2$$

$$\sum_{m=1}^{M}\left(-\frac{\alpha}{2} w_m^T w_m\right) = -\,\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}$$

$$\therefore \quad p(\bar{\mathbf{t}}|\alpha,\beta) = \left(\frac{\beta}{2\pi}\right)^{N/2}\left(\frac{\alpha}{2\pi}\right)^{M/2}\int \exp\{-E(\mathbf{w})\}d\mathbf{w}$$
$$\text{where } E(\mathbf{w}) = \beta E_D(\mathbf{w}) + \alpha E_w(\mathbf{w}) = \frac{\beta}{2}\|\bar{\mathbf{t}} - \Phi\mathbf{w}\|^2 + \frac{\alpha}{2}\mathbf{w}^T\mathbf{w}$$

Chapter 4. Linear Models for Classification

Eq 4.15: (PRML p.185)

$$E_D(\widetilde{\mathbf{W}}) = \frac{1}{2}\,\mathrm{Tr}\left\{(\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{T})^T\,(\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{T})\right\}$$

Proof :

$\mathbf{w}_k$: column vector for a class K = k . (m x 1 dim)

$\mathbf{x}_n$: column vector for a sample #n. (m x 1 dim)

$\widetilde{\mathbf{W}} = (\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_K)$. (m x K dim) ($\mathbf{w}_k$ is a column vector)

$$\widetilde{\mathbf{X}} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_N^T \end{bmatrix}$$

$\widetilde{\mathbf{X}}$ has N x m dimension, and $\mathbf{x}_n^T$ is a row vector.

$\mathbf{t}_n$: column vector for a sample #n. (K x 1 dim)

$$\widetilde{\mathbf{T}} = \begin{bmatrix} \mathbf{t}_1^T \\ \mathbf{t}_2^T \\ \vdots \\ \mathbf{t}_N^T \end{bmatrix}$$

$\widetilde{\mathbf{T}}$ has N x K dimension, and $\mathbf{t}_n^T$ is a row vector.

$\widetilde{\mathbf{X}}\widetilde{\mathbf{W}}$ = (N x m) · (m x K) = N x K

Therefore $(\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{T})^T \cdot (\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{T})$ = (N x K)^T· (N x K) = K x K (a square matrix)

If you take a trace of $(\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{T})^T \cdot (\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{T})$, you will get

$$\sum_{k=1}^{K} \|\bar{\mathbf{t}}_{(k,k)} - (\widetilde{\mathbf{X}}\widetilde{\mathbf{W}})_{(k,k)}\|^2$$

Eq 4.16: (PRML p.185)

$$\widetilde{\mathbf{W}} = (\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^T\mathbf{T}$$

Proof :

From Eq (4.15),

$$\begin{aligned}
E_D(\widetilde{\mathbf{W}}) &= \frac{1}{2}\mathrm{Tr}\left\{(\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{T})^T(\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{T})\right\} \\
&= \frac{1}{2}\mathrm{Tr}\left\{(\widetilde{\mathbf{W}}^T\widetilde{\mathbf{X}}^T - \mathbf{T}^T)(\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \mathbf{T})\right\} \\
&= \frac{1}{2}\mathrm{Tr}\left\{\widetilde{\mathbf{W}}^T\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} - \widetilde{\mathbf{W}}^T\widetilde{\mathbf{X}}^T\mathbf{T} - \mathbf{T}^T\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} + \mathbf{T}^T\mathbf{T})\right\}
\end{aligned}$$

Taking a derivative w.r.t. $\widetilde{\mathbf{W}}$,

$$\frac{\partial E_D(\widetilde{\mathbf{W}})}{\partial \widetilde{\mathbf{W}}} = \frac{1}{2}\left\{(\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}}\widetilde{\mathbf{W}} + \widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}}\widetilde{\mathbf{W}}) - \widetilde{\mathbf{X}}^T\mathbf{T} - (\mathbf{T}^T\widetilde{\mathbf{X}})^T + 0\right\} \tag{1}$$

Used were the matrix derivative formula from Matrix Cookbook section 2.4, p.11.

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{X}}\mathrm{Tr}(\mathbf{X}^T\mathbf{B}\mathbf{X}) &= \mathbf{B}\mathbf{X} + \mathbf{B}^T\mathbf{X} \\
\frac{\partial}{\partial \mathbf{X}}\mathrm{Tr}(\mathbf{B}\mathbf{X}) &= \mathbf{B}^T \\
\frac{\partial}{\partial \mathbf{X}}\mathrm{Tr}(\mathbf{X}^T\mathbf{C}) &= \mathbf{C}
\end{aligned}$$

Eq (1) becomes

$$(\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}} + \widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}})\widetilde{\mathbf{W}} = 2\widetilde{\mathbf{X}}^T\mathbf{T}$$

$$\therefore\ \widetilde{\mathbf{W}} = (\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^T\mathbf{T}$$

Eq 4.29: (PRML p.189)

$$(\mathbf{w}^T\mathbf{S}_B\mathbf{w})\mathbf{S}_W\mathbf{w} = (\mathbf{w}^T\mathbf{S}_W\mathbf{w})\mathbf{S}_B\mathbf{w}$$

Proof :

From Matrix Cookbook p.11,

$$\frac{\partial \mathbf{x}^T\mathbf{B}\mathbf{x}}{\partial \mathbf{x}} = (\mathbf{B} + \mathbf{B}^T)\mathbf{x}$$

$$\begin{aligned}
J(\mathbf{w}) &= \frac{\mathbf{w}^T\mathbf{S}_B\mathbf{w}}{\mathbf{w}^T\mathbf{S}_W\mathbf{w}} \\
\frac{\partial J(\mathbf{w})}{\partial \mathbf{w}} &= \frac{\mathbf{w}^T\mathbf{S}_W\mathbf{w}(\mathbf{S}_B + \mathbf{S}_\mathbf{S}^T)\mathbf{w} - \mathbf{w}^T\mathbf{S}_B\mathbf{w}(\mathbf{S}_W + \mathbf{S}_W^T)\mathbf{w}}{(\mathbf{w}^T\mathbf{S}_W\mathbf{w})^2} \\
&= 0
\end{aligned}$$

Since $\mathbf{S}_B^T = \mathbf{S}_B$ and $\mathbf{S}_W^T = \mathbf{S}_W$,

$$\therefore \quad (\mathbf{w}^T\mathbf{S}_B\mathbf{w})\mathbf{S}_W\mathbf{w} = (\mathbf{w}^T\mathbf{S}_W\mathbf{w})\mathbf{S}_B\mathbf{w}$$

Eq 4.65: (PRML p.198)

$$p(\mathbf{C}_1|\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x} + w_0)$$

$$\begin{aligned}
\text{where } \mathbf{w} &= \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\
w_0 &= -\frac{1}{2}\mu_1^T\Sigma^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^T\Sigma^{-1}\boldsymbol{\mu}_2 + \ln\frac{p(C_1)}{p(C_2)}
\end{aligned}$$

Proof :

Eqs (4.57) and (4.58) say,

$$\begin{aligned}
p(C_1|\mathbf{x}) &= \sigma(a) \\
\text{where } a = &\ \ln\frac{p(\mathbf{x}|C_1)p(C_1)}{p(\mathbf{x}|C_2)p(C_2)}
\end{aligned}$$

Let's calculate $\dfrac{p(\mathbf{x}|C_1)}{p(\mathbf{x}|C_2)}$

$$
\begin{aligned}
p(\mathbf{x}|C_1) &= \frac{1}{(2\pi)^{D/2}} \frac{1}{|\mathbf{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\right\} \\
p(\mathbf{x}|C_2) &= \frac{1}{(2\pi)^{D/2}} \frac{1}{|\mathbf{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right\} \\
\frac{p(\mathbf{x}|C_1)}{p(\mathbf{x}|C_2)} &= \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right\} \qquad (1)
\end{aligned}
$$

Let's calculate the exponent inside the exponential function,

$$
\begin{aligned}
&-\frac{1}{2}(\mathbf{x}^T-\boldsymbol{\mu}_1)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\mathbf{x}^T-\boldsymbol{\mu}_2)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) \\
&= -\frac{1}{2}(\mathbf{x}^T\mathbf{\Sigma}^{-1}-\boldsymbol{\mu}_1^T\mathbf{\Sigma}^{-1})(\mathbf{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\mathbf{x}^T\mathbf{\Sigma}^{-1}-\boldsymbol{\mu}_2^T\mathbf{\Sigma}^{-1})(\mathbf{x}-\boldsymbol{\mu}_2) \\
&= -\frac{1}{2}(\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x}-\mathbf{x}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_1-\boldsymbol{\mu}_1^T\mathbf{\Sigma}^{-1}\mathbf{x}+\boldsymbol{\mu}_1^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_1) \\
&\quad + \frac{1}{2}(\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x}-\mathbf{x}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_2-\boldsymbol{\mu}_2^T\mathbf{\Sigma}^{-1}\mathbf{x}+\boldsymbol{\mu}_2^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_2) \\
&= \frac{1}{2}(\boldsymbol{\mu}_1^T\mathbf{\Sigma}^{-1}-\boldsymbol{\mu}_2^T\mathbf{\Sigma}^{-1})\mathbf{x} + \frac{1}{2}\mathbf{x}^T(\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_1-\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_2) - \frac{1}{2}\boldsymbol{\mu}_1^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_2 \\
&= \frac{1}{2}(\boldsymbol{\mu}_1^T-\boldsymbol{\mu}_2^T)\mathbf{\Sigma}^{-1}\mathbf{x} + \frac{1}{2}\mathbf{x}^T\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) - \frac{1}{2}\boldsymbol{\mu}_1^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_2 \\
&\text{(Since } \mathbf{x}^T\mathbf{M}\mathbf{y} = \mathbf{y}^T\mathbf{M}\mathbf{x} \text{ if } \mathbf{M} \text{ is symmetric)} \\
&= (\boldsymbol{\mu}_1^T-\boldsymbol{\mu}_2^T)\mathbf{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_1^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_2
\end{aligned}
$$

If we define $\mathbf{w}^T = (\boldsymbol{\mu}_1^T-\boldsymbol{\mu}_2^T)\mathbf{\Sigma}^{-1}$, then Eq (1) becomes

$$
\frac{p(\mathbf{x}|C_1)}{p(\mathbf{x}|C_2)} = \exp\left\{\mathbf{w}^T\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_1^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_2\right\}
$$

Now we can calculate $p(C_1|\mathbf{x})$,

$$
\begin{aligned}
p(C_1|\mathbf{x}) &= \sigma\left[\ln\frac{p(\mathbf{x}|C_1)p(C_1)}{p(\mathbf{x}|C_2)p(C_2)}\right] \\
&= \sigma\left[\ln\frac{p(\mathbf{x}|C_1)}{p(\mathbf{x}|C_2)} + \ln\frac{p(C_1)}{p(C_2)}\right] \\
&= \sigma\left[\mathbf{w}^T\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_1^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_2 + \ln\frac{p(C_1)}{p(C_2)}\right] \\
&= \sigma\left[\mathbf{w}^T\mathbf{x} + w_0\right]
\end{aligned}
$$

$$
\text{where} \quad w_0 = -\frac{1}{2}\boldsymbol{\mu}_1^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_2 + \ln\frac{p(C_1)}{p(C_2)}
$$

Eq 4.75: (PRML p.201)

$$\boldsymbol{\mu}_1 = \frac{1}{N_1}\sum_{n=1}^{N} t_n \mathbf{x}_n$$

Proof :

From Eq (4.74)

$$\ln p(\bar{\mathbf{t}}|\pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \{t_n \ln [\pi \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma})] + (1-t_n)\ln[(1-\pi)\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_2, \boldsymbol{\Sigma})]\}$$

Differentiate w.r.t. $\boldsymbol{\mu}_1$,

$$\frac{\partial \ln p}{\partial \boldsymbol{\mu}_1} = \sum_{n=1}^{N} t_n \frac{\pi \frac{\partial \mathcal{N}}{\partial \boldsymbol{\mu}_1}}{\pi \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma})} \tag{1}$$

$$\text{where} \quad \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) = C \exp\left\{-\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_1)\right\}$$

$$\frac{\partial \mathcal{N}}{\partial \boldsymbol{\mu}_1} = \mathcal{N} \cdot \frac{\partial}{\partial \boldsymbol{\mu}_1}\left\{-\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_1)\right\} \tag{2}$$

Plugging Eq (2) into (1),

$$\begin{aligned}
\frac{\partial \ln p}{\partial \boldsymbol{\mu}_1} &= \sum_{n=1}^{N} t_n \frac{\partial}{\partial \boldsymbol{\mu}_1}\left\{-\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_1)\right\} \\
&= \sum_{n=1}^{N} t_n \frac{\partial}{\partial \boldsymbol{\mu}_1}\left\{-\frac{1}{2}(\mathbf{x}_n^T \boldsymbol{\Sigma}^{-1}\mathbf{x}_n - \mathbf{x}_n^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1}\mathbf{x}_n + \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1)\right\}
\end{aligned}$$

Using the matrix derivative formula,

$$\begin{aligned}
\frac{\partial \mathbf{x}^T \mathbf{a}}{\partial \mathbf{x}} &= \mathbf{a} \\
\frac{\partial \mathbf{x}^T \mathbf{B} \mathbf{x}}{\partial \mathbf{x}} &= (\mathbf{B} + \mathbf{B}^T)\mathbf{x} \\
\frac{\partial \mathbf{a}^T \mathbf{x}}{\partial \mathbf{x}} &= \mathbf{a}
\end{aligned}$$

We have

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\mu}_1}(\boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1}\mathbf{x}_n) &= \boldsymbol{\Sigma}^{-1}\mathbf{x}_n \\
\frac{\partial}{\partial \boldsymbol{\mu}_1}(\boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1) &= (\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1})\boldsymbol{\mu}_1 = 2\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 \\
\frac{\partial}{\partial \boldsymbol{\mu}_1}(\mathbf{x}_n^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1) &= \boldsymbol{\Sigma}^{-1}\mathbf{x}_n
\end{aligned}$$

$$\frac{\partial \ln p}{\partial \boldsymbol{\mu}_1} = -\frac{1}{2}\sum_{n=1}^{N} t_n(-2\boldsymbol{\Sigma}^{-1}\mathbf{x}_n + 2\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1) = 0$$

$$
\Rightarrow \quad \sum_{n=1}^{N} t_n \mathbf{x}_n = \sum_{n=1}^{N} t_n \boldsymbol{\mu}_1 = N_1 \boldsymbol{\mu}_1
$$

$$
\therefore \quad \boldsymbol{\mu}_1 = \frac{1}{N_1} \sum_{n=1}^{N} t_n \mathbf{x}_n
$$

Eqs 4.77: (PRML p.201)

$$
\begin{aligned}
& -\frac{1}{2}\sum_{n=1}^{N} t_n \ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N} t_n (\mathbf{x}_n - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_1) \\
& -\frac{1}{2}\sum_{n=1}^{N} (1-t_n) \ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N} (1-t_n) (\mathbf{x}_n - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_2) \\
& = -\frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{N}{2}\mathrm{Tr}\{\boldsymbol{\Sigma}^{-1}\mathbf{S}\} \\
\text{where} \quad & \mathbf{S} = \frac{N_1}{N}\mathbf{S}_1 + \frac{N_2}{N}\mathbf{S}_2 \\
& \mathbf{S}_1 = \frac{1}{N_1}\sum_{n \in C_1} (\mathbf{x}_n - \boldsymbol{\mu}_1)(\mathbf{x}_n - \boldsymbol{\mu}_1)^T \\
& \mathbf{S}_2 = \frac{1}{N_2}\sum_{n \in C_2} (\mathbf{x}_n - \boldsymbol{\mu}_2)(\mathbf{x}_n - \boldsymbol{\mu}_2)^T
\end{aligned}
$$

Proof :

From Eq (4.71),

$$
\ln p = \sum_{n=1}^{N} \Big\{ t_n \ln\left[\pi \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_1, \boldsymbol{\Sigma}) + (1-t_n)\ln\left[(1-\pi)\mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_2, \boldsymbol{\Sigma})\right]\Big\}
$$

Since $\mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}) = C \dfrac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{ -\dfrac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k) \right\}$,

$$
\begin{aligned}
\ln p = \sum_{n=1}^{N} \Big\{ & t_n [\ln \pi + \ln C - \frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_1) \\
& + (1-t_n)[\ln(1-\pi) + \ln C - \frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_2)] \Big\}
\end{aligned}
$$

Picking out the terms that depend on $\mathbf{\Sigma}$,

$$
\begin{aligned}
&-\frac{1}{2}\sum_{n=1}^{N} t_n \ln|\mathbf{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N} t_n(\mathbf{x}_n-\boldsymbol{\mu}_1)^T\mathbf{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu}_1) - \frac{1}{2}\sum_{n=1}^{N}(1-t_n)\ln|\mathbf{\Sigma}| \\
&-\frac{1}{2}\sum_{n=1}^{N}(1-t_n)(\mathbf{x}_n-\boldsymbol{\mu}_2)^T\mathbf{\Sigma}^{-1}(x_n-\boldsymbol{\mu}_2) \\
&= -\frac{1}{2}\ln|\mathbf{\Sigma}| - \frac{1}{2}\sum_{n\in C_1}^{N}(\mathbf{x}_n-\boldsymbol{\mu}_1)^T\mathbf{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu}_1) \\
&\quad -\frac{1}{2}\sum_{n\in C_2}^{N}(\mathbf{x}_n-\boldsymbol{\mu}_2)^T\mathbf{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu}_2) \qquad (1)
\end{aligned}
$$

Since $\mathbf{x}^T\mathbf{M}\mathbf{x} = \mathrm{Tr}(\mathbf{M}\mathbf{x}\mathbf{x}^T)$.

$$
(\mathbf{x}_n-\boldsymbol{\mu}_k)^T\mathbf{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu}_k) = \mathrm{Tr}[\,\mathbf{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu}_k)(\mathbf{x}_n-\boldsymbol{\mu}_k)^T]
$$

The Eq (1) becomes

$$
\begin{aligned}
&-\frac{N}{2}\ln|\mathbf{\Sigma}| - \frac{1}{2}\sum_{n\in C_1}^{N}\mathrm{Tr}[\mathbf{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu}_1)(\mathbf{x}_n-\boldsymbol{\mu}_1)^T] - \frac{1}{2}\sum_{n\in C_2}^{N}\mathrm{Tr}[\mathbf{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu}_2)(\mathbf{x}_n-\boldsymbol{\mu}_2)^T] \\
&= -\frac{N}{2}\ln|\mathbf{\Sigma}| - \frac{N}{2}\mathrm{Tr}\left\{\Sigma^{-1}\left[\frac{N_1}{N}\frac{1}{N_1}\sum_{n\in C_1}^{N}(\mathbf{x}_n-\boldsymbol{\mu}_1)(\mathbf{x}_n-\boldsymbol{\mu}_1)^T\right.\right. \\
&\qquad\qquad \left.\left.+\frac{N_2}{N}\frac{1}{N_2}\sum_{n\in C_2}^{N}(\mathbf{x}_n-\boldsymbol{\mu}_2)(\mathbf{x}_n-\boldsymbol{\mu}_2)^T\right]\right\} \\
&= -\frac{N}{2}\ln|\mathbf{\Sigma}| - \frac{N}{2}\mathrm{Tr}\{\mathbf{\Sigma}^{-1}\mathbf{S}\}
\end{aligned}
$$

where

$$
\begin{aligned}
\mathbf{S} &= \frac{N_1}{N}\mathbf{S}_1 + \frac{N_2}{N}\mathbf{S}_2 \\
\mathbf{S}_1 &= \frac{1}{N_1}\sum_{n\in C_1}(\mathbf{x}_n-\boldsymbol{\mu}_1)(\mathbf{x}_n-\boldsymbol{\mu}_1)^T \\
\mathbf{S}_2 &= \frac{1}{N_2}\sum_{n\in C_2}(\mathbf{x}_n-\boldsymbol{\mu}_2)(\mathbf{x}_n-\boldsymbol{\mu}_2)^T
\end{aligned}
$$

Eq 4.107: (PRML p.209)

$$
p(\mathbf{T}|\mathbf{w}_1,\cdots,\mathbf{w}_k) = \prod_{n=1}^{N}\prod_{k=1}^{K} p(C_k|\boldsymbol{\phi}_n)^{t_{nk}} = \prod_{n=1}^{N}\prod_{k=1}^{K} y_{nk}^{t_{nk}}
$$

Proof :

$\boldsymbol{\phi}_n \equiv \boldsymbol{\phi}(\mathbf{x}_n)$

n is the number for a set of input $\mathbf{x}$. In other words, n is one of the N data sets.

For each $\mathbf{x}, \boldsymbol{\phi}(\mathbf{x})$ is calculated for calculation in $\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x})$. $\mathbf{x}$ is the raw input values.

For K = 1,

$$p(C_1|\boldsymbol{\phi}_1)^{t_{11}} \cdot p(C_1|\boldsymbol{\phi}_2)^{t_{21}} \cdots \; p(C_1|\boldsymbol{\phi}_N)^{t_{N1}}$$

where $t_{11} = 1, t_{21} = 0, \cdots, t_{N1} = 1$ for example.

For K = 2,

$$p(C_2|\boldsymbol{\phi}_1)^{t_{12}} \cdot p(C_2|\boldsymbol{\phi}_2)^{t_{22}} \cdots \; p(C_2|\boldsymbol{\phi}_N)^{t_{N2}}$$

$\vdots$

Putting all these together,

$$\begin{aligned} p(\mathbf{T}|\mathbf{w}_1, \cdots, \mathbf{w}_k) &= \prod_{n=1}^{N}\prod_{k=1}^{K} p(C_k|\boldsymbol{\phi}_n)^{t_{nk}} \\ &\text{(since } \; y_k(\boldsymbol{\phi}_n) = p(C_k|\boldsymbol{\phi}_n)) \\ &= \prod_{n=1}^{N}\prod_{k=1}^{K} [y_k(\boldsymbol{\phi}_n)]^{t_{nk}} \\ &= \prod_{n=1}^{N}\prod_{k=1}^{K} y_{nk}^{t_{nk}} \\ &\text{where } y_{nk} \equiv y_k(\boldsymbol{\phi}_n) \end{aligned}$$

Eq 4.119: (PRML p.212)

$$y \equiv \mathbb{E}[t|\eta] = -s\frac{d}{d\eta}\ln g(\eta)$$

Proof :

$$p(t|\eta, s) = \frac{1}{s}h\left(\frac{t}{s}\right)g(\eta)\exp\left\{\frac{\eta t}{s}\right\}$$

$$\int p(t|\eta, s)dt = \int \frac{1}{s}h\left(\frac{t}{s}\right)g(\eta)\exp\left\{\frac{\eta t}{s}\right\}dt = 1$$

Taking a derivative w.r.t η,

$$\frac{dg(\eta)}{d\eta}\int \frac{1}{s}h\left(\frac{t}{s}\right)\exp\left\{\frac{\eta t}{s}\right\}dt + \int \frac{1}{s}h\left(\frac{t}{s}\right)g(\eta)\exp\left\{\frac{\eta t}{s}\right\}\cdot\frac{t}{s}dt = 0$$

$$\text{Since} \int \frac{1}{s} h\left(\frac{t}{s}\right) \exp\left\{\frac{\eta t}{s}\right\} dt = \frac{1}{g(\eta)}$$

$$\text{and} \int \frac{1}{s} h\left(\frac{t}{s}\right) g(\eta)\text{exp}\left\{\frac{\eta t}{s}\right\} \cdot \frac{t}{s} dt = p(t|\eta, s)$$

$$-\frac{1}{g(\eta)} \frac{dg(\eta)}{d\eta} = \frac{1}{s} \int p(t|\eta, s) t dt = \frac{1}{s} \mathbb{E}[t|\eta]$$

$$\therefore \quad \mathbb{E}[t|\eta] = -s \frac{d}{d\eta} \ln g(\eta)$$

Eq 4.124: (PRML p.213)

$$\nabla \ln E(\mathbf{w}) = \frac{1}{s} \sum_{n=1}^{N} \{y_n - t_n\} \boldsymbol{\phi}$$

Proof :

From Eq (3.11),

$$\begin{aligned} \ln p(\overline{\mathbf{t}}|\mathbf{w}, \beta) &= \sum_{n=1}^{N} \ln \mathcal{N}(t_n | \mathbf{w}^T \phi(\mathbf{x}_n), \beta^{-1}) \\ &= \frac{N}{2} \ln \beta - \frac{N}{2} \ln (2\pi) - \beta E_D(\mathbf{w}) \end{aligned}$$

The error function $E_D(\mathbf{w})$ is defined as,

$$E_D(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{t_n - \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_n)\}^2$$

Since $\ln p(\overline{\mathbf{t}}|\mathbf{w}, \beta)$ depends on $\mathbf{w}$ through $E_D(\mathbf{w})$ only,

$$\nabla_{\mathbf{w}} \ln p(\overline{\mathbf{t}}|\mathbf{w}, \beta) = -\nabla_{\mathbf{w}} E_D(\mathbf{w})$$

Therefore, to obtain the Eq (4.124), you need to take a derivative $\ln p(t|\eta, s)$ of Eq (4.122).

$$\begin{aligned} \Rightarrow \qquad \nabla_{\mathbf{w}} \ln E(\mathbf{w}) &= -\nabla_{\mathbf{w}} \ln p(t|\eta, s) \\ &= -\sum_{n=1}^{N} \frac{1}{s} \{t_n - y_n\} \Psi'(y_n) f'(a_n) \boldsymbol{\phi}_n \\ &\text{(since } \Psi'(y_n) f'(a_n) = 1) \\ &= \frac{1}{s} \sum_{n=1}^{N} \{y_n - t_n\} \boldsymbol{\phi} \end{aligned}$$

Eq 4.149: (PRML p.219)

$$\mu_a = \mathbb{E}[a] = \int p(a)a\,da = \int q(\mathbf{w})\mathbf{w}^T\boldsymbol{\phi}d\mathbf{w} = \mathbf{w}_{MAP}^T\boldsymbol{\phi}$$

Proof :

$$\begin{aligned}
\mu_a &= \int p(a)ada \\
&= \int \left[\int \delta(a - \mathbf{w}^T\boldsymbol{\phi})q(\mathbf{w})d\mathbf{w}\right] a\,da \\
&= \int \left[\int \delta(a - \mathbf{w}^T\boldsymbol{\phi})a\,da\right] q(\mathbf{w})d\mathbf{w} \\
&= \int \mathbf{w}^T\boldsymbol{\phi}q(\mathbf{w})d\mathbf{w} \\
&= \int \mathbf{w}^T\boldsymbol{\phi}\mathcal{N}(\mathbf{w}|\mathbf{w}_{MAP}, \mathbf{S}_N)d\mathbf{w} \\
&= \boldsymbol{\phi}\int \mathbf{w}^T\mathcal{N}(\mathbf{w}|\mathbf{w}_{MAP}, \mathbf{S}_N)d\mathbf{w} \\
&= \boldsymbol{\phi}\,\mathbf{w}_{MAP}^T
\end{aligned}$$

Here we used

$$\int \mathbf{x}\,\mathcal{N}(\mathbf{x}|\mathbf{m}_\mathbf{x}, \mathbf{S})d\mathbf{x} = \mathbf{m}_\mathbf{x}$$

Eq 4.150: (PRML p.219)

$$\begin{aligned}
\sigma_a^2 = \mathrm{var}[a] &= \int p(a)\{a^2 - \mathbb{E}[a]^2\}da \\
&= \int q(\mathbf{w})\{(\mathbf{w}^T\boldsymbol{\phi})^2 - (\mathbf{m}_N^T\,\boldsymbol{\phi})^2\}d\mathbf{w} = \boldsymbol{\phi}^T\mathbf{S}_N\boldsymbol{\phi}
\end{aligned}$$

Proof :

$$\begin{aligned}
\sigma_a^2 &= \int p(a)\{a^2 - \mathbb{E}[a]^2\}da \\
&= \iint \delta(a - \mathbf{w}^T\boldsymbol{\phi})q(\mathbf{w})d\mathbf{w}\{a^2 - (\mathbf{w}_{MAP}^2\boldsymbol{\phi})^2\}da \\
&= \int \left[\int \delta(a - \mathbf{w}^T\boldsymbol{\phi})\{a^2 - (\mathbf{w}_{MAP}^2\boldsymbol{\phi})^2\}da\right] q(\mathbf{w})d\mathbf{w} \\
&= \int [(\mathbf{w}^T\phi)^2 - (\mathbf{w}_{MAP}^T)^2]q(\mathbf{w})d\mathbf{w} \\
&= \boldsymbol{\phi}^2 \int ((\mathbf{w}^T)^2 - (\mathbf{w}_{MAP}^T)^2)\mathcal{N}(\mathbf{w}|\mathbf{w}_{MAP}, \mathbf{S}_N)d\mathbf{w} \\
&= \boldsymbol{\phi}^T\mathbf{S}_N\boldsymbol{\phi}
\end{aligned}$$

Here we used

$$\int (\mathbf{x}^2 - \mathbf{m}_x^2)\mathcal{N}(\mathbf{x}|\mathbf{m}_x, \sigma^2)d\mathbf{x} = \sigma^2$$

Chapter 5. Neural Networks

Eq 5.18: (PRML p.234)

$$\frac{\partial E}{\partial a_k} = y_k - t_k$$

Proof :

Using Eq (4.122),

$$\nabla_y \ln p(t|\eta, s) = \sum_{n=1}^{N}\left\{\frac{d}{d\eta_n}\ln g(\eta_n) + \frac{t_n}{s}\right\}\frac{d\eta_n}{dy_n}$$

where $\eta = \psi(y)$

If we look at y_k term only, (k-th data)

$$\nabla_{y_k} \ln p(\overline{\mathbf{t}}|\eta, s) = \left\{\frac{d}{d\eta_n}\ln g(\eta_k) + \frac{t_k}{s}\right\}\frac{d\eta_k}{dy_k} \tag{1}$$

Since y = $\psi(y)$ (from Eq (4.123) & f = 1),

$$\frac{d\eta}{dy} = \frac{d\psi}{dy} = 1$$

$$\text{Eq (1)} = \frac{1}{s}(t_k - y_k)$$

Since s = 1 for regression,

$$\therefore \quad \nabla_{y_k} E = \frac{\partial E}{\partial a_k} = -\frac{\partial}{\partial y_k} \ln p(\bar{\mathbf{t}}|\eta, s) = y_k - t_k$$

Eq 5.47: (PRML p.242)

$$\frac{\partial E_n}{\partial w_{ji}} = (y_{nj} - t_{nj})x_{ni}$$

Proof :

$$E_n = \frac{1}{2} \sum_k (y_{nk} - t_{nk})^2$$

$$(\text{Since } y_{nk} = \sum_i w_{ki} x_{ni})$$

$$= \frac{1}{2} \sum_k (\sum_i w_{ki} x_{ni} - t_{nk})^2$$

$$\frac{\partial E_n}{\partial w_{jl}} = \sum_k \left[\left(\sum_i w_{ki} x_{ni} - t_{nk} \right) \cdot \frac{\partial (\sum_i w_{ki} x_{ni})}{\partial w_{jl}} \right]$$

(Here, $\frac{\partial (\sum_i w_{ki} x_{ni})}{\partial w_{jl}}$ survives only when k = j and i = l.)

$$= \left(\sum_i w_{ji} x_{ni} - t_{nj} \right) \cdot x_{nl}$$

(The first term inside the parenthesis $= y_{nj}$)

$$= (y_{nj} - t_{nj}) x_{nl}$$

Eq 5.56: (PRML p.244)

$$\delta_j = h'(a_j) \sum_k w_{kj} \delta_k$$

Proof :

$$\delta_j \equiv \frac{\partial E_n}{\partial a_j} = \sum_k \frac{\partial E_n}{\partial a_k} \cdot \frac{\partial a_k}{\partial a_j} \tag{5.55}$$

$$\frac{\partial E_n}{\partial a_k} \equiv \delta_k \tag{5.51}$$

$$a_k = \sum_j w_{kj} z_j \tag{5.48}$$

$$= \sum_j w_{kj} h(a_j) \tag{5.49}$$

$$\frac{\partial a_k}{\partial a_j} = w_{kj} \frac{\partial h(a_j)}{\partial a_j} = h'(a_j) w_{kj}$$

$$\Rightarrow \quad \delta_j = \sum_k \delta_k \cdot h'(a_j) w_{kj} = h'(a_j) \sum_k w_{kj} \delta_k$$

Eqs 5.68 & 5.69 : (PRML p.246)

$$\frac{\partial E_n}{\partial w_{ji}} = \frac{E_n(w_{ji} + \epsilon) - E_n(w_{ji})}{\epsilon} + O(\epsilon) \tag{5.68}$$

$$\frac{\partial E_n}{\partial w_{ji}} = \frac{E_n(w_{ji} + \epsilon) - E_n(w_{ji} - \epsilon)}{2\epsilon} + O(\epsilon^2) \tag{5.69}$$

Proof :

Talylor expansion:

$$f(x) = f(a) + \left.\frac{df(x)}{dx}\right|_{x=a} (x-a) + \frac{1}{2} \left.\frac{d^2 f(x)}{dx^2}\right|_{x=a} (x-a)^2 + \cdots \tag{1}$$

$$f(x+a) = f(x+a)\Big|_{x=0} + \left.\frac{df(x+a)}{dx}\right|_{x=0} x + \frac{1}{2} \left.\frac{d^2 f(x+a)}{dx^2}\right|_{x=0} x^2 + \cdots \tag{2}$$

Eq (5.68):

Using Eq (2),

$$
\begin{aligned}
E_n(w_{ji}+\epsilon) &= E_n(w_{ji}) + \left.\frac{\partial E_n(w_{ji}+\epsilon)}{\partial \epsilon}\right|_{\epsilon=0} \epsilon + \frac{1}{2}\left.\frac{\partial^2 E_n(w_{ji}+\epsilon)}{\partial \epsilon^2}\right|_{\epsilon=0} \epsilon^2 + \cdots \\
\frac{E_n(w_{ji}+\epsilon) - E(w_{ji})}{\epsilon} &= \frac{\partial E_n(w_{ji})}{\partial \epsilon} + O(\epsilon) \\
&\text{(Defining } \frac{\partial E_n(w_{ji})}{\partial \epsilon} \equiv \frac{\partial E_n}{\partial w_{ji}}\text{)} \\
\therefore \quad \frac{\partial E_n}{\partial w_{ji}} &= \frac{E_n(w_{ji}+\epsilon) - E_n(w_{ji})}{\epsilon} + O(\epsilon)
\end{aligned}
$$

Eq (5.69):

$$
\begin{aligned}
E_n(w_{ji}+\epsilon) &= E_n(w_{ji}) + \left.\frac{\partial E_n(w_{ji}+\epsilon)}{\partial \epsilon}\right|_{\epsilon=0} \cdot \epsilon + \frac{1}{2}\left.\frac{\partial^2 E_n(w_{ji}+\epsilon)}{\partial \epsilon^2}\right|_{\epsilon=0} \cdot \epsilon^2 \\
&\quad + \frac{1}{6}\frac{\partial^3 E_n(w_{ji}+\epsilon)}{\partial \epsilon^3} \cdot \epsilon^3 + \cdots \\
E_n(w_{ji}-\epsilon) &= E_n(w_{ji}) - \left.\frac{\partial E_n(w_{ji}+\epsilon)}{\partial \epsilon}\right|_{\epsilon=0} \cdot \epsilon + \frac{1}{2}\left.\frac{\partial^2 E_n(w_{ji}+\epsilon)}{\partial \epsilon^2}\right|_{\epsilon=0} \cdot \epsilon^2 \\
&\quad - \frac{1}{6}\frac{\partial^3 E_n(w_{ji}+\epsilon)}{\partial \epsilon^3} \cdot \epsilon^3 + \cdots \\
\frac{E_n(w_{ji}+\epsilon) - E_n(w_{ji}-\epsilon)}{2\epsilon} &= \frac{\partial E_n(w_{ji})}{\partial \epsilon} + \frac{1}{6}\frac{\partial^3 E_n(w_{ji})}{\partial \epsilon^2} + \cdots \\
&= \frac{\partial E_n(w_{ji})}{\partial \epsilon} + O(\epsilon^3) \\
\therefore \quad \frac{\partial E_n}{\partial w_{ji}} &= \frac{E_n(w_{ji}+\epsilon) - E_n(w_{ji}-\epsilon)}{2\epsilon} + O(\epsilon^2)
\end{aligned}
$$

Eq 5.94: (PRML p.254)

$$
\begin{aligned}
\frac{\partial^2 E_n}{\partial w_{ji}^{(1)} \partial w_{j'i'}^{(1)}} &= x_i x_i' h''(a_j) I_{jj'} \sum_k w_{kj'}^{(2)} \delta_k \\
&\quad + x_i x_i' h'(a_{j'}) h'(a_j) \sum_k \sum_{k'} w_{k'j'}^{(2)} w_{kj}^{(2)} M_{kk'}
\end{aligned}
$$

Proof :

$$
\begin{aligned}
\frac{\partial^2 E_n}{\partial w_{ji}^{(1)} \partial w_{j'i}^{(1)}} &= \frac{\partial}{\partial w_{ji}^{(1)}} \left\{ \frac{\partial E_n}{\partial w_{j'i'}^{(1)}} \right\} \\
\frac{\partial E_n}{\partial w_{j'i'}^{(1)}} &= \frac{\partial E_n}{\partial a_{j'}} \cdot \frac{\partial a_{j'}}{\partial w_{j'i'}^{(1)}}
\end{aligned}
$$

Notice that there is no $\sum_{j'}$ in front. This is because you are looking at the destination node j' only, as defined in $w^{(1)}_{j'i'}$. $w^{(1)}_{j'i'}$ relates only to j' and no other nodes.

Since

$$a_{j'} = \sum_{i'} w^{(1)}_{j'i'} x_{i'}, \quad \frac{\partial a_{j'}}{\partial w^{(1)}_{j'i'}} = x_{i'}$$

Therefore,

$$\frac{\partial E_n}{\partial w^{(1)}_{j'i'}} = \frac{\partial E_n}{\partial a_{j'}} \cdot x_{i'}$$

$$\frac{\partial}{\partial w^{(1)}_{ji}} \left(\frac{\partial E_n}{\partial a_{j'}} x_{i'} \right) = x_{i'} \frac{\partial}{\partial w^{(1)}_{ji}} \left(\frac{\partial E_n}{\partial a_{j'}} \right)$$

Now let's calculate $\frac{\partial}{\partial w^{(1)}_{ji}} \left(\frac{\partial E_n}{\partial a_{j'}} \right)$.

$$\frac{\partial E_n}{\partial a_{j'}} = \sum_{k'} \frac{\partial E_n}{\partial a_{k'}} \cdot \frac{\partial a_{k'}}{\partial a_{j'}}$$

Since $a_{k'} = \sum_{j'} w^{(2)}_{k'j'} z_{j'}$ and $z_{j'} = h(a_{j'})$,

$$\begin{aligned}
\frac{\partial a_{k'}}{\partial a_{j'}} &= \frac{\partial}{\partial a_{j'}} \left[\sum_{j'} w^{(2)}_{k'j'} h(a_{j'}) \right] \\
&= w^{(2)}_{k'j'} h'(a_{j'})
\end{aligned}$$

Then,

$$\frac{\partial E_n}{\partial a_{j'}} = \sum_{k'} \frac{\partial E_n}{\partial a_{k'}} \cdot w^{(2)}_{k'j'} h'(a_{j'})$$

$$\begin{aligned}
\frac{\partial}{\partial w^{(1)}_{ji}} \left(\frac{\partial E_n}{\partial a_{j'}} \right) &= \frac{\partial}{\partial w^{(1)}_{ji}} \left[\sum_{k'} \frac{\partial E_n}{\partial a_{k'}} \cdot w^{(2)}_{k'j'} h'(a_{j'}) \right] \\
&= \frac{\partial}{\partial a_j} \left[\sum_{k'} \frac{\partial E_n}{\partial a_{k'}} \cdot w^{(2)}_{k'j'} h'(a_{j'}) \right] \cdot \frac{\partial a_j}{\partial w^{(1)}_{ji}} \\
&\text{(where } \frac{\partial a_j}{\partial w^{(1)}_{ji}} = x_i) \\
&= \sum_{k'} \left[\frac{\partial}{\partial a_j} \left(\frac{\partial E_n}{\partial a_{k'}} \right) \right] w^{(2)}_{k'j'} h'(a_{j'}) x_i \\
&\quad + \sum_{k'} \left(\frac{\partial E_n}{\partial a_{k'}} \right) \cdot \frac{\partial}{\partial a_j} \left[w^{(2)}_{k'j'} h'(a_{j'}) \right] x_i
\end{aligned}$$

The first term,

$$\begin{aligned}
\frac{\partial}{\partial a_j}\left(\frac{\partial E_n}{\partial a_{k'}}\right) &= \sum_k \frac{\partial}{\partial a_k}\left(\frac{\partial E_n}{\partial a_{k'}}\right)\cdot\frac{\partial a_k}{\partial a_j} \\
(\text{Since } \frac{\partial a_k}{\partial a_j} &= \frac{\partial}{\partial a_j}\left[\sum_j w_{kj}^{(2)} h(a_j)\right] = w_{kj}^{(2)} h'(a_j)) \\
&= \sum_k \frac{\partial}{\partial a_k}\left(\frac{\partial E_n}{\partial a_{k'}}\right)\cdot w_{kj}^{(2)} h'(a_j) \\
(\text{where } \frac{\partial}{\partial a_k}\left(\frac{\partial E_n}{\partial a_{k'}}\right) &= M_{kk'})
\end{aligned}$$

The second term,

$$\sum_{k'}\left(\frac{\partial E_n}{\partial a_{k'}}\right)\cdot\frac{\partial}{\partial a_j}\left[w_{k'j'}^{(2)} h'(a_{j'})\right] x_i = \sum_{k'} x_i \frac{\partial E_n}{\partial a_{k'}}\delta_{jj'} w_{k'j'}^{(2)} h''(a_{j'})$$

Therefore,

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

Eq 5.101, 5.102, & 5.103: (PRML p.255)

$$\begin{aligned}
\mathcal{R}\{a_j\} &= \sum_i v_{ji} x_i \\
\mathcal{R}\{z_j\} &= h'(a_j)\mathcal{R}\{a_j\} \\
\mathcal{R}\{y_k\} &= \sum_j w_{kj}\mathcal{R}\{z_j\} \sum_j v_{kj} z_j
\end{aligned}$$

Proof :

$$
\begin{aligned}
\mathcal{R}\{\cdot\} &= v^T\nabla = \sum_i v_{ji}\frac{\partial}{\partial w_{ji}} + \sum_j v_{kj}\frac{\partial}{\partial w_{kj}} \\
\mathcal{R}\{a_j\} &= \sum_i v_{ji}\frac{\partial}{\partial w_{ji}}\left(\sum_{i'} w_{ji'}x_{i'}\right) + 0 \\
&= \sum_i v_{ji}x_i \\
\mathcal{R}\{z_j\} &= \sum_i v_{ji}\frac{\partial}{\partial w_{ji}}\left[h(a_j)\right] + 0 \\
&= \sum_i v_{ji}\frac{\partial}{\partial a_j}\left[h(a_j)\right]\cdot\frac{\partial a_j}{\partial w_{ji}} \\
&= \sum_i v_{ji}h'(a_j)\cdot x_i \\
&= h'(a_j)\mathcal{R}\{a_j\} \\
\mathcal{R}\{y_k\} &= \sum_i v_{ji}\frac{\partial}{\partial w_{ji}}\left(\sum_{j'} w_{kj'}z_{j'}\right) + \sum_j v_{kj}\frac{\partial}{\partial w_{kj}}\left(\sum_{j'} w_{kj'}z_{j'}\right) \\
&= \sum_i v_{ji}\sum_{j'} w_{kj'}\frac{\partial}{\partial w_{ji}}z_{j'} + \sum_j v_{kj}z_j
\end{aligned}
$$

$$
\begin{aligned}
\text{Since the first term } &= \sum_i v_{ji}w_{kj}h'(a_j)x_i = \sum_i w_{kj}v_{ji}h'(a_j)x_i \\
&= \sum_j w_{kj}h'(a_j)\sum_i v_{j'i}x_i \\
&= \sum_j w_{kj}\mathcal{R}\{z_j\}
\end{aligned}
$$

$$
\therefore\quad \mathcal{R}\{y_k\} = \sum_j w_{kj}\mathcal{R}\{z_j\}\sum_j v_{kj}z_j
$$

Eq 5.107: (PRML p.255)

$$
\mathcal{R}\{\delta_j\} = h''(a_j)\mathcal{R}\{a_j\}\sum_k w_{kj}\delta_k + h'(a_j)\sum_k v_{kj}\delta_k + h'(a_j)\sum_k w_{kj}\mathcal{R}\{\delta_k\}
$$

Proof :

$$
\delta_j = h'(a_j)\sum_k w_{kj}\delta_k
$$

$$\begin{aligned}
\mathcal{R}\{\delta_j\} &= \sum_i v_{ji}\frac{\partial \delta_j}{\partial w_{ji}} + \sum_{j'} v_{kj'}\frac{\partial \delta_j}{\partial w_{jk'}} \\
&= \sum_i v_{ji}\frac{\partial}{\partial w_{ji}}\left[h'(a_j)\sum_k w_{kj}\delta_k\right] + \sum_{j'} v_{kj'}\frac{\partial}{\partial w_{kj'}}\left[h'(a_j)\sum_k w_{kj}\delta_k\right] \qquad (1)
\end{aligned}$$

First term in Eq (1):

$$\begin{aligned}
\sum_i v_{ji}\frac{\partial}{\partial w_{ji}}\left[h'(a_j)\sum_k w_{kj}\delta_k\right] &= \sum_i v_{ji}\frac{\partial}{\partial a_j}\left[h'(a_j)\sum_k w_{kj}\delta_k\right]\frac{\partial a_j}{\partial w_{ji}} \\
&= \sum_i v_{ji}h''(a_j)\sum_k w_{kj}\delta_k x_i \\
&= h''(a_j)\sum_k w_{kj}\delta_k \cdot \sum_i v_{ji}x_i \\
\text{(Since } &\sum_i v_{ji}x_i = \mathcal{R}\{a_j\}\text{)} \\
&= h''(a_j)\sum_k w_{kj}\delta_k \cdot \mathcal{R}\{a_j\}
\end{aligned}$$

Second term in Eq (1):

$$\begin{aligned}
&\sum_{j'} v_{kj'}\frac{\partial}{\partial w_{kj'}}\left[h'(a_j)\sum_k w_{kj}\delta_k\right] \\
&= \sum_k h'(a_j)\delta_k \sum_{j'}\frac{\partial w_{kj}}{\partial w_{kj'}}v_{kj'} + \sum_k h'(a_j)w_{kj}\sum_{j'}\frac{\partial \delta_k}{\partial w_{kj'}}v_{kj'} \qquad (2)
\end{aligned}$$

First term in Eq (2):

$$\begin{aligned}
\sum_k h'(a_j)\delta_k \sum_{j'}\frac{\partial w_{kj}}{\partial w_{kj'}}v_{kj'} &= \sum_k h'(a_j)\delta_k v_{kj} \\
&= h'(a_j)\sum_k \delta_k v_{kj}
\end{aligned}$$

Second term in Eq (2):

$$\begin{aligned}
\sum_k h'(a_j)w_{kj}\sum_{j'}\frac{\partial \delta_k}{\partial w_{kj'}}v_{kj'} &= h'(a_j)\sum_k w_{kj}\sum_{j'} v_{kj'}\frac{\partial}{\partial w_{kj'}}\delta_k \\
\text{(Since } &\sum_{j'} v_{kj'}\frac{\partial}{\partial w_{jk'}}\delta_k = \mathcal{R}\{\delta_k\}\text{)} \\
&= h'(a_j)\sum_k w_{kj}\mathcal{R}\{\delta_k\}
\end{aligned}$$

$$\Rightarrow \quad \sum_{j'} v_{kj'}\frac{\partial}{\partial w_{kj'}}\left[h'(a_j)\sum_k w_{kj}\delta_k\right] = h'(a_j)\sum_k \delta_k v_{kj} + h'(a_j)\sum_k w_{kj}\mathcal{R}\{\delta_k\}$$

$$\therefore \quad \mathcal{R}\{\delta_j\} = h''(a_j)\sum_k w_{kj}\delta_k\mathcal{R}\{a_j\} + h'(a_j)\sum_k \delta_k v_{kj} + h'(a_j)\sum_k w_{kj}\mathcal{R}\{\delta_k\}$$

PRML p.266 :

$$y(\mathbf{s}(\mathbf{x},\boldsymbol{\xi})) = y(\mathbf{x}) + \boldsymbol{\xi}\boldsymbol{\tau}^T\nabla y(\mathbf{x}) + \frac{\boldsymbol{\xi}^2}{2}\left[(\boldsymbol{\tau}')^T\nabla y(\mathbf{x}) + \boldsymbol{\tau}^T\nabla\nabla y(\mathbf{x})\boldsymbol{\tau}\right] + O(\boldsymbol{\xi}^3)$$

Proof :

$$y(\mathbf{s}(\mathbf{x},\boldsymbol{\xi})) = y(\mathbf{s}(\mathbf{x},0)) + \boldsymbol{\xi}\left.\frac{\partial y}{\partial\boldsymbol{\xi}}\right|_{\boldsymbol{\xi}=0} + \frac{1}{2}\boldsymbol{\xi}^2\left.\frac{\partial^2 y}{\partial\boldsymbol{\xi}^2}\right|_{\boldsymbol{\xi}=0} + O(\boldsymbol{\xi}^3) \tag{1}$$

$$\left.\frac{\partial y}{\partial\boldsymbol{\xi}}\right|_{\boldsymbol{\xi}=0} = \left.\frac{\partial y}{\partial\mathbf{s}}\frac{\partial\mathbf{s}}{\partial\boldsymbol{\xi}}\right|_{\boldsymbol{\xi}=0} = \frac{\partial y}{\partial\mathbf{s}}\boldsymbol{\tau}^T \tag{2}$$

$$\left.\frac{\partial^2 y}{\partial\boldsymbol{\xi}^2}\right|_{\boldsymbol{\xi}=0} = \frac{\partial}{\partial\boldsymbol{\xi}}\left(\frac{\partial y}{\partial\mathbf{s}}\frac{\partial\mathbf{s}}{\partial\boldsymbol{\xi}}\right)\bigg|_{\boldsymbol{\xi}=0} = \frac{\partial}{\partial\boldsymbol{\xi}}\left(\frac{\partial y}{\partial\mathbf{s}}\right)\left.\frac{\partial\mathbf{s}}{\partial\boldsymbol{\xi}}\right|_{\boldsymbol{\xi}=0} + \left.\frac{\partial y}{\partial\mathbf{s}}\frac{\partial^2\mathbf{s}}{\partial\boldsymbol{\xi}^2}\right|_{\boldsymbol{\xi}=0} \tag{3}$$

First term in Eq (3):

$$\frac{\partial}{\partial\boldsymbol{\xi}}\left(\frac{\partial y}{\partial\mathbf{s}}\right)\left.\frac{\partial\mathbf{s}}{\partial\boldsymbol{\xi}}\right|_{\boldsymbol{\xi}=0} = \frac{\partial\mathbf{s}}{\partial\boldsymbol{\xi}}\frac{\partial}{\partial\mathbf{s}}\left(\frac{\partial y}{\partial\mathbf{s}}\right)\left.\frac{\partial\mathbf{s}}{\partial\boldsymbol{\xi}}\right|_{\boldsymbol{\xi}=0} \tag{4}$$

$$\text{where} \quad \frac{\partial\mathbf{s}}{\partial\boldsymbol{\xi}} = \boldsymbol{\tau}^T, \quad \frac{\partial\mathbf{s}}{\partial\boldsymbol{\xi}} = \boldsymbol{\tau} \tag{5}$$

Second term in Eq (3):

$$\left.\frac{\partial^2\mathbf{s}}{\partial\boldsymbol{\xi}^2}\right|_{\boldsymbol{\xi}=0} = (\boldsymbol{\tau}')^T \tag{6}$$

From Eqs (4) $\sim$ (6), Eq (3) becomes

$$\left.\frac{\partial^2 y}{\partial\boldsymbol{\xi}^2}\right|_{\boldsymbol{\xi}=0} = \boldsymbol{\tau}\frac{\partial^2 y}{\partial\mathbf{s}^2}\boldsymbol{\tau}^T + \frac{\partial y}{\partial\mathbf{s}}(\boldsymbol{\tau}')^T \tag{7}$$

Plugging Eqs (2) and (7) into Eq (1) gives

$$\therefore y(\mathbf{s}(\mathbf{x},\boldsymbol{\xi})) = y(\mathbf{x}) + \boldsymbol{\xi}\boldsymbol{\tau}^T\frac{\partial y}{\partial\mathbf{s}} + \frac{1}{2}\boldsymbol{\xi}^2\left[(\boldsymbol{\tau}')^T\frac{\partial y}{\partial\mathbf{s}} + \boldsymbol{\tau}^T\frac{\partial^2 y}{\partial\mathbf{s}^2}\boldsymbol{\tau}\right] + O(\boldsymbol{\xi}^3)$$

This can be easily generalized to multi-dimension case.

PRML p.270:

"Recall that the simple weight decay regularizer, given in (5.112), can be viewed as the negative log of a Gaussian prior distribution over the weights. ”

Proof :

Eq (5.112):

$$\widetilde{E}(\mathbf{w}) = E(\mathbf{w}) + \frac{\lambda}{2}\mathbf{w}^{\mathbf{w}}$$

Eq (1.62):

$$\ln p(\bar{\mathbf{t}}|\mathbf{x}, \mathbf{w}, \beta) = -\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi)$$

Eq (1.65):

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|0, \alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2}\exp\left\{-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}\right\}$$

Eq (1.66):

$$p(\mathbf{w}|\mathbf{x}, \bar{\mathbf{t}}, \alpha, \beta) \propto p(\bar{\mathbf{t}}|\mathbf{x}, \mathbf{w}, \beta)\, p(\mathbf{w}|\alpha)$$

Taking the negative logarithm of Eq (1.66) gives,

$$\begin{aligned}
&-\ln[(\mathbf{w}|\mathbf{x}, \bar{\mathbf{t}}, \alpha, \beta) \propto -\ln p(\bar{\mathbf{t}}|\mathbf{x}, \mathbf{w}, \beta) - \ln p(\mathbf{w}|\alpha) \\
&= \frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 - \frac{N}{2}\ln\beta + \frac{N}{2}\ln(2\pi) + \frac{\alpha}{2}\mathbf{w}^T\mathbf{w} - \frac{M+1}{2}\ln\alpha + \frac{M+1}{2}\ln(2\pi) \\
&= \frac{\beta}{2}\sum_{n=1}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\alpha}{2}\mathbf{w}^T\mathbf{w} + C
\end{aligned}$$

$\therefore\ \frac{\lambda}{2}\mathbf{w}^T\mathbf{w}$ in Eq (5.112) comes from the negative logarithm of a Gaussian distribution.

Eq 5.164: (PRML p.278)

$$p(\mathbf{w}|D, \alpha, \beta) \propto p(\mathbf{w}|\alpha)\, p(D|\mathbf{w}, \beta)$$

Proof :

$$
\begin{aligned}
p(\mathbf{w}|D,\alpha,\beta) &= \frac{p(\mathbf{w}\cap(D\cap\alpha\cap\beta))}{p(D\cap\alpha\cap\beta)} \\
&= \frac{p(D\cap(\mathbf{w}\cap\alpha\cap\beta))}{p(D\cap\alpha\cap\beta)} \\
&= \frac{p(D|\mathbf{w},\alpha,\beta)\,p(\mathbf{w},\alpha,\beta)}{p(D,\alpha,\beta)} \\
&\text{(Since } p(\mathbf{w},\alpha,\beta) = p(\mathbf{w}|\alpha,\beta)\,p(\alpha,\beta)) \\
&= \frac{p(D|\mathbf{w},\alpha,\beta)\,p(\mathbf{w}|\alpha,\beta)\,p(\alpha,\beta}{p(D,\alpha,\beta)}
\end{aligned}
$$

From Eq (5.163)

$$
\begin{aligned}
p(D|\mathbf{w},\beta) &= \prod_{n=1}^{N}\mathcal{N}(t_n|y(x_n,\mathbf{w}),\beta^{-1}) \\
\Rightarrow \quad & p(D|\mathbf{w},\beta) \perp \alpha
\end{aligned}
$$

From Eq (5.162)

$$
\begin{aligned}
p(\mathbf{w}|\alpha) &= \mathcal{N}(\mathbf{w}|0,\alpha^{-1}I),\beta^{-1}) \\
\Rightarrow \quad & p(\mathbf{w}|\alpha) \perp \beta
\end{aligned}
$$

$$
\therefore \quad p(\mathbf{w}|D,\alpha,\beta) = p(\mathbf{w}|\alpha)\,p(D|\mathbf{w},\beta)\,\frac{p(\alpha,\beta)}{p(D,\alpha,\beta)}
$$

Eq 5.181: (PRML p.282)

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

Proof :

Eq (5.181) can be derived from Eq (4.89),

$$
p(\bar{\mathbf{t}}|\mathbf{w}) = \prod_{n=1}^{N} y_n^{t_n}\{1-y_n\}^{1-t_n}
$$

where $\bar{\mathbf{t}} = (t_1,\ldots,t_N)^T$ and $y_n = p(C_1|\boldsymbol{\phi}_n)$.

Replacing $\bar{\mathbf{t}}$ with D, and taking log of Eq (4.89) will give us Eq (5.181).

Eq 5.188: (PRML p.283)

$$\sigma_a^2(\mathbf{x}) = \mathbf{b}^T(\mathbf{x})\mathbf{A}^{-1}\mathbf{b}(\mathbf{x})$$

Proof :

$$p(a|\mathbf{x}, D) = \int \delta(a - a_{MAP}(\mathbf{x}) - \mathbf{b}^T(\mathbf{x})(\mathbf{w} - \mathbf{w}_{MAP}))q(\mathbf{w}|D)d\mathbf{w} \tag{5.187}$$

$$\begin{aligned}
\mu_a &= \int p(a|\mathbf{x}, D)\, a\, da \\
&= \int \left[a_{MAP}(\mathbf{x}) + \mathbf{b}^T(\mathbf{x})(\mathbf{w} - \mathbf{w}_{MAP})\right]\; q(\mathbf{w}|D)\, d\mathbf{w} \\
&= a_{MAP}(\mathbf{x}) + \mathbf{b}^T(\mathbf{x})(\mathbf{w}_{MAP} - \mathbf{w}_{MAP}) \\
&= a_{MAP}(\mathbf{x}, \mathbf{w}_{MAP})
\end{aligned}$$

$$\begin{aligned}
\sigma_a^2 &= \int p(a|\mathbf{x}, D)\, \{a^2 - \mathbb{E}[a]^2\} da \\
&= \int \left\{[a_{MAP}(\mathbf{x}) + \mathbf{b}^T(\mathbf{x})(\mathbf{w} - \mathbf{w}_{MAP})]^2 - a_{MAP}^2\right\} q(\mathbf{w}|D)d\mathbf{w} \\
&= \int \left[2a_{MAP}(x)\mathbf{b}^T(\mathbf{x})(\mathbf{w} - \mathbf{w}_{MAP}) + \mathbf{b}^T(\mathbf{w} - \mathbf{w}_{MAP})^2\mathbf{b}\right] q(\mathbf{w}|D)d\mathbf{w} \\
&\text{(The first term in the integrand becomes zero.)} \\
&= \mathbf{b}^T \left[\int (\mathbf{w} - \mathbf{w}_{MAP})^2 q(\mathbf{w}|D)d\mathbf{w}\right] \mathbf{b}
\end{aligned}$$

Since $q(\mathbf{w}|D) = \mathcal{N}(\mathbf{w}|\mathbf{w}_{MAP}, A^{-1})$

$$\therefore\;\; \sigma_a^2 = \mathbf{b}^T(\mathbf{x})A^{-1}\mathbf{b}(\mathbf{x})$$

Remember

$$\sigma^2 = \int p(x)(x-\mu)^2 dx = \int p(x)x^2 ds - \mu^2 = \int p(x)[x^2 - \mu^2]dx$$

Chapter 6. Kernel Methods

Eq 6.3: (PRML p.293)

$$\sum_{n=1}^{N} a_n \boldsymbol{\phi}(\mathbf{x}_n) = \mathbf{\Phi}^T \mathbf{a}$$

where

$$\mathbf{\Phi} = \begin{bmatrix} \boldsymbol{\phi}(\mathbf{x}_1)^T \\ \boldsymbol{\phi}(\mathbf{x}_2)^T \\ \vdots \\ \boldsymbol{\phi}(\mathbf{x}_N)^T \end{bmatrix} \qquad \text{and} \quad \mathbf{a} = (a_1, \cdots, a_N)^T$$

Proof :

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

$$\begin{aligned}
\mathbf{\Phi}^T \mathbf{a} &= \begin{bmatrix} \phi_1(\mathbf{x}_1) & \phi_1(\mathbf{x}_2) & \cdots & \phi_1(\mathbf{x}_N) \\ \phi_2(\mathbf{x}_1) & \phi_2(\mathbf{x}_2) & \cdots & \phi_2(\mathbf{x}_N) \\ \vdots & & & \\ \phi_M(\mathbf{x}_1) & \phi_M(\mathbf{x}_2) & \cdots & \phi_M(\mathbf{x}_N) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{bmatrix} \\
&= \begin{bmatrix} a_1\phi_1(\mathbf{x}_1) + a_2\phi_1(\mathbf{x}_2) + \cdots + a_N\phi_1(\mathbf{x}_N) \\ a_1\phi_2(\mathbf{x}_1) + a_2\phi_2(\mathbf{x}_2) + \cdots + a_N\phi_2(\mathbf{x}_N) \\ \vdots \\ a_1\phi_M(\mathbf{x}_1) + a_2\phi_M(\mathbf{x}_2) + \cdots + a_N\phi_M(\mathbf{x}_N) \end{bmatrix} \\
&= a_1 \begin{bmatrix} \phi_1(\mathbf{x}_1) \\ \phi_2(\mathbf{x}_1) \\ \vdots \\ \phi_M(\mathbf{x}_1) \end{bmatrix} + a_2 \begin{bmatrix} \phi_1(\mathbf{x}_2) \\ \phi_2(\mathbf{x}_2) \\ \vdots \\ \phi_M(\mathbf{x}_2) \end{bmatrix} \cdots + a_N \begin{bmatrix} \phi_1(\mathbf{x}_N) \\ \phi_2(\mathbf{x}_N) \\ \vdots \\ \phi_M(\mathbf{x}_N) \end{bmatrix}
\end{aligned}$$

$$\begin{aligned}
&= a_1\boldsymbol{\phi}(\mathbf{x}_1) + a_2\boldsymbol{\phi}(\mathbf{x}_2) + \cdots + a_N\boldsymbol{\phi}(\mathbf{x}_N) \\
&= \sum_{n=1}^{N} a_n \boldsymbol{\phi}(\mathbf{x}_n)
\end{aligned}$$

Eq 6.45: (PRML p.302)

$$y(\mathbf{x}) = \frac{\sum_n g(\mathbf{x} - \mathbf{x}_n) t_n}{\sum_m g(\mathbf{x} - \mathbf{x}_m)}$$
$$\text{where } g(\mathbf{x}) = \int_{-\infty}^{\infty} f(\mathbf{x}, t)\, dt$$

Proof :

Starting from Eq (6.43),

$$y(\mathbf{x}) = \frac{\sum_n \int t f(\mathbf{x} - \mathbf{x}_n, t - t_n) dt}{\sum_m \int f(\mathbf{x} - \mathbf{x}_m, t - t_m) dt}$$

By changing $t - t_n = p, dt = dp$ and $t = p + t_n$, the numerator becomes

$$\begin{aligned}
\int t f(\mathbf{x} - \mathbf{x}_n, t - t_n) dt &= \int (p + t_n) f(\mathbf{x} - \mathbf{x}_n, p) dp \\
&= \int p\, f(\mathbf{x} - \mathbf{x}_n, p) dp + t_n \int f(\mathbf{x} - \mathbf{x}_n, p) dp \\
&\text{(from Eq (6.44), the first term} = 0) \\
&= t_n \int f(\mathbf{x} - \mathbf{x}_n, t) dt \\
&= t_n\, g(\mathbf{x} - \mathbf{x}_n)
\end{aligned}$$

The denominator is,

$$\begin{aligned}
\int f(\mathbf{x} - \mathbf{x}_m, t - t_m) dt &= \int f(\mathbf{x} - \mathbf{x}_m, p) dp \\
&= g(\mathbf{x} - \mathbf{x}_m)
\end{aligned}$$

$$\therefore \quad y(\mathbf{x}) = \frac{\sum_n g(\mathbf{x} - \mathbf{x}_n) t_n}{\sum_m g(\mathbf{x} - \mathbf{x}_m)}$$

Eq 6.66 & 6.67: (PRML p.308)

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

Proof :

From Eqs (2.81) and (2.82)

$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b) \tag{2.81}$$

$$\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba} \tag{2.82}$$

The above equations are based on

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix} \tag{1}$$

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \tag{2}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix} \tag{3}$$

where b is a condition and a is what we are looking for.

Eq (6.65) is in the reversed order for the condition and result,

$$\mathbf{C}_{N+1} = \begin{pmatrix} \mathbf{C}_N & \mathbf{k} \\ \mathbf{k}^T & c \end{pmatrix} \tag{6.65}$$

Here $\mathbf{C}_N$ is the condition.

Conditional Gaussian distributions for the reverse case like this can be derived using Eqs (2.75) to (2.82), where the condition is on $\mathbf{x}_b$.

$$\begin{aligned} \boldsymbol{\mu}_{b|a} &= \boldsymbol{\mu}_b - \boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}(\mathbf{x}_a - \boldsymbol{\mu}_a) \\ &= \boldsymbol{\mu}_b + \boldsymbol{\Sigma}_{ba}\boldsymbol{\Sigma}_{aa}^{-1}(\mathbf{x}_a - \boldsymbol{\mu}_a) \qquad (4) \end{aligned}$$

$$\boldsymbol{\Sigma}_{b|a} = \boldsymbol{\Sigma}_{bb} - \boldsymbol{\Sigma}_{ba}\boldsymbol{\Sigma}_{aa}^{-1}\boldsymbol{\Sigma}_{ab} \tag{5}$$

Using Eqs (4) and (5), and identifying the matching elements as follows, (back to their original place after deriving (4) and (5))

$$\begin{pmatrix} \mathbf{C}_N & \mathbf{k} \\ \mathbf{k}^T & c \end{pmatrix} \Longleftrightarrow \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}$$

$$\begin{pmatrix} \bar{\mathbf{t}} \\ t_{N+1} \end{pmatrix} \Longleftrightarrow \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix}$$

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} \Longleftrightarrow \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}$$

$$\begin{aligned} \therefore \quad \boldsymbol{\mu}_{N+1|1\sim N} &= 0 + \mathbf{k}^T \mathbf{C}_N^{-1}(\bar{\mathbf{t}} - 0) \\ &= \mathbf{k}^T \mathbf{C}_N^{-1} \bar{\mathbf{t}} \\ \therefore \quad \boldsymbol{\Sigma}_{N+1|1\sim N} &= c - \mathbf{k}^T \mathbf{C}_N^{-1} k \end{aligned}$$

Eq 6.69: (PRML p.311)

$$\ln p(\bar{\mathbf{t}}|\boldsymbol{\theta}) = -\frac{1}{2}\ln|\mathbf{C}_N| - \frac{1}{2}\bar{\mathbf{t}}^T \mathbf{C}_N^{-1}\bar{\mathbf{t}} - \frac{N}{2}\ln(2\pi)$$

Proof :

Eq (2.43):

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

where D is the dimension of $\mathbf{x}$ and $\boldsymbol{\Sigma}$.

Eq (6.61):

$$p(\bar{\mathbf{t}}) = \mathcal{N}(\bar{\mathbf{t}}|0, \mathbf{C}) \tag{2}$$

There is a fundamental difference between (1) and (2).

In (1), $\mathbf{x}$ is a vector in feature space. $\rightarrow$ D-dimension.

In (2), $\bar{\mathbf{t}}$ is a vector in number of training data set. $\rightarrow$ N-dimension.

In (1), $\boldsymbol{\Sigma}$ is a covariance matrix related to the intrinsic error in measurement, ϵ.

$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon$$

In (2), $\mathbf{C}$ reflects two sources of Gaussian randomness; that associated with ϵ and that associated with $y(\mathbf{x})$.

$$\mathbf{C}_{nm} = \mathbf{C}(\mathbf{x}_n, \mathbf{x}_m) = k(\mathbf{x}_n, \mathbf{x}_m) + \beta^{-1}\delta_{nm} \tag{6.62}$$

One widely used kernel function for Gaussian process regression is

$$k(\mathbf{x}_n, \mathbf{x}_m) = \theta_0 \exp\left\{-\frac{\theta_1}{2}\|\mathbf{x}_n - \mathbf{x}_m\|^2\right\} + \theta_2 + \theta_3 \mathbf{x}_n^T \mathbf{x}_m \tag{6.63}$$

$$\mathbf{C} = \begin{bmatrix} C(\mathbf{x}_1, \mathbf{x}_1) & C(\mathbf{x}_1, \mathbf{x}_2) & \cdots & C(\mathbf{x}_1, \mathbf{x}_N) \\ C(\mathbf{x}_2, \mathbf{x}_1) & C(\mathbf{x}_2, \mathbf{x}_2) & \cdots & C(\mathbf{x}_2, \mathbf{x}_N) \\ \vdots & & & \\ C(\mathbf{x}_N, \mathbf{x}_1) & C(\mathbf{x}_N, \mathbf{x}_2) & \cdots & C(\mathbf{x}_N, \mathbf{x}_N) \end{bmatrix}$$

In (1), $\mathbf{x} = (x_1, x_2, \ldots, x_D)$

In (2), $\bar{\mathbf{t}} = (t_1, t_2, \ldots, t_N)$ $\leftarrow$ considered as an N-dim vector.

So we are ready to write down a Gaussian equation similar to (1) and (2).

$$\mathcal{N}(\bar{\mathbf{t}}|0, \mathbf{C}) = \frac{1}{(2\pi)^{N/2}} \frac{1}{|\mathbf{C}|^{1/2}} \exp\left\{-\frac{1}{2}\bar{\mathbf{t}}^T \mathbf{C}^{-1} \bar{\mathbf{t}}\right\}$$

$p(\bar{\mathbf{t}})$ is not marginalized over $\boldsymbol{\theta}$, since $\mathbf{C}$ depends on $\boldsymbol{\theta}$.

$\Rightarrow \quad p(\bar{\mathbf{t}}) = p(\bar{\mathbf{t}}|\boldsymbol{\theta})$

$$\therefore \quad \ln p(\bar{\mathbf{t}}|\boldsymbol{\theta}) = -\frac{1}{2}\ln|\mathbf{C}_N| - \frac{1}{2}\bar{\mathbf{t}}^T \mathbf{C}_N^{-1} \bar{\mathbf{t}} - \frac{N}{2}\ln(2\pi)$$

Eq 6.79: (PRML p.316)

$$p(\bar{\mathbf{t}}_N|\mathbf{a}_N) = \prod_{n=1}^{N} \sigma(a_n)^{t_n}(1-\sigma(a_n))^{1-t_n} = \prod_{n=1}^{N} e^{a_n t_n}\sigma(-a_n)$$

Proof :

$$\sigma(a_n) = \frac{1}{1+e^{-a_n}}$$

$$
\begin{aligned}
\sigma(a_n)^{t_n}\left[1-\sigma(a_n)\right]^{1-t_n} &= \left[\frac{1}{1+e^{-a_n}}\right]^{t_n}\left[1-\frac{1}{1+e^{-a_n}}\right]^{(1-t_n)} \\
&= \left[\frac{1}{1+e^{-a_n}}\right]^{t_n}\left[\frac{e^{-a_n}}{1+e^{-a_n}}\right]^{(1-t_n)} \\
&= \left[\frac{e^{a_n}}{1+e^{a_n}}\right]^{t_n}\left[\frac{1}{1+e^{a_n}}\right]^{(1-t_n)} \\
&= e^{a_n t_n}\left(e^{a_n}+1\right)^{-t_n}\left(e^{a_n}+1\right)^{t_n-1} \\
&= e^{a_n t_n}\left(e^{a_n}+1\right)^{-1} \\
&= e^{a_n t_n}\,\sigma(-a_n)
\end{aligned}
$$

$$
\therefore\;\; p(\overline{\mathbf{t}}_N|\mathbf{a}_N) = \prod_{n=1}^{N} e^{a_n t_n}\sigma(-a_n)
$$

Eq 6.90: (PRML p.317)

$$
\begin{aligned}
\ln p(\overline{\mathbf{t}}_N|\boldsymbol{\theta}) &= \Psi(\mathbf{a}_N^*) - \frac{1}{2}\ln\left|\mathbf{W}_N + \mathbf{C}_N^{-1}\right| + \frac{N}{2}\ln(2\pi) \\
\text{where } \Psi(\mathbf{a}_N^*) &= \ln p(\mathbf{a}_N^*|\boldsymbol{\theta}) + \ln p(\overline{\mathbf{t}}_N|\mathbf{a}_N^*)
\end{aligned}
$$

Proof :

$$
p(\overline{\mathbf{t}}_N|\boldsymbol{\theta}) = \int p(\overline{\mathbf{t}}_N|\mathbf{a}_N)\,p(\mathbf{a}_N|\boldsymbol{\theta})d\mathbf{a}_N \tag{6.89}
$$

Let's call the integrand as $f(\mathbf{a}_N) = p(\overline{\mathbf{t}}_N|\mathbf{a}_N)\,p(\mathbf{a}_N|\boldsymbol{\theta})$

$$
\begin{aligned}
\ln f(\mathbf{a}_N) &\simeq \ln f(\mathbf{a}_N^*) - \frac{1}{2}(\mathbf{a}_N - \mathbf{a}_N^*)^T\mathbf{A}(\mathbf{a}_N - \mathbf{a}_N^*) \\
\text{where } \mathbf{A} &= -\nabla\nabla\ln f(\mathbf{a}_N)\big|_{a_N=\mathbf{a}_N^*} \\
\Rightarrow f(\mathbf{a}_N) &\simeq f(\mathbf{a}_N^*)\exp\left\{-\frac{1}{2}(\mathbf{a}_N - \mathbf{a}_N^*)^T\mathbf{A}(\mathbf{a}_N - \mathbf{a}_N^*)\right\}
\end{aligned}
$$

Utilizing this Laplace approximation,

$$
\begin{aligned}
p(\overline{\mathbf{t}}_N|\theta) &= \int f(\mathbf{a}_N)d\mathbf{a}_N \\
&= f(\mathbf{a}_N^*)\int \exp\left\{-\frac{1}{2}(\mathbf{a}_N - \mathbf{a}_N^*)^T\mathbf{A}(\mathbf{a}_N - \mathbf{a}_N^*)\right\}d\mathbf{a}_N \\
&= f(\mathbf{a}_N^*)\,(2\pi)^{N/2}\frac{1}{|\mathbf{A}|^{1/2}}
\end{aligned}
$$

$$f(\mathbf{a}_N^*) = p(\overline{\mathbf{t}}_N|\mathbf{a}_N^*)\, p(\mathbf{a}_N^*|\boldsymbol{\theta}) \quad \Rightarrow \quad \Psi(\mathbf{a}_N^*) = \ln f(\mathbf{a}_N^*)$$

And to find $\mathbf{A}$,

$$\ln f(\mathbf{a}_N) = \ln\left[p(\overline{\mathbf{t}}_N|\mathbf{a}_N)\, p(\mathbf{a}_N|\boldsymbol{\theta})\right]$$

This is identical to $\Psi(\mathbf{a}_N)$ in Eq (6.80).

$$\Psi(\mathbf{a}_N) = \ln\left[p(\overline{\mathbf{t}}_N|\mathbf{a}_N)\, p(\mathbf{a}_N)\right] \tag{6.80}$$

We can see that $p(\mathbf{a}_N|\boldsymbol{\theta}) = p(\mathbf{a}_N)$.

Therefore,

$$\begin{aligned}\mathbf{A} &= -\nabla\nabla\Psi(\mathbf{a}_N)\big|_{\mathbf{a}_N=\mathbf{a}_N^*} \\ &= \mathbf{W}_N + \mathbf{C}_N^{-1} \end{aligned} \tag{6.82}$$

$$\begin{aligned}\therefore\quad \ln p(\overline{\mathbf{t}}_N|\boldsymbol{\theta}) &= \ln f(\mathbf{a}_N^*) - \frac{1}{2}\ln\left|\mathbf{W}_N + \mathbf{C}_N^{-1}\right| + \frac{N}{2}\ln(2\pi) \\ &= \Psi(\mathbf{a}_N^*) - \frac{1}{2}\ln\left|\mathbf{W}_N + \mathbf{C}_N^{-1}\right| + \frac{N}{2}\ln(2\pi) \\ \text{where } \Psi(\mathbf{a}_N^*) &= \ln p(\mathbf{a}_N^*|\boldsymbol{\theta}) + \ln p(\overline{\mathbf{t}}_N|\mathbf{a}_N^*)\end{aligned}$$

Eq 6.91: (PRML p.318)

$$\begin{aligned}\frac{\partial \ln p(\overline{\mathbf{t}}_N|\boldsymbol{\theta})}{\partial\theta_j} &= \frac{1}{2}(\mathbf{a}_N^*)^T\mathbf{C}_N^{-1}\frac{\partial\mathbf{C}_N}{\partial\theta_j}\mathbf{C}_N^{-1}\mathbf{a}_N^* \\ &\quad - \frac{1}{2}\mathrm{Tr}\left[(\mathbf{I}+\mathbf{C}_N\mathbf{W}_N)^{-1}\mathbf{W}_N\frac{\partial\mathbf{C}_N}{\partial\theta_j}\right]\end{aligned}$$

Typo in the text: the second term should be,

$$-\frac{1}{2}\mathrm{Tr}\left[(\mathbf{I}+\mathbf{C}_N\mathbf{W}_N)^{-1}\mathbf{C}_N\frac{\partial\mathbf{C}_N}{\partial\theta_j}\right]$$

Proof :

Eq (6.90):

$$\ln p(\overline{\mathbf{t}}_N|\boldsymbol{\theta}) = \Psi(\mathbf{a}_N^*) - \frac{1}{2}\ln\left|\mathbf{W}_N + \mathbf{C}_N^{-1}\right| + \frac{N}{2}\ln(2\pi)$$

To calculate the derivative on Eq (6.90), we need to calculate the following.

Using Eqs (C.21) and (C.22),

$$\frac{\partial\ln\left|\mathbf{W}_N+\mathbf{C}_N^{-1}\right|}{\partial\theta_j} = \mathrm{Tr}\left[(\mathbf{W}_N+\mathbf{C}_N^{-1})^{-1}\frac{\partial(\mathbf{W}_N+\mathbf{C}_N^{-1})}{\partial\theta_j}\right]$$

$$\frac{\partial(\mathbf{W}_N + \mathbf{C}_N^{-1})}{\partial\theta_j} = \frac{\partial\mathbf{W}_N}{\partial\theta_j} + \frac{\partial\mathbf{C}_N^{-1}}{\partial\theta_j}$$

$$\frac{\partial\mathbf{W}_N}{\partial\theta_j} = 0 \quad (\because \mathbf{W}_N = \sigma(\mathbf{a}_n)(1-\sigma(\mathbf{a}_n)))$$

$$\frac{\partial\mathbf{C}_N^{-1}}{\partial\theta_j} = -\mathbf{C}_N^{-2}\frac{\partial\mathbf{C}_N}{\partial\theta_j}$$

$$\Rightarrow \quad \frac{\partial \ln\left|\mathbf{W}_N + \mathbf{C}_N^{-1}\right|}{\partial\theta_j} = \mathrm{Tr}\left[(\mathbf{W}_N + \mathbf{C}_N^{-1})^{-1}(-\mathbf{C}_N^{-2})\frac{\partial\mathbf{C}_N}{\partial\theta_j}\right]$$
$$= -\mathrm{Tr}\left[(\mathbf{I} + \mathbf{C}_N\mathbf{W}_N)^{-1}\mathbf{C}_N^{-1}\frac{\partial\mathbf{C}_N}{\partial\theta_j}\right]$$

This proves the second term in Eq (6.91). The first term can be easily calculated from Eq (6.80).

Eq 6.92: (PRML p.318)

$$\frac{\partial \ln\left|\mathbf{W}_N + \mathbf{C}_N^{-1}\right|}{\partial a_n^*} = \mathrm{Tr}\left[(\mathbf{I} + \mathbf{C}_N\mathbf{W}_N)^{-1}\mathbf{C}_N\sigma_n^*(1-\sigma_n^*)(1-2\sigma_n^*)\right]$$

Proof :

$$\frac{\partial \ln\left|\mathbf{W}_N + \mathbf{C}_N^{-1}\right|}{\partial a_n^*} = \mathrm{Tr}\left[(\mathbf{W}_N + \mathbf{C}_N^{-1})^{-1}\frac{\partial(\mathbf{W}_N + \mathbf{C}_N^{-1})}{\partial a_n^*}\right]$$
$$= \mathrm{Tr}\left[(\mathbf{I} + \mathbf{C}_N\mathbf{W}_N)^{-1}\mathbf{C}_N(\frac{\partial\mathbf{W}_N}{\partial a_n^*} + \frac{\partial\mathbf{C}_N^{-1}}{\partial a_n^*})\right]$$

$$\text{Here } \frac{\partial\mathbf{C}_N^{-1}}{\partial a_n^*} = 0$$

$$\frac{\partial\mathbf{W}_N}{\partial a_n^*} = \frac{\partial\sigma(a_n^{-1})}{\partial a_n^*}(1-\sigma(a_n^*)) + \sigma(a_n^*)\left(-\frac{\partial\sigma(a_n^*)}{\partial a_n^*}\right)$$
$$(\text{Since } \frac{\partial\sigma}{\partial a} = \sigma(1-\sigma))$$
$$= \sigma(1-\sigma) + \sigma(-1)\sigma(1-\sigma)$$
$$= \sigma(1-\sigma)(1-2\sigma)$$

$$\therefore \quad \frac{\partial \ln\left|\mathbf{W}_N + \mathbf{C}_N^{-1}\right|}{\partial a_n^*} = \mathrm{Tr}\left[(\mathbf{I} + \mathbf{C}_N\mathbf{W}_N)^{-1}\mathbf{C}_N\sigma_n^*(1-\sigma_n^*)(1-2\sigma_n^*)\right]$$

PRML p.318:

"The Laplace approximation has been constructed such that $\Psi(\mathbf{a}_N)$ has zero gradient at $a_N = a_N^*$, and so $\Psi(\mathbf{a}_N)$ gives no contribution to the gradient as a result of its dependence on $\mathbf{a}_N^*$."

Proof :

Laplace approximation is made on Eq (6.89).

$$p(\overline{\mathbf{t}}_N|\theta) = \int p(\overline{\mathbf{t}}_N|\mathbf{a}_N)p(\mathbf{a}_N|\theta)d\mathbf{a}_N \tag{6.89}$$

$$\begin{aligned}
\text{Let} \quad f(\mathbf{a}_N) &= p(\overline{\mathbf{t}}_N|\mathbf{a}_N)p(\mathbf{a}_N|\theta) \\
\ln f(\mathbf{a}_N) &\simeq \ln f(\mathbf{a}_N^*) - \frac{1}{2}(\mathbf{a}_N - \mathbf{a}_N^*)^T A(\mathbf{a}_N - \mathbf{a}_N^*) \\
\text{Here} \quad \nabla(\ln f(\mathbf{a}_N))\big|_{\mathbf{a}_N=\mathbf{a}_N^*} &= 0
\end{aligned}$$

$\Psi(\mathbf{a}_N)$ in Bishop is defined as $\ln f(\mathbf{a}_N)$ above.

$$\begin{aligned}
\Psi(\mathbf{a}_N) &= \ln f(\mathbf{a}_N) \\
\Rightarrow \quad \nabla\Psi(\mathbf{a}_N)\big|_{\mathbf{a}_N=\mathbf{a}_N^*} &= 0
\end{aligned} \tag{1}$$

Applying this Laplace approximation, we get

$$\ln p(\overline{\mathbf{t}}_N|\theta) = \Psi(\mathbf{a}_N^*) - \frac{1}{2}\ln\left|\mathbf{W}_N + \mathbf{C}_N^{-1}\right| + \frac{N}{2}\ln(2\pi)$$

If we want to take a derivative of $\ln p(\overline{\mathbf{t}}_N|\theta)$ on $\mathbf{a}_N^*$, we know that from Eq (1) we will get

$$\nabla\Psi(\mathbf{a}_N^*) = 0$$

Chapter 7. Sparse Kernel Machines

Eq 7.56: (PRML p.341)

$$L = C\sum_{n=1}^{N}(\xi_n + \widehat{\xi}_n) + \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{n=1}^{N}(\mu_n\xi_n + \widehat{\mu}_n\widehat{\xi}_n)$$
$$- \sum_{n=1}^{N} a_n(\epsilon + \xi_n + y_n - t_n) - \sum_{n=1}^{N}\widehat{a}_n(\epsilon + \widehat{\xi}_n - y_n + t_n)$$

Proof :

The error function is given at Eq (7.55).

$$E_\epsilon = C\sum_{n=1}^{N}(\xi_n + \widehat{\xi}_n) + \frac{1}{2}\|\mathbf{w}\|^2$$

There are four constraints:

1. $\xi_n \geq 0$
2. $\widehat{\xi}_n \geq 0$
3. $t_n = y(x_n) + \epsilon + \xi_n$
4. $t_n = y(x_n) - \epsilon - \xi_n$

By assigning Lagrange multipliers for the constraints correspondingly,

1. μ_n
2. $\widehat{\mu}_n$
3. a_n
4. $\widehat{a}_n$

The Lagranian becomes (minimizing)

$$\therefore \;\; L = C\sum_{n=1}^{N}(\xi_n + \widehat{\xi}_n) + \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{n=1}^{N}(\mu_n\xi_n + \widehat{\mu}_n\widehat{\xi}_n)$$
$$- \sum_{n=1}^{N} a_n(\epsilon + \xi_n + y_n - t_n) - \sum_{n=1}^{N}\widehat{a}_n(\epsilon + \widehat{\xi}_n - y_n + t_n)$$

Eq 7.61: (PRML p.342)

$$\widetilde{L}(\mathbf{a},\widehat{\mathbf{a}}) = -\frac{1}{2}\sum_{n=1}^{N}\sum_{n=1}^{N}(a_n-\widehat{a}_n)(a_m-\widehat{a}_m)k(\mathbf{x}_n,\mathbf{x}_m) \\ -\epsilon\sum_{n=1}^{N}(a_n+\widehat{a}_n)+\sum_{n=1}^{N}(a_n-\widehat{a}_n)t_n$$

Proof :

$$L = C\sum_{n=1}^{N}(\xi_n+\widehat{\xi}_n)+\frac{1}{2}\|\mathbf{w}\|^2-\sum_{n=1}^{N}(\mu_n\xi_n+\widehat{\mu}_n\widehat{\xi}_n) \\ -\sum_{n=1}^{N}a_n(\epsilon+\xi_n+y_n-t_n)-\sum_{n=1}^{N}\widehat{a}_n(\epsilon+\widehat{\xi}_n-y_n+t_n)$$

$$\mathbf{w} = \sum_{n=1}^{N}(a_n-\widehat{a}_n)\boldsymbol{\phi}(x_n) \tag{7.57}$$

$$a_n+\mu_n = C \quad \rightarrow \quad \mu_n = C-a_n$$

$$\widehat{a}_n+\widehat{\mu}_n = C \quad \rightarrow \quad \widehat{\mu}_n = C-\widehat{a}_n$$

Substituting these into L,

$$\widetilde{L} = C\sum_{n=1}^{N}(\xi_n+\widehat{\xi}_n)+\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}(a_n-\widehat{a}_n)(a_m-\widehat{a}_m)\boldsymbol{\phi}(\mathbf{x}_n)\boldsymbol{\phi}(\mathbf{x}_m) \\ -\sum_{n=1}^{N}[(C-a_n)\xi_n+(C-\widehat{a}_n)\widehat{\xi}_n]-\sum_{n=1}^{N}[a_n(\epsilon+\xi_n)+\widehat{a}_n(\epsilon+\widehat{\xi}_n)] \\ -\sum_{n=1}^{N}(a_n-\widehat{a}_n)y_n+\sum_{n=1}^{N}(a_n-\widehat{a}_n)t_n \\ = C\sum_{n=1}^{N}(\xi_n+\widehat{\xi}_n)+\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}(a_n-\widehat{a}_n)(a_m-\widehat{a}_m)k(\mathbf{x}_n,\mathbf{x}_m) \\ -C\sum_{n=1}^{N}(\xi_n+\widehat{\xi}_n)+\sum_{n=1}^{N}(a_n\xi_n+\widehat{a}_n\widehat{\xi}_n)-\epsilon\sum_{n=1}^{N}(a_n+\widehat{a}_n) \\ -\sum_{n=1}^{N}(a_n\xi_n+\widehat{a}_n\widehat{\xi}_n)-\sum_{n=1}^{N}(a_n-\widehat{a}_n)y_n+\sum_{n=1}^{N}(a_n-\widehat{a}_n)t_n$$

Since $y_n = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_n) + b$,

$$\begin{aligned}
\sum_{n=1}^{N}(a_n - \widehat{a}_n)y_n &= \sum_{n=1}^{N}(a_n - \widehat{a}_n)[\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n) + b] \\
&= \sum_{n=1}^{N}(a_n - \widehat{a}_n)\left[\sum_{m=1}^{N}(a_m - \widehat{a}_m)\boldsymbol{\phi}(\mathbf{x}_m)\right]\boldsymbol{\phi}(\mathbf{x}_n) + \sum_{n=1}^{N}(a_n - \widehat{a}_n)b \\
&\quad \text{(From the constraint Eq (7.58), } \sum_{n=1}^{N}(a_n - \widehat{a}_n) = 0\text{)} \\
&= \sum_{n=1}^{N}\sum_{n=1}^{N}(a_n - \widehat{a}_n)(a_m - \widehat{a}_m)k(\mathbf{x}_n, \mathbf{x}_m)
\end{aligned}$$

$$\begin{aligned}
\therefore \quad \widetilde{L}(\mathbf{a}, \widehat{\mathbf{a}}) = &-\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}(a_n - \widehat{a}_n)(a_m - \widehat{a}_m)k(\mathbf{x}_n, \mathbf{x}_m) \\
&- \epsilon\sum_{n=1}^{N}(a_n + \widehat{a}_n) + \sum_{n=1}^{N}(a_n - \widehat{a}_n)t_n
\end{aligned}$$

PRML p.342:

"The support vectors are those data points that contribute to predictions given by (7.64), in other words those for which either $a_n \neq 0$ or $\widehat{a}_n \neq 0$. These are points that lie on the boundary of the ϵ-tube or outside the tube."

Proof :

$$y(\mathbf{x}) = \sum_{n=1}^{N}(a_n - \widehat{a}_n)k(\mathbf{x}, \mathbf{x}_n) + b \tag{7.64}$$

$$a_n(\epsilon + \xi_n + y_n - t_n) = 0 \tag{7.65}$$

$$\widehat{a}_n(\epsilon + \widehat{\xi}_n - y_n + t_n) = 0 \tag{7.66}$$

$$\epsilon + \xi_n + y_n - t_n = 0 \tag{1}$$

$$\epsilon + \widehat{\xi}_n - y_n + t_n = 0 \tag{2}$$

Equations (1) and (2) are not compatible. If we add them together, (assuming (1) and (2) are compatible)

$$2\epsilon + \xi_n + \widehat{\xi}_n = 0$$

However, $\epsilon > 0$ and $\xi_n \geq 0$ and $\widehat{\xi}_n \geq 0$. Therefore, (1) and (2) are not compatible. This means that either a_n or $\widehat{a}_n$ (or both) must be zero.

$$
\begin{aligned}
\epsilon + \xi_1 + y_1 - t_1 = 0 \quad &\rightarrow \quad a_1 = 0 \text{ or } a_1 \neq 0 \\
&\rightarrow \quad \epsilon + \widehat{\xi}_1 + y_1 - t_1 \neq 0 \quad \text{(compatibility)} \\
&\rightarrow \quad \widehat{a}_1 = 0 \qquad (3) \\
\epsilon + \widehat{\xi}_2 - y_2 + t_2 = 0 \quad &\rightarrow \quad a_2 \neq 0 \text{ or } a_2 = 0 \\
&\rightarrow \quad \epsilon + \xi_2 + y_2 - t_2 \neq 0 \quad \text{(compatibility)} \\
&\rightarrow \quad \widehat{a}_2 = 0 \qquad (4)
\end{aligned}
$$

In case (3), $\widehat{a}_1 = 0$ so we consider the upper side only.

$$
\begin{aligned}
\epsilon + \xi_1 + y_1 - t_1 = 0 \quad &\rightarrow \quad t_1 = y_1 + \epsilon + \xi_1 \geq y_1 + \epsilon \\
&\Rightarrow \quad t_1 \text{ is on the boundary or above the } \epsilon\text{-tube.}
\end{aligned}
$$

In case (4), $a_1 = 0$, so we consider the lower side only.

$$
\begin{aligned}
\epsilon + \widehat{\xi}_2 - y_2 + t_2 = 0 \quad &\rightarrow \quad t_2 = y_2 - \epsilon - \xi_2 \leq y_2 - \epsilon \\
&\Rightarrow \quad t_2 \text{ is on the boundary or below the } \epsilon\text{-tube.}
\end{aligned}
$$

Eq 7.95: (PRML p.351)

$$
\mathbf{C}^{-1} = \mathbf{C}_{-i}^{-1} - \frac{\mathbf{C}_{-i}^{-1}\boldsymbol{\varphi}_i\boldsymbol{\varphi}_i^T\mathbf{C}_{-i}^{-1}}{\alpha_i + \boldsymbol{\varphi}_i^T\mathbf{C}_{-i}^{-1}\boldsymbol{\varphi}_i}
$$

Proof :

Starting from Eq (7.93),

$$
\begin{aligned}
\mathbf{C} &= \mathbf{C}_{-i} + \alpha_i^{-1}\boldsymbol{\varphi}_i\boldsymbol{\varphi}_i^T \\
&= \mathbf{C}_{-i}(1 + \alpha_i^{-1}\boldsymbol{\varphi}_i\mathbf{C}_{-i}^{-1}\boldsymbol{\varphi}_i^T) \\
\mathbf{C}^{-1} &= (1 + \alpha_i^{-1}\boldsymbol{\varphi}_i\mathbf{C}_{-i}^{-1}\boldsymbol{\varphi}_i^T)^{-1}\mathbf{C}_{-i}^{-1} \\
&\quad (\because (\mathbf{A}\cdot\mathbf{B})^{-1} = \mathbf{B}^{-1}\cdot\mathbf{A}^{-1})
\end{aligned}
$$

We also know that $\mathbf{C}$ is symetric.

$$\mathbf{C} = \frac{1}{\beta}\mathbf{I} + \frac{1}{\alpha}\boldsymbol{\varphi}\boldsymbol{\varphi}^T$$

(where $\boldsymbol{\varphi}\boldsymbol{\varphi}^T$ is symmetric.)

Eq (C.7): $(\mathbf{A} + \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{D} + \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1}$ (Woodbury identity)

Applying the Woodbury identity by identifying $\alpha_i^{-1}\boldsymbol{\varphi}_i \Leftrightarrow \mathbf{B},\ \mathbf{C}_{-i}^{-1} \Leftrightarrow \mathbf{D}^{-1},$ and $\boldsymbol{\varphi}_i^T \Leftrightarrow \mathbf{C},$

$$\begin{aligned}(1 + \alpha_i^{-1}\boldsymbol{\varphi}_i\mathbf{C}_{-i}^{-1}\boldsymbol{\varphi}_i^T)^{-1} &= 1 - 1 \cdot \alpha_i^{-1}(\mathbf{C}_{-i} + \boldsymbol{\varphi}_i \cdot 1 \cdot \alpha_i^{-1}\boldsymbol{\varphi}_i)^{-1}\boldsymbol{\varphi}_i^T \cdot 1 \\ &= 1 - \frac{\alpha_i^{-1}\boldsymbol{\varphi}_i\boldsymbol{\varphi}_i^T}{\mathbf{C}_{-i} + \boldsymbol{\varphi}_i^T\alpha_i^{-1}\boldsymbol{\varphi}_i}\end{aligned}$$

$$\therefore\ \mathbf{C}^{-1} = \mathbf{C}_{-i}^{-1} - \frac{\alpha_i^{-1}\boldsymbol{\varphi}_i\boldsymbol{\varphi}_i^T\mathbf{C}_{-i}^{-1}}{\mathbf{C}_{-i} + \boldsymbol{\varphi}_i^T\alpha_i^{-1}\boldsymbol{\varphi}_i}$$

(Multiplying $\alpha_i\mathbf{C}_{-i}^{-1}$ on both numerator and denominator.)

$$= \mathbf{C}_{-i}^{-1} - \frac{\mathbf{C}_{-i}^{-1}\boldsymbol{\varphi}_i\boldsymbol{\varphi}_i^T\mathbf{C}_{-i}^{-1}}{\alpha_i + \boldsymbol{\varphi}_i^T\mathbf{C}_{-i}^{-1}\boldsymbol{\varphi}_i}$$

(We know that $\boldsymbol{\varphi}_i^T\mathbf{C}_{-i}^{-1}\boldsymbol{\varphi}_i = \boldsymbol{\varphi}_i\mathbf{C}_{-i}^{-1}\boldsymbol{\varphi}_i^T$.)

Eq 7.104 & 7.105: (PRML p.353)

$$\begin{aligned} q_i &= \frac{\alpha_i Q_i}{\alpha_i - S_i} \\ s_i &= \frac{\alpha_i S_i}{\alpha_i - S_i}\end{aligned}$$

Proof :

$$\mathbf{C}^{-1} = \mathbf{C}_{-i}^{-1} - \frac{\mathbf{C}_{-i}^{-1}\boldsymbol{\varphi}_i\boldsymbol{\varphi}_i^T\mathbf{C}_{-i}^{-1}}{\alpha_i + \boldsymbol{\varphi}_i^T\mathbf{C}_{-i}^{-1}\boldsymbol{\varphi}_i} \tag{7.95}$$

$$s_i = \boldsymbol{\varphi}_i^T\mathbf{C}_{-i}^{-1}\boldsymbol{\varphi}_i \tag{7.98}$$

$$q_i = \boldsymbol{\varphi}_i^T\mathbf{C}_{-i}^{-1}\bar{\mathbf{t}} \tag{7.99}$$

$$\begin{aligned}
Q_i &= \boldsymbol{\varphi}_i^T \mathbf{C}^{-1} \overline{\mathbf{t}} \\
&= \boldsymbol{\varphi}_i^T \mathbf{C}_{-i}^{-1} \overline{\mathbf{t}} - \frac{\boldsymbol{\varphi}_i^T \mathbf{C}_{-i}^{-1} \boldsymbol{\varphi}_i \boldsymbol{\varphi}_i^T \mathbf{C}_{-i}^{-1} \overline{\mathbf{t}}}{\alpha_i + \boldsymbol{\varphi}_i^T \mathbf{C}_{-i}^{-1} \boldsymbol{\varphi}_i} \\
&= q_i - \frac{s_i q_i}{\alpha_i + s_i} \\
&= \frac{q_i \alpha_i}{\alpha_i + s_i} \qquad (1) \\
S_i &= \boldsymbol{\varphi}_i^T \mathbf{C}^{-1} \boldsymbol{\varphi}_i \\
&= \boldsymbol{\varphi}_i^T \mathbf{C}_{-i}^{-1} \boldsymbol{\varphi}_i - \frac{\boldsymbol{\varphi}_i^T \mathbf{C}_{-i}^{-1} \boldsymbol{\varphi}_i \boldsymbol{\varphi}_i^T \mathbf{C}_{-i}^{-1} \boldsymbol{\varphi}_i}{\alpha_i + s_i} \\
&= s_i - \frac{s_i s_i^T}{\alpha_i + s_i} \\
&= \frac{s_i \alpha_i}{\alpha_i + s_i} \qquad (2)
\end{aligned}$$

Solving for s_i from Eq (2),

$$\begin{aligned}
S_i(\alpha_i + s_i) &= s_i \alpha_i \\
\therefore \quad s_i &= \frac{\alpha_i S_i}{\alpha_i - S_i}
\end{aligned}$$

Plugging this into Eq (1),

$$\begin{aligned}
q_i &= \frac{1}{\alpha_i} Q_i(\alpha_i + s_i) = \frac{1}{\alpha_i} Q_i \left(\alpha_i + \frac{\alpha_i S_i}{\alpha_i - S_i} \right) \\
\therefore \quad q_i &= \frac{\alpha_i Q_i}{\alpha_i - S_i}
\end{aligned}$$

Eq 7.109: (PRML p.354)

$$\begin{aligned}
\ln p(\mathbf{w}|\overline{\mathbf{t}}, \boldsymbol{\alpha}) &= \ln \{ p(\overline{\mathbf{t}}|\mathbf{w}) p(\mathbf{w}|\boldsymbol{\alpha}) \} - \ln p(\overline{\mathbf{t}}|\boldsymbol{\alpha}) \\
&= \sum_{n=1}^{N} \{ t_n \ln y_n + (1 - t_n) \ln (1 - y_n) \} - \frac{1}{2} \mathbf{w}^T \mathbf{A} \mathbf{w} + const
\end{aligned}$$

Proof :

In section 7.2.3, we are considering two-class problems with a binary target t $\in \{0, 1\}$

as in Chap.4. The difference here is that $\boldsymbol{\alpha}$ (prior parameter) is a vector.

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

$$p(\bar{\mathbf{t}}|\mathbf{w}) = \prod_{n=1}^{N} y_n^{t_n}\{1 - y_n\}^{1-t_n}$$

From 1-D Gaussian,

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

Eq (1) becomes

$$\begin{aligned}
p(\mathbf{w}|\boldsymbol{\alpha}) &= \prod_{i=1}^{M} \frac{1}{(2\pi\alpha_i^{-1})^{1/2}} \exp\left\{-\frac{1}{2\alpha_i^{-1}} w_i^2\right\} \\
&= \frac{1}{(2\pi)^{M/2}} \prod_{i=1}^{M} (\alpha_i)^{1/2} \exp\left\{-\frac{\alpha_i}{2} w_i^2\right\} \\
&= \frac{1}{(2\pi)^{M/2}} \left[\prod_{i=1}^{M} (\alpha_i)^{1/2}\right] \exp\left\{\sum_{i=1}^{M} (-\frac{1}{2}\alpha_i w_i^2)\right\}
\end{aligned}$$

We can identify the square brackets and curly brackets above as follows,

$$\prod_{i=1}^{M} (\alpha_i)^{1/2} = \left|\mathbf{A}\right|^{1/2}$$

$$\sum_{i=1}^{M} (-\frac{1}{2}\alpha_i w_i^2) = -\frac{1}{2}\mathbf{w}^T\mathbf{A}\mathbf{w}$$

$$\mathbf{A} = \begin{bmatrix} \alpha_1 & & & \\ & \alpha_2 & & \\ & & \ddots & \\ & & & \alpha_M \end{bmatrix}$$

Where the off-diagonal elements are zeros.

$$
\begin{aligned}
\therefore \quad \ln p(\mathbf{w}|\overline{\mathbf{t}},\boldsymbol{\alpha}) &= \ln p(\overline{\mathbf{t}}|\mathbf{w}) + \ln(\mathbf{w}|\boldsymbol{\alpha}) - \ln p(\overline{\mathbf{t}}|\boldsymbol{\alpha}) \\
&= \sum_{n=1}^{N}\{t_n \ln y_n + (1-t_n)\ln(1-y_n)\} - \frac{1}{2}\mathbf{w}^T A\mathbf{w} \\
&\quad + \ln\left[\frac{\mathbf{A}^{1/2}}{(2\pi)^{M/2}}\right] - \ln p(\overline{\mathbf{t}}|\boldsymbol{\alpha})
\end{aligned}
$$

where the last two terms are constant.

Eq 7.114: (PRML p.355)

$$
\begin{aligned}
p(\overline{\mathbf{t}}|\boldsymbol{\alpha}) &= \int p(\overline{\mathbf{t}}|\mathbf{w})p(\mathbf{w}|\boldsymbol{\alpha})d\mathbf{w} \\
&\simeq p(\overline{\mathbf{t}}|\mathbf{w}^*)p(\mathbf{w}^*|\boldsymbol{\alpha})(2\pi)^{M/2}|\boldsymbol{\Sigma}|^{1/2}
\end{aligned}
$$

Proof :

Using the sum rule,

$$
\begin{aligned}
p(\mathbf{X}) &= \sum_i p(\mathbf{X},\mathbf{Y}_i) = \int p(\mathbf{X},\mathbf{Y})d\mathbf{Y} = \int p(\mathbf{X}|\mathbf{Y})p(\mathbf{Y})d\mathbf{Y} \\
p(\overline{\mathbf{t}}|\boldsymbol{\alpha}) &= \int p(\overline{\mathbf{t}},\mathbf{w}|\boldsymbol{\alpha})d\mathbf{w} \\
p(\overline{\mathbf{t}},\mathbf{w}|\boldsymbol{\alpha}) &= \frac{p(\overline{\mathbf{t}}\mathbf{w},\boldsymbol{\alpha})}{p(\boldsymbol{\alpha})} = \frac{p(\overline{\mathbf{t}}|\mathbf{w},\boldsymbol{\alpha})p(\mathbf{w},\boldsymbol{\alpha})}{p(\boldsymbol{\alpha})} \\
&= \frac{p(\overline{\mathbf{t}}|\mathbf{w})p(\mathbf{w}|\boldsymbol{\alpha})p(\boldsymbol{\alpha})}{p(\boldsymbol{\alpha})} \\
&= p(\overline{\mathbf{t}}|\mathbf{w})p(\mathbf{w}|\boldsymbol{\alpha}) \\
\Rightarrow \quad p(\overline{\mathbf{t}}|\boldsymbol{\alpha}) &= \int p(\overline{\mathbf{t}}|\mathbf{w})p(\mathbf{w}|\boldsymbol{\alpha})d\mathbf{w} \qquad (1)
\end{aligned}
$$

Using Eq (4.135), normalization constant Z in Laplace approximation,

$$
Z = \int f(\mathbf{z})d\mathbf{z} = f(\mathbf{z}_0)\frac{(2\pi)^{M/2}}{|\mathbf{A}|^{1/2}}
$$

We can identify Z and f(z) in Eq (1) as follows,

$$
p(\overline{\mathbf{t}}|\boldsymbol{\alpha}) \Leftrightarrow Z, \quad p(\overline{\mathbf{t}}|\mathbf{w})p(\mathbf{w}|\boldsymbol{\alpha}) \Leftrightarrow f(\mathbf{z})
$$

$$
\begin{aligned}
f(\mathbf{w}) &= p(\overline{\mathbf{t}}|\mathbf{w})p(\mathbf{w}|\boldsymbol{\alpha}) \\
f(\mathbf{w}^*) &= p(\overline{\mathbf{t}}|\mathbf{w}^*)p(\mathbf{w}^*|\boldsymbol{\alpha}) \quad (\mathbf{w}^* : \text{mode from } \nabla f(\mathbf{w}) = 0) \\
&\therefore \;\; p(\overline{\mathbf{t}}|\boldsymbol{\alpha}) \simeq p(\overline{\mathbf{t}}|\mathbf{w}^*)p(\mathbf{w}^*|\boldsymbol{\alpha})(2\pi)^{M/2}|\boldsymbol{\Sigma}|^{1/2} \\
\text{where } \; \boldsymbol{\Sigma}^{-1} &= -\nabla\nabla \ln f(\mathbf{w})\Big|_{\mathbf{w}=\mathbf{w}^*}
\end{aligned}
$$

Chapter 8. Graphical Models

Eq 8.16: (PRML p.371)

$$
\begin{aligned}
\text{cov}[x_i, x_j] &= \mathbb{E}[(x_i - \mathbb{E}[x_i])(x_j - \mathbb{E}[x_j])] \\
&= \mathbb{E}\left\{(x_i - \mathbb{E}[x_i])\left[\sum_{k \in pa_j} w_{jk}(x_k - \mathbb{E}[x_k]) + \sqrt{v_j}\epsilon_j\right]\right\} \\
&= \sum_{k \in pa_j} w_{jk}\text{cov}[x_i, x_k] + I_{ij}v_j
\end{aligned}
$$

Proof :

$$x_i = \sum_{j \in pa_i} w_{ij}x_j + b_i + \sqrt{v_i}\epsilon_i \tag{8.14}$$

$$\mathbb{E}[x_i] = \sum_{j \in pa_i} w_{ij}\mathbb{E}[x_j] + b_i \tag{8.15}$$

$$\mathbb{E}[\epsilon_i\epsilon_j] = I_{ij} \tag{1}$$

The tricky thing in this problem is the definition of x_i and x_j.

x_i: i = 1, . . . , D vector component (e.g. x, y, z, . . .)

x_{il}: l = 1, . . . , n data point #

Using

$$
\begin{aligned}
x_i &= \sum_{j \in pa_i} w_{ij}x_j + b_i + \sqrt{v_i}\epsilon_i \\
x_{il} &= \sum_{j \in pa_i} w_{ij}x_{jl} + b_i + \sqrt{v_i}\epsilon_i \\
x_{jl} &= \sum_{k \in pa_j} w_{jk}x_{kl} + b_j + \sqrt{v_j}\epsilon_j
\end{aligned}
$$

$$\begin{aligned}
\text{cov}[x_i, x_j] &= \mathbb{E}(x_{il} - \mathbb{E}[x_{il}])(x_{jl} - \mathbb{E}[x_{jl}]) \\
&= \mathbb{E}(x_{il} - \mathbb{E}[x_{il}]) \left[\sum_{k \in pa_j} w_{jk} x_{kl} + b_j + \sqrt{v_j} \epsilon_j - \sum_{k \in pa_j} w_{jk} \mathbb{E}[x_{kl}] - b_j - 0 \right] \\
&= \mathbb{E} \sum_{k \in pa_j} (x_{il} - \mathbb{E}[x_{il}]) w_{jk} (x_{kl} - \mathbb{E}[x_{kl}]) + \mathbb{E}(x_{il} - \mathbb{E}[x_{il}]) \sqrt{v_j} \epsilon_j \\
&= \sum_{k \in pa_j} w_{jk} \mathbb{E}(x_{il} - \mathbb{E}[x_{il}])(x_{kl} - \mathbb{E}[x_{kl}]) + \mathbb{E}(x_{il} - \mathbb{E}[x_{il}]) \sqrt{v_j} \epsilon_j \\
&= \sum_{k \in pa_j} w_{jk} \text{cov}(x_i, x_k) + \mathbb{E}[(x_{il} - \mathbb{E}[x_{il}]) \sqrt{v_j} \epsilon_j] \\
&= \sum_{k \in pa_j} w_{jk} \text{cov}(x_i, x_k) + \sqrt{v_i v_j}\, \mathbb{E}[\epsilon_i \epsilon_j] \\
&\text{(from Eq (1))} \\
&= \sum_{k \in pa_j} w_{jk} \text{cov}(x_i, x_k) + v_j I_{ij}
\end{aligned}$$

PRML p.374:

$$p(a, b) = p(a) \sum_c p(c|a) p(b|c) = p(a) p(b|a)$$

Proof :

$$p(b|a) = \sum_c p(b, c|a) \qquad \text{(sum rule)}$$

$$\begin{aligned}
p(b, c|a) &= \frac{p(a, b, c)}{p(a)} \\
&= \frac{p(b|a, c) p(a, c)}{p(a)} \\
&= \frac{p(b|a, c) p(c|a) p(a)}{p(a)} \\
&= p(b|a, c) p(c|a)
\end{aligned}$$

Since b is not dependent on a (see Fig 8.17),

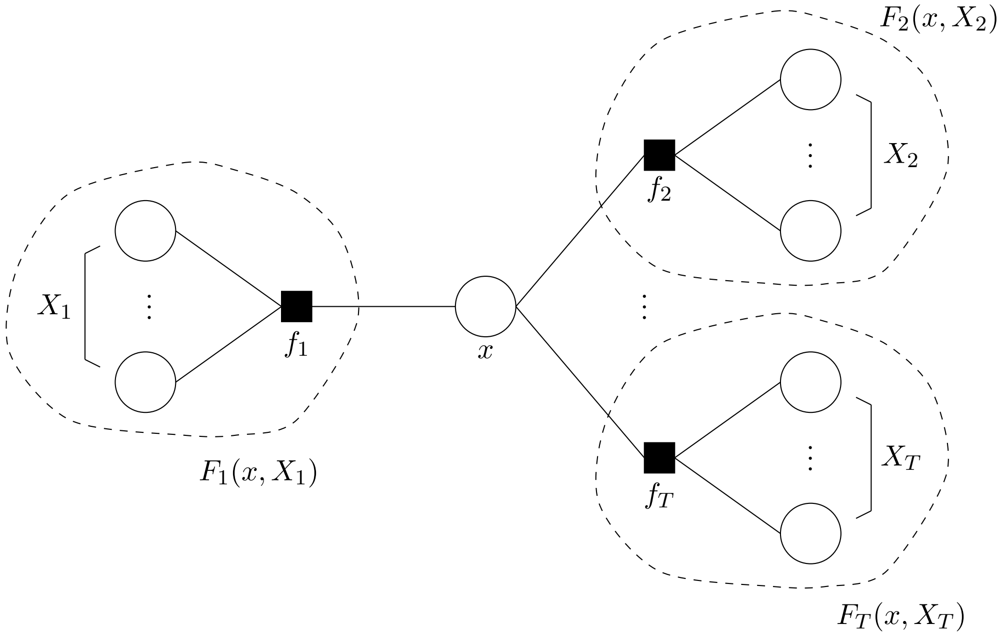
$$p(b|a, c) = p(b|c)$$

$$\begin{aligned}
\therefore \quad p(b|a) &= \sum_c p(b,c|a) \\
&= \sum_c p(b|a,c)p(c|a) \\
&= \sum_c p(c|a)p(b|c)
\end{aligned}$$

Eq 8.63: (PRML p.404)

$$\begin{aligned}
p(x) &= \sum_{\mathbf{x}\backslash x} \prod_{s\in ne(x)} F_s(x, X_s) \\
&= \prod_{s\in ne(x)} \left[\sum_{X_s} F_s(x, X_s) \right]
\end{aligned}$$

Proof :



$\mathbf{x}_n \backslash x = X1, X2, \cdots, X_T$ (except for x)

$$
\begin{aligned}
p(x) &= \sum_{\mathbf{x}\backslash x} p(\mathbf{x}) \\
&= \sum_{\mathbf{x}\backslash x} \left[\prod_{s \in ne(x)} F_s(x, X_s) \right] \\
&= \sum_{\mathbf{x}\backslash x} [F_1(x, X1) \cdot F_2(x, X_2) \cdots F_T(x, X_T)]
\end{aligned}
$$

Since

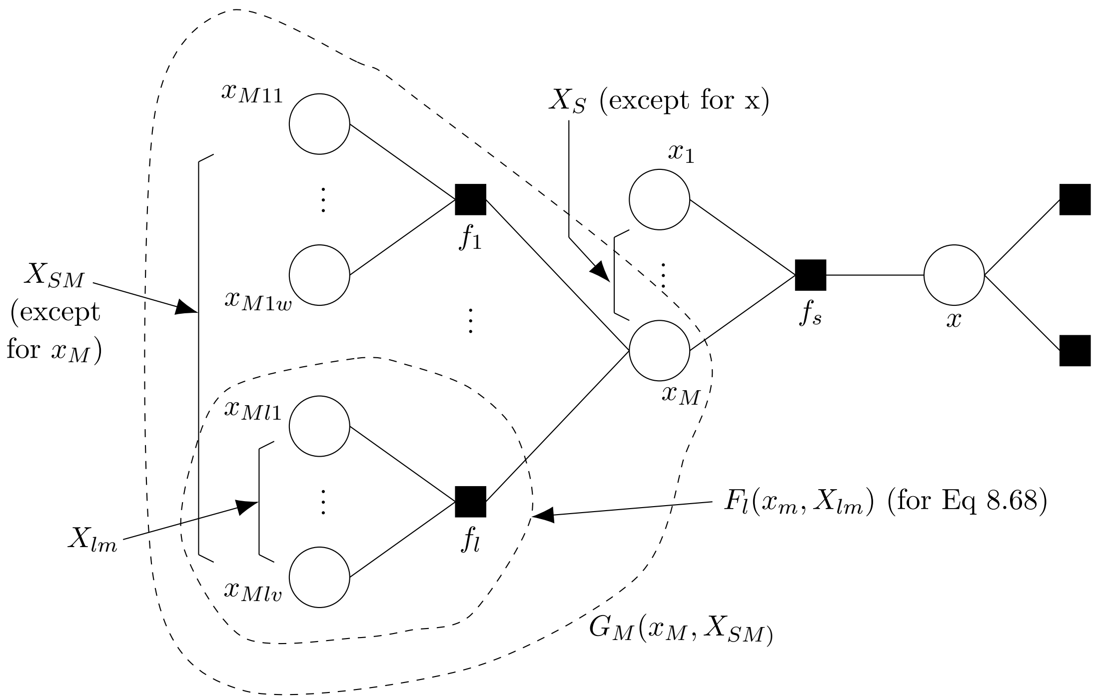
$$
\begin{aligned}
&X_1 = \{x_1, x_2, \cdots, x_k\} \\
&X_2 = \{x_{k+1}, x_{k+2}, \cdots, x_l\} \\
&\quad \vdots
\end{aligned}
$$

$$
\begin{aligned}
p(x) &= \sum_{x_1 \in X_1} \sum_{x_2 \in X_2} \cdots \sum_{x_T \in X_T} [F_1(x, X_1) \cdots F_T(x, X_T)] \\
&= \left[\sum_{X_1} F_1(x, X_1) \right] \cdot \left[\sum_{x_2} \cdots \sum_{X_T} F(x, X_2) \cdots F_T(x, X_T) \right] \\
&= \left[\sum_{X_1} F_1(x, X_1) \right] \cdot \left[\sum_{X_2} F_2(x, X_2) \right] \cdots \left[\sum_{X_T} F_T(x, X_T) \right] \\
&= \prod_{s \in ne(x)} \left[\sum_{X_s} F_s(x, X_s) \right]
\end{aligned}
$$

Eq 8.66: (PRML p.404)

$$
\mu_{f_s \to x}(x) = \sum_{x_1} \cdots \sum_{x_M} f_x(x, x_1, \cdots, x_M) \prod_{m \in ne(f_s)\backslash x} \left[\sum_{X_{sm}} G_m(x_m, X_{sm}) \right]
$$

Proof :



$$X_{SM1} = (x_{M11}, x_{M12}, \cdots, x_{M1w})$$

$$X_{SMl} = (x_{Ml1}, x_{Ml2}, \cdots, x_{Mlv})$$

$$\begin{aligned}
\mu_{f_s \to x}(x) &= \sum_{X_s} F_s(x, X_s) \qquad\qquad (\leftarrow \text{Eq } 8.64) \\
&= \sum_{X_s} f_s(x, x_1, \cdots, x_M) \cdot G_1(x_1, X_{S1}) \cdots G_M(x_M, X_{SM}) \\
&= \sum_{x_1} \sum_{x_2} \cdots \sum_{x_M} \cdot \sum_{X_{S1}} \sum_{X_{S2}} \cdots \sum_{X_{SM}} f_s(x, x_1, \cdots, x_M) \cdot G_1(x_1, X_{S1}) \\
&\quad \cdots G_M(x_M, X_{SM}) \\
&= \sum_{x_1} \cdots \sum_{x_M} f_s(x, x_1, \cdots, x_M) \sum_{X_{S1}} \cdots \sum_{X_{SM}} \prod_{m \in ne(f_s) \setminus x} G_m(x_m, X_{Sm})
\end{aligned}$$

Using Eq (8.63) to switch Σ and $\prod$,

$$\therefore \ \mu_{f_s \to x} = \sum_{x_1} \cdots \sum_{x_M} f_x(x, x_1, \cdots, x_M) \prod_{m \in ne(f_s) \setminus x} \left[\sum_{X_{sm}} G_m(x_m, X_{sm}) \right]$$

Eq 8.69: (PRML p.406)

$$\mu_{x_m \to f_s}(x_m) = \prod_{l \in ne(x_m) \setminus f_s} \left[\sum_{X_{lm}} F_l(x_m, X_{lm}) \right]$$

Proof :

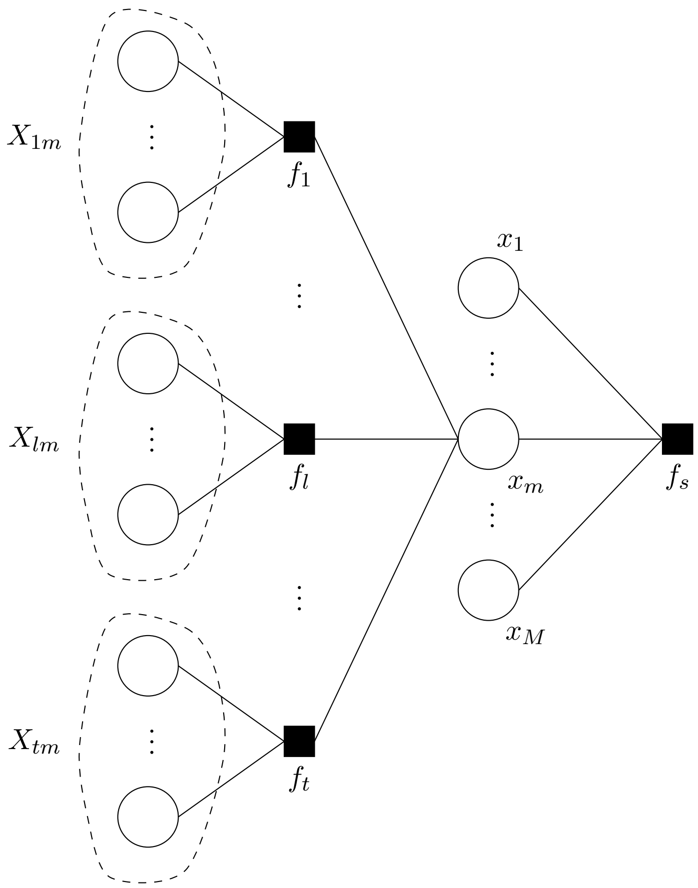
$$\mu_{x_m \to f_x}(x_m) = \sum_{X_{sm}} G_m(x_m, X_{sm}) \tag{8.67}$$

$$G_m(x_m, X_{sm}) = \prod_{l \in ne(x_m) \setminus f_s} F_l(x_m, X_{lm}) \tag{8.68}$$

Plugging Eq (8.68) into (8.67),

$$\mu_{x_m \to f_x}(x_m) = \sum_{X_{sm}} \left[\prod_{l \in ne(x_m) \setminus f_s} F_l(x_m, X_{lm}) \right] \tag{1}$$

We need to understand X_{sm} and X_{lm}. From the graph in the proof of Eq (8.66),



$$X_{sm} = \{X_{1m}, X_{2m}, \cdots, X_{lm}, \cdots, X_{tm}\}$$

To be rigorous, X_{1m} should be denoted as X_{s1m}, but it is considered as a short handed notation.

$$\begin{aligned}
\text{Eq (1)} &= \sum_{X_{1m}} \sum_{X_{2m}} \cdots \sum_{X_{tm}} [F_1(x_m, X_{1m}) \cdot F_2(x_m, X_{2m}) \cdots F_t(x_m, X_{tm})] \\
&= \left[\sum_{X_{1m}} F_1(x_m, X_{1m})\right] \cdot \left[\sum_{X_{2m}} F_2(x_m, X_{2m})\right] \cdots \left[\sum_{X_{tm}} F_t(x_m, X_{tm})\right] \\
&= \prod_{l \in ne(x_m) \setminus f_s} \left[\sum_{X_{lm}} F_l(x_m, X_{lm})\right]
\end{aligned}$$

Chapter 9. Mixture Models and EM

Eq 9.19: (PRML p.436)

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T$$

Proof :

$$\ln p(\mathbf{X}|\pi, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

Taking a derivative w.r.t. $\boldsymbol{\Sigma}_k$ and set it to zero to find the maximum.

$$\begin{aligned}
\frac{\partial \ln p}{\partial \boldsymbol{\Sigma}_k} &= \sum_{n=1}^{N} \frac{\frac{\partial}{\partial \boldsymbol{\Sigma}_k}\left[\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right]}{\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \\
&= \sum_{n=1}^{N} \frac{\pi_k \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \\
&= 0
\end{aligned}$$

Gaussian is

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

Derivative of Gaussian w.r.t. $\boldsymbol{\Sigma}_k$ is,

$$\begin{aligned}
\frac{\partial \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\partial \boldsymbol{\Sigma}_k} &= \frac{1}{(2\pi)^{D/2}} \left(-\frac{1}{2}\right) \frac{1}{\boldsymbol{\Sigma}_k^{3/2}} \exp\left[-\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k)\right] \\
&+ \frac{1}{(2\pi)^{D/2}} \frac{1}{\boldsymbol{\Sigma}_k^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k)\right] \left[\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-2} (\mathbf{x}_n - \boldsymbol{\mu}_k)\right] \\
&= \frac{1}{(2\pi)^{D/2}} \frac{1}{\boldsymbol{\Sigma}_k^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k) \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k)\right] \\
&\quad \cdot \left[-\frac{1}{2}\frac{1}{\boldsymbol{\Sigma}_k} + \frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-2} (\mathbf{x}_n - \boldsymbol{\mu}_k)\right] \\
&= \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \cdot \left[-\frac{1}{2}\frac{1}{\boldsymbol{\Sigma}_k} + \frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-2} (\mathbf{x}_n - \boldsymbol{\mu}_k)\right]
\end{aligned}$$

Therefore,

$$\begin{aligned}
\frac{\partial \ln p}{\partial \boldsymbol{\Sigma}_k} &= \sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \cdot \left(\frac{1}{2\boldsymbol{\Sigma}_k}\right) \cdot \left[1 + (\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k)\right] \\
&= 0
\end{aligned}$$

Multiplying by $\boldsymbol{\Sigma}_k$, and using

$$\gamma(z_{nk}) \equiv \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \tag{9.16}$$

$$N_k \equiv \sum_{n=1}^{N} \gamma(z_{nk}) \tag{9.18}$$

$$\begin{aligned}
&\sum_{n=1}^{N} \gamma(z_{nk}) \left[1 + (\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k)\right] = 0 \\
\Rightarrow \quad & N_k + \sum_{n=1}^{N} \gamma(z_{nk}) (\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k) = 0 \\
\therefore \quad & \boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T
\end{aligned}$$

Eq 9.22: (PRML p.436)

$$\pi_k = \frac{N_k}{N}$$

Proof :

Using the Lagrangian Eq (E.4),

$$L = \ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) + \lambda(\sum_{k=1}^{K} \pi_k - 1)$$

$$\frac{\partial L}{\partial \pi_k} = \sum_{n=1}^{N} \frac{\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} + \lambda = 0 \tag{1}$$

Multipy Eq (1) by π_k and sum over k,

$$\sum_{n=1}^{N} \sum_{k=1}^{K} \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} + \sum_{k=1}^{K} \lambda \pi_k = 0$$

$$\text{where} \quad \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} = \gamma(z_{nk})$$

$$\Rightarrow \quad \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) + \lambda \sum_{k=1}^{K} \pi_k = 0$$

Since the first term = N,

$$\lambda = -N$$

Multipying Eq (1) by π_k and using this λ,

$$\sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} - N\pi_k = 0$$

$$\Rightarrow \quad \sum_{n=1}^{N} \gamma(z_{nk}) - N\pi_k = 0$$

Since $N_k = \sum_{n=1}^{N} \gamma(z_{nk})$ (Eq 9.18),

$$\therefore \quad \pi_k = \frac{N_k}{N}$$

Eq 9.37: (PRML p.442)

$$\pi_k = \frac{1}{N} \sum_{n=1}^{N} z_{nk}$$

Proof :

$$\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \{\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\} \tag{9.36}$$

Lagrangian is,

$$L = \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right)$$

$$\frac{\partial L}{\partial \pi_k} = \sum_n z_{nk} \frac{1}{\pi_k} + \lambda = 0 \tag{1}$$
$$\rightarrow \quad \pi_k = -\frac{1}{\lambda} \sum_n z_{nk}$$

To calculate λ, multiply Eq (1) by π_k and sum over k,

$$\sum_n \sum_k z_{nk} + \lambda \sum_k \pi_k = 0$$

Since $\sum_k \pi_k = 1$,

$$\lambda = -\sum_n \sum_k z_{nk} = -N$$
$$\therefore \quad \pi_k = \frac{1}{N} \sum_n z_{nk}$$

Eq 9.39: (PRML p.443)

$$\begin{aligned}
\mathbb{E}[z_{nk}] &= \frac{\sum_{z_n} z_{nk} \prod_{k'} [\pi_{k'} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_{k'}, \boldsymbol{\Sigma}_{k'})]^{z_{nk'}}}{\sum_{z_n} \prod_j [\pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)]^{z_{nj}}} \\
&= \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}
\end{aligned}$$

Proof :

$\mathbb{E}[z_{nk}]$: Expectation value of z_{nk} over z_n, where n is fixed.

$z_n \rightarrow \{(z_{n1} = 1), (z_{n2} = 1), \ldots, (z_{nk} = 1)\}$

We also need to understand the meaning of $\sum_{z_n}$.

$$\sum_{z_n} = (z_{n1} = 1\text{case }) + (z_{n2} = 1\text{case }) + \ldots + (z_{nk} = 1\text{case })$$

When $z_{n1} = 1$ case, all the other $z_{nk} = 0$.

Numerator:

$$\begin{aligned}
\sum_{z_n} z_{nk} \prod_{k'} [\pi_{k'} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_{k'}, \boldsymbol{\Sigma}_{k'})]^{z_{nk'}} = \sum_{z_n} z_{nk} [\pi_1 \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)]^{z_{n1}} \\
\cdot [\pi_2 \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)]^{z_{n2}} \cdots [\pi_K \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K)]^{z_{nK}} \quad (1)
\end{aligned}$$

Out of all the cases of z_n, only $z_{nk} = 1$ case survies because of the z_{nk} term in the numerator in Eq (9.39).

Therefore,

$$\text{Eq (1)} = \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Denominator:

$$\sum_{z_n} \prod_j [\pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)]^{z_{nj}}$$

In this case, at each case of $z_n (z_{n1} = 1), (z_{n2} = 1), \dots, (z_{nk} = 1)$ the terms inside the bracket survive.

$$\begin{aligned}
\text{Denominator } &= \prod_j [\pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)]^{z_{nj}}\Big|_{z_{n1}=1, \text{ all others } = 0} \\
&+ \prod_j [\pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)]^{z_{nj}}\Big|_{z_{n2}=1, \text{ all others } = 0} \\
&+ \cdots \\
&+ \prod_j [\pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)]^{z_{nj}}\Big|_{z_{nK}=1, \text{ all others } = 0} \\
&= \pi_1 \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + \pi_2 \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) + \cdots + \pi_K \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K) \\
&= \sum_j \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)
\end{aligned}$$

$$\therefore \quad \mathbb{E}[z_{nk}] = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

Eq 9.62: (PRML p.449)

$$\begin{aligned}
\mathbb{E}]\ln p(\mathbf{t}, \mathbf{w}|\alpha, \beta)] = \frac{M}{2}\ln\left(\frac{\alpha}{2\pi}\right) - \frac{\alpha}{2}\mathbb{E}[\mathbf{w}^T\mathbf{w}] + \frac{N}{2}\ln\left(\frac{\beta}{2\pi}\right) \\
- \frac{\beta}{2}\sum_{n=1}^{N}\mathbb{E}[(t_n - \mathbf{w}^T\boldsymbol{\phi}_n)^2]
\end{aligned}$$

Proof :

Eq (3.56): (q = 2)

$$\begin{aligned}
p(\mathbf{w}|\alpha) &= \left[\left(\frac{\alpha}{2}\right)^{1/2} \frac{1}{\Gamma(1/2)}\right]^M \exp\left(-\frac{\alpha}{2}\sum_{j=0}^{M-1}|w_j|^2\right) \\
&= \left(\frac{\alpha}{2\pi}\right)^{M/2} \exp\left(-\frac{\alpha}{2}\sum_{j=0}^{M-1}|w_j|^2\right)
\end{aligned}$$

where $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$ is used.

Eq (3.11):

$$\ln p(\bar{\mathbf{t}}|\mathbf{w},\beta) = \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) - \frac{\beta}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n)\}^2$$

$$\begin{aligned}\mathbb{E}[\ln p(\bar{\mathbf{t}},\mathbf{w}|\alpha,\beta)] &= \mathbb{E}[\ln p(\bar{\mathbf{t}}|\mathbf{w},\beta)] + \mathbb{E}[\ln p(\mathbf{w}|\alpha)] \\ &= \frac{M}{2}\ln\left(\frac{\alpha}{2\pi}\right) - \frac{\alpha}{2}\mathbb{E}[\mathbf{w}^T\mathbf{w}] + \frac{N}{2}\ln\left(\frac{\beta}{2\pi}\right) \\ &\quad - \frac{\beta}{2}\sum_{n=1}^{N}\mathbb{E}[(t_n - \mathbf{w}^T\boldsymbol{\phi}_n)^2]\end{aligned}$$

Chapter 10. Approximate Inference

Eq 10.25: (PRML p.471)

$$\ln q_\mu^*(\mu) = -\frac{\mathbb{E}[\tau]}{2}\left\{\lambda_0(\mu-\mu_0)^2 + \sum_{n=1}^{N}(x_n-\mu)^2\right\} + const$$

Proof :

$$p(D|\mu,\tau) = \left(\frac{\tau}{2\pi}\right)^{N/2}\exp\left\{-\frac{\tau}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right\} \tag{10.21}$$

$$p(\mu|\tau) = \mathcal{N}(\mu|\mu_0,(\lambda_0\tau)^{-1}) \tag{10.22}$$

$$p(\tau) = \mathrm{Gam}(\tau|a_0,b_0) \tag{10.23}$$

From Eq (10.9):

$$\begin{aligned}\ln q_\mu^*(\mu) &= \mathbb{E}_{\tau(i\neq\mu)}[\ln p(D,\mu,\tau)] + const \\ &= \mathbb{E}_\tau[\ln p(D|\mu,\tau)\cdot p(\mu|\tau)\cdot p(\tau)] + const \\ &= \mathbb{E}_\tau[\ln p(D|\mu,\tau)] + \mathbb{E}_\tau[p(\mu|\tau)] + \mathbb{E}_\tau[p(\tau)] + const \\ &= \left\{\mathbb{E}_\tau\left[\frac{N}{2}\ln\left(\frac{\tau}{2\pi}\right)\right] + \mathbb{E}_\tau[\tau]\cdot\left(-\frac{1}{2}\right)\sum_{n=1}^{N}(x_n-\mu)^2\right\} \\ &\quad + \left\{\mathbb{E}_\tau\left[\frac{1}{2}\ln\left(\frac{\lambda_0\tau}{2\pi}\right)\right] + \mathbb{E}_\tau[\tau]\cdot\left(-\frac{\lambda_0}{2}\right)(\mu-\mu_0)^2\right\} \\ &\quad + \mathbb{E}_\tau[p(\tau)] + const\end{aligned}$$

Since

$$\mathbb{E}_\tau\left[\frac{N}{2}\ln\left(\frac{\tau}{2\pi}\right)\right] = const$$

$$\mathbb{E}_\tau\left[\frac{1}{2}\ln\left(\frac{\lambda_0\tau}{2\pi}\right)\right] = const$$

$$\mathbb{E}_\tau[p(\tau)] = const$$

$$\therefore \quad \ln q_\mu^*(\mu) = -\frac{\mathbb{E}[\tau]}{2}\left\{\lambda_0(\mu-\mu_0)^2 + \sum_{n=1}^{N}(x_n-\mu)^2\right\} + const$$

Eq 10.28: (PRML p.471)

$$\begin{aligned}\ln q_\mu^*(\tau) = (a_0-1)\ln\tau - b_0\tau + \frac{N}{2}\ln\tau + \frac{1}{2}\ln\tau \\ -\frac{\tau}{2}\mathbb{E}_\mu\left[\sum_{n=1}^{N}(x_n-\mu)^2 + \lambda_0(\mu-\mu_0)^2\right] + const\end{aligned}$$

Proof :

$$\begin{aligned}\ln p(D,\mu,\tau) = &\left\{\frac{N}{2}\ln\left(\frac{\tau}{2\pi}\right) + \left[-\frac{\tau}{2}\sum(x_n-\mu)^2\right]\right\} + \left\{\ln\left(\frac{\lambda_0\tau}{2\pi}\right)^{1/2}\right. \\ &\left. + \left(-\frac{\lambda_0\tau}{2}\right)(\mu-\mu_0)^2\right\} + \ln\left[\frac{1}{\Gamma(a_0)}b_0^{a_0}\right] + (a_0-1)\ln\tau + (-b_0\tau) \\ &= (a_0-1)\ln\tau - b_0\tau + \frac{N}{2}\ln\tau + \frac{1}{2}\ln\tau - \frac{\tau}{2}\left[\sum(x_n-\mu)^2\right. \\ &\left. +\lambda_0(\mu-\mu_0)^2\right] + const\end{aligned}$$

$$\begin{aligned}\therefore \quad \ln q_\mu^*(\tau) &= \mathbb{E}_\mu[\ln p(D,\mu,\tau)] \\ &= (a_0-1)\ln\tau - b_0\tau + \frac{N}{2}\ln\tau + \frac{1}{2}\ln\tau \\ &\quad -\frac{\tau}{2}\mathbb{E}_\mu\left[\sum_{n=1}^{N}(x_n-\mu)^2 + \lambda_0(\mu-\mu_0)^2\right] + const\end{aligned}$$

Eq 10.29 & 10.30: (PRML p.471)

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

$$b_N = b_0 + \frac{1}{2}\mathbb{E}_\mu \ \sum(x_n-\mu)^2 + \lambda_0(\mu-\mu_0)^2]$$

Proof :

From Eq (10.28), the coefficients of $\ln\tau$ are,

$$(a_0-1)\ln\tau+\frac{N}{2}\ln\tau+\frac{1}{2}\ln\tau=\left[a_0+\frac{N-1}{2}\right]\ln\tau$$

Comparing to Gamma function,

$$\begin{aligned}\mathrm{Gam}(\tau|a_0,b_0)&=\frac{1}{\Gamma(a_0)}b_0^{a_0}\tau^{a_0-1}\exp(-b_0\tau)\\ \rightarrow\quad \ln\mathrm{Gam}(\tau|a_0,b_0)&=(a_0-1)\ln\tau-b_0\tau+const\\ \Rightarrow\quad a_N-1&=a_0+\frac{N-1}{2}\\ \therefore\quad a_N&=a_0+\frac{N+1}{2}\end{aligned}$$

From Eq (10.28), coefficients of τ are,

$$\begin{aligned}&-b_0\tau-\frac{\tau}{2}\mathbb{E}_\mu\left[\sum_{n=1}^{N}(x_n-\mu)^2-\lambda_0(\mu-\mu_0)^2\right]\\ &=-\left\{b_0+\frac{1}{2}\mathbb{E}_\mu\left[\sum_{n=1}^{N}(x_n-\mu)^2-\lambda_0(\mu-\mu_0)^2\right]\right\}\tau\\ \Rightarrow\quad & b_N=b_0+\frac{1}{2}\mathbb{E}_\mu\left[\sum(x_n-\mu)^2-\lambda_0(\mu-\mu_0)^2\right]\end{aligned}$$

Eq 10.50: (PRML p.477)

$$\mathbb{E}[z_{nk}]=r_{nk}$$

Proof :

$$q^*(\mathbf{Z})=\prod_{n=1}^{N}\prod_{k=1}^{K}r_{nk}^{z_{nk}} \tag{10.48}$$

To see how the sum over **Z** works, let's take a look at N = 1 case.

$$q^*(\mathbf{z}_1)=\prod_{k=1}^{K}r_{1k}^{z_{1k}}$$

$$\mathbb{E}[z_{1k}]=\sum_{\mathbf{z}_1}z_{1k}q^*(\mathbf{z}_1)=\sum_{\{z_1,z_2,\ldots,z_K\}}z_{1k}\prod_{k=1}^{K}r_{1k}^{z_{1k}}$$

where the summation is over

$$z_1 = [1, 0, 0, \dots, 0]$$
$$z_2 = [0, 1, 0, \dots, 0]$$
$$\vdots$$
$$z_K = [0, 0, 0, \dots, 1]$$

$$\Rightarrow \quad \mathbb{E}[z_{1k}] = z_{1k} r_{1k}^1 = r_{1k}$$

where all other z_i will set $z_{1k} = 0$. ([0,0, …, 1, 0, …, 0], where 1 occurs at the i-th position.)

In general case,

$$q^*(\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_N) = \prod_{n=1}^{N} \prod_{k=1}^{K} r_{nk}^{z_{nk}}$$

$$\mathbb{E}[z_{nk}] = \sum_{\mathbf{z}_1} \sum_{\mathbf{z}_2} \cdots \sum_{\mathbf{z}_N} z_{nk} \prod_{n=1}^{N} \prod_{k=1}^{K} r_{nk}^{z_{nk}}$$

(where z_{nk} determines which n and k will survive.)

$$= \sum_{z_n} z_{nk} \prod_{k=1}^{K} r_{nk}^{z_{nk}}$$

(all other $\sum_{\mathbf{z}_i, (i \neq n)}$ will give $z_{nk} = 0$)

$$= z_{nk} r_{nk}^1$$
$$= r_{nk}$$

Eq 10.54: (PRML p.477)

$$\ln q^*(\boldsymbol{\pi}, \boldsymbol{\mu}, \Lambda) = \ln p(\boldsymbol{\pi}) + \sum_{k=1}^{K} \ln p(\boldsymbol{\mu}_k, \mathbf{\Lambda}_k) + \mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{Z}|\boldsymbol{\pi})] + \sum_{k=1}^{K} \sum_{n=1}^{N} \mathbb{E}[z_{nk}] \ln \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \mathbf{\Lambda}_k^{-1}) + const$$

Proof :

$$\ln q^*(\boldsymbol{\pi}, \boldsymbol{\mu}, \Lambda) = \mathbb{E}_{\mathbf{Z}, (i \neq j)}[\ln p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \mathbf{\Lambda})] + const$$

Using Eq (10.41):

$$p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \Lambda) = p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})\, p(\mathbf{Z}|\boldsymbol{\pi})\, p(\boldsymbol{\pi})\, p(\boldsymbol{\mu}|\boldsymbol{\Lambda})\, p(\boldsymbol{\Lambda})$$

$$\begin{aligned}\ln q^*(\boldsymbol{\pi}, \boldsymbol{\mu}, \Lambda) &= \mathbb{E}_{\mathbf{Z}}[\ln \{p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})\, p(\mathbf{Z}|\boldsymbol{\pi})\, p(\boldsymbol{\pi})\, p(\boldsymbol{\mu}|\boldsymbol{\Lambda})\, p(\boldsymbol{\Lambda})\}] + const \\ &= \mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] + \mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{Z}|\boldsymbol{\pi})] + \ln p(\boldsymbol{\pi}) + \ln p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) + const\end{aligned}$$

Since

$$p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1})^{z_{nk}} \qquad \leftarrow \text{from Eq (10.38)}$$

$$\begin{aligned}\therefore\quad \ln q^*(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = &\sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{E}_{\mathbf{Z},(i \neq j)}[z_{nk}] \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}) + \mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{Z}|\boldsymbol{\pi})] \\ &+ \ln p(\pi) + \sum_{k=1}^{K} \ln p(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) + const\end{aligned}$$

$$\text{where} \quad \mathbb{E}_{\mathbf{Z},(i \neq j)}[z_{nk}] = \int z_{nk} \prod_{i \neq j} q_i d\mathbf{Z}_i$$

$$q_j \equiv q_j(\mathbf{Z}_j) \quad \leftarrow \text{q distribution}$$

Eq 10.56: (PRML p.478)

$$\ln q^*(\boldsymbol{\pi}) = (\alpha_0 - 1) \sum_{k=1}^{K} \ln \pi_k + \sum_{k=1}^{K} \sum_{n=1}^{N} r_{nk} \ln \pi_k + const$$

Proof :

$$p(\mathbf{Z}|\boldsymbol{\pi}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \tag{10.37}$$

$$p(\boldsymbol{\pi}) = C(\boldsymbol{\alpha}_0) \prod_{k=1}^{K} \pi_k^{\alpha_0 - 1} \tag{10.39}$$

$$
\begin{aligned}
\ln q^*(\boldsymbol{\pi}) &= \ln p(\boldsymbol{\pi}) + \mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{Z}|\boldsymbol{\pi})] + const \\
&= \mathbb{E}_{\mathbf{Z}}\left[\ln \prod_{n=1}^{N}\prod_{k=1}^{K}\pi_k^{z_{nk}}\right] + \ln\left[C(\boldsymbol{\alpha}_0)\prod_{k=1}^{K}\pi_k^{\alpha_0-1}\right] + const \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{E}_{\mathbf{Z}}[z_{nk}]\ln\pi_k + (\alpha_0-1)\sum_{k=1}^{K}\ln\pi_k + const \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\ln\pi_k + (\alpha_0-1)\sum_{k=1}^{K}\ln\pi_k + const \\
&\text{where} \quad \mathbb{E}_z(z_{nk}) = r_{nk} \text{ is used.} \quad \text{(Eq 10.50)}
\end{aligned}
$$

Eq 10.57: (PRML p.478)

$$q^*(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha})$$

$$\text{where} \quad \alpha_k = \alpha_0 + N_k$$

Proof :

Continuing from Eq (10.56),

$$
\begin{aligned}
\ln q^*(\boldsymbol{\pi}) &= \sum_{k=1}^{K}\left[\sum_{n=1}^{N}\mathbb{E}_{\mathbf{Z}}[z_{nk}]\ln\pi_k + (\alpha_0-1)\ln\pi_k\right] + const \\
&\left(\text{Since } N_k = \sum_{n=1}^{N}\mathbb{E}_{\mathbf{Z}}[z_{nk}] \quad : \text{Eq (10.51)}\right) \\
&= \sum_{k=1}^{K}(N_k + \alpha_0 - 1)\ln\pi_k + const \\
&= \ln\left(\prod_{k=1}^{K}\pi_k^{N_k+\alpha_0-1}\right) + const
\end{aligned}
$$

Referring to Eq (10.39):

$$\mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}_0) = C(\boldsymbol{\alpha}_0)\prod_{k=1}^{K}\pi_k^{\alpha_0-1}$$

$$\therefore \quad q^*(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha})$$

$$\text{where} \quad \alpha_k = \alpha_0 + N_k$$

Eq 10.92: (PRML p.487)

$$\begin{aligned}\ln q^*(\alpha) &= \ln p(\alpha) + \mathbb{E}_w[\ln [(\mathbf{w}|\alpha) + const \\ &= (a_0 - 1)\ln\alpha - b_0\alpha + \frac{M}{2}\ln\alpha - \frac{\alpha}{2}\mathbb{E}[\mathbf{w}^T\mathbf{w}] + const\end{aligned}$$

Proof :

$$q(\mathbf{w}, \alpha) = q(\mathbf{w}) \cdot q(\alpha) \tag{10.91}$$

$\Rightarrow \quad q(\mathbf{w}, \alpha)$ factorizes into $q(\mathbf{w}) \cdot q(\alpha)$

According to Eqs (10.5) $\sim$ (10.9), we can find $q_j(\alpha) = \widetilde{p}(\bar{\mathbf{t}}, \mathbf{w}, \alpha)$, which maximizes the lower bound $\mathcal{L}(q)$.

Using Eqs (10.9) and (10.90),

$$\ln q_j^*(\mathbf{Z}_j) = \mathbb{E}_{i\neq j}[\ln p(\mathbf{X}, \mathbf{Z})] + const \tag{10.9}$$

$$p(\bar{\mathbf{t}}, \mathbf{w}, \alpha) = p(\bar{\mathbf{t}}|\mathbf{w})p(\mathbf{w}|\alpha)p(\alpha) \tag{10.90}$$

$$\ln q^*(\alpha) = \mathbb{E}_{\mathbf{w}}\{\ln [p(\bar{\mathbf{t}}|\mathbf{w})p(\mathbf{w}|\alpha)p(\alpha)]\} + const$$

where the expectation is calculated on w only, since w is the only parameter that corresponds to $i \neq j$ condition in Eq (10.9). (There are only w and α.)

$$\ln q^*(\alpha) = \ln p(\alpha) + \mathbb{E}_{\mathbf{w}}[\ln p(\mathbf{w}|\alpha)] + const$$

where $\mathbb{E}_{\mathbf{w}}[t|\mathbf{w}]$ is absorbed into the constant term, because it is independent of α.

Since

$$\begin{aligned}p(\alpha) &= \mathrm{Gam}(\alpha|a_0, b_0) \\ &= \frac{1}{\Gamma(a_0)} b_0^{a_0}\alpha^{\alpha_0 - 1}\mathrm{e}^{-b_0\alpha}\end{aligned}$$

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

$$\therefore \quad \ln q^*(\alpha) = (a_0 - 1)\ln\alpha - b_0\alpha + \frac{M}{2}\ln\alpha - \frac{\alpha}{2}\mathbb{E}[\mathbf{w}^T\mathbf{w}] + const$$

where

$$\frac{M}{2}\ln\alpha - \frac{\alpha}{2}\mathbb{E}[\mathbf{w}^T\mathbf{w}] \text{ is from } p(\mathbf{w}|\alpha).$$

Eq 10.144: (PRML p.497)

$$\sigma(x) \geq \sigma(\xi)\exp\{(x-\xi)/2 - \lambda(\xi)(x^2-\xi^2)\}$$

Proof :

$$\ln\sigma(x) = \frac{x}{2} + f(x) \tag{10.138}$$

$$(\text{where} \quad f(x) = -\ln(\mathrm{e}^{x/2} + \mathrm{e}^{-x/2})$$

$$f(x) \geq -\lambda x^2 + \lambda\xi^2 - \ln(\mathrm{e}^{\xi/2} + \mathrm{e}^{-\xi/2}) \tag{10.143}$$

From Eqs (10.138) and (10.143),

$$\ln\sigma(x) \geq \frac{x}{2} - \lambda x^2 + \lambda\xi^2 - \ln(\mathrm{e}^{\xi/2} + \mathrm{e}^{-\xi/2})$$

$$\begin{aligned}
\sigma(x) &\geq \exp\left\{\frac{x}{2} - \lambda x^2 + \lambda\xi^2 - \ln(\mathrm{e}^{\xi/2} + \mathrm{e}^{-\xi/2})\right\} \\
&= \frac{1}{\mathrm{e}^{\xi/2} + \mathrm{e}^{-\xi/2}}\exp\left[\frac{x}{2} - \lambda x^2 + \lambda\xi^2\right] \\
&= \frac{\mathrm{e}^{-\xi/2}}{1 + \mathrm{e}^{-\xi}}\exp\left[\frac{x}{2} - \lambda x^2 + \lambda\xi^2\right] \\
&= \sigma(\xi)\exp\left[-\frac{\xi}{2} + \frac{x}{2} - \lambda x^2 + \lambda\xi^2\right] \\
&= \sigma(\xi)\exp\left[-\lambda(x^2-\xi^2) + \frac{1}{2}(x-\xi)\right]
\end{aligned}$$

Eq 10.223: (PRML p.513)

$$p(D) \simeq (2\pi v^{new})^{D/2}\exp(B/2)\prod_{n=1}^{N}\{s_n(2\pi v_n)^{-D/2}\}$$

$$\text{where} \quad B = \frac{(\mathbf{m}^{new})^T\mathbf{m}^{new}}{v} - \sum_n \frac{\mathbf{m}_n^T\mathbf{m}_n}{v_n}$$

Proof :

$$p(D) \simeq \int \prod_i \widetilde{f}_i(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{10.208}$$

$$\begin{aligned} \int \prod_n \widetilde{f}_n(\boldsymbol{\theta}) d\boldsymbol{\theta} &= \int \prod_n s_n \mathcal{N}(\boldsymbol{\theta}|\mathbf{m}_n, v_n \mathbf{I}) d\boldsymbol{\theta} \qquad \text{(from Eq (10.213))} \\ &= \prod_n s_n \int \mathcal{N}(\boldsymbol{\theta}|\mathbf{m}_n, v_n \mathbf{I}) d\boldsymbol{\theta} \qquad (1) \end{aligned}$$

where

$$\frac{1}{v_n} = \frac{1}{v^{new}} - \frac{1}{v^{\backslash n}} \tag{2}$$

$$\mathbf{m}_n = \mathbf{m}^{\backslash n} + \frac{v^{\backslash n}}{v_n + v^{\backslash n}} (\mathbf{m}^{new} - \mathbf{m}^{\backslash n}) \tag{3}$$

$$s_n = \frac{z_n}{(2\pi v_n)^{D/2} \mathcal{N}(\mathbf{m}_n | \mathbf{m}^{\backslash n}, (v_n + v^{\backslash n})\mathbf{I})} \tag{4}$$

$$\begin{aligned} \prod_n \mathcal{N}(\boldsymbol{\theta}|\mathbf{m}_n, v_n \mathbf{I})) &= \prod_n \frac{1}{(2\pi v_n)^{D/2}} \exp\left[-\frac{1}{2v_n} (\boldsymbol{\theta} - \mathbf{m}_n)^T \cdot (\boldsymbol{\theta} - \mathbf{m}_n) \right] \\ &= \prod_n \frac{1}{(2\pi v_n)^{D/2}} \exp\left[\sum_n \left(-\frac{1}{2v_n} \right) (\boldsymbol{\theta}^2 - 2\mathbf{m}_n \boldsymbol{\theta} + \mathbf{m}_n^2) \right] \\ &= \prod_n \frac{1}{(2\pi v_n)^{D/2}} \exp\left\{ \sum_n \left[-\frac{1}{2} \sum_n \left(\frac{1}{v_n} \right) \boldsymbol{\theta}^2 + \left(\sum_n \frac{\mathbf{m}_n}{v_n} \right) \boldsymbol{\theta} \right. \right. \\ &\quad \left. \left. -\frac{1}{2} \sum \frac{\mathbf{m}_n^2}{v_n} \right] \right\} \end{aligned}$$

From Eq (2),

$$\frac{1}{v^{new}} = \frac{1}{v_n} + \frac{1}{v^{\backslash n}} = \sum_n \frac{1}{v_n}$$

And let's define a new $\mathbf{m}^{new}$ (different from Eq (3))

$$\frac{\mathbf{m}^{new}}{v^{new}} = \sum_n \frac{\mathbf{m}_n}{v_n}$$

$$\prod_n \mathcal{N}(\boldsymbol{\theta}|\mathbf{m}_n, v_n\mathbf{I})) = \prod_n \frac{1}{(2\pi v_n)^{D/2}} \exp\left[-\frac{1}{2v^{new}}\boldsymbol{\theta}^2 + \frac{\mathbf{m}^{new}}{v^{new}}\boldsymbol{\theta} - \frac{1}{2}\sum_n \frac{\mathbf{m}_n^2}{v_n}\right]$$
$$= \prod_n \frac{1}{(2\pi v_n)^{D/2}} \exp\left[-\frac{1}{2v^{new}}(\boldsymbol{\theta}^2 - 2\mathbf{m}^{new}\boldsymbol{\theta}) - \frac{1}{2}\sum_n \frac{\mathbf{m}_n^2}{v_n}\right]$$
$$= \prod_n \frac{1}{(2\pi v_n)^{D/2}} \exp\left[-\frac{1}{2v^{new}}(\boldsymbol{\theta} - \mathbf{m}^{new})^2 + \frac{1}{2v^{new}}(\mathbf{m}^{new})^2 - \frac{1}{2}\sum_n \frac{\mathbf{m}_n^2}{v_n}\right]$$

Therefore Eq (1) becomes

$$\int \prod_n \widetilde{f}_n(\boldsymbol{\theta})d\boldsymbol{\theta} = \prod_n s_n \int \mathcal{N}(\boldsymbol{\theta}|\mathbf{m}_n, v_n\mathbf{I})d\boldsymbol{\theta}$$
$$= \prod_n \left[s_n \frac{1}{(2\pi v_n)^{D/2}}\right] \exp\left[\frac{1}{2v^{new}}(\mathbf{m}^{new})^2 - \frac{1}{2}\sum_n \frac{\mathbf{m}_n^2}{v_n}\right] \cdot \int \exp\left[-\frac{1}{2v^{new}}(\boldsymbol{\theta} - \mathbf{m}^{new})^2\right] d\boldsymbol{\theta}$$
$$= \prod_n \left[\frac{s_n}{(2\pi v_n)^{D/2}}\right] \exp\left[\frac{1}{2}\left(\frac{(\mathbf{m}^{new})^2}{v^{new}} - \sum_n \frac{\mathbf{m}_n^2}{v_n}\right)\right] \cdot (2\pi v^{new})^{D/2}$$

Chapter 12. Continuous Latent Variables

Eq 12.12 & 12.13: (PRML p.564)

$$z_{nj} = \mathbf{x}_n^T\mathbf{u}_j, \qquad b_j = \overline{\mathbf{x}}^T\mathbf{u}_j$$

Proof :

$$J = \frac{1}{N}\sum_{n=1}^{N} \|\mathbf{x}_n - \widetilde{\mathbf{x}}_n\|^2 \tag{12.11}$$

$$\widetilde{\mathbf{x}}_n = \sum_{i=1}^{M} z_{ni}\mathbf{u}_i + \sum_{i=M+1}^{D} b_i\mathbf{u}_i \tag{12.10}$$

$$\begin{aligned}
J &= \frac{1}{N}\sum_{n=1}^{N}\|\mathbf{x}_n - \sum_{i=1} z_{ni}\mathbf{u}_i - \sum_{i=M+1}^{D} b_i\mathbf{u}_i\|^2 \\
&= \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n^T - \sum_{i=1}^{M} z_{ni}\mathbf{u}_i^T - \sum_{i=M+1}^{D} b_i\mathbf{u}_i^T)\cdot(\mathbf{x}_n - \sum_{i=1}^{M} z_{ni}\mathbf{u}_i - \sum_{i=M+1}^{D} b_i\mathbf{u}_i) \\
\frac{\partial J}{\partial z_{ni}} &= \frac{1}{N}\sum_{n=1}^{N}\Bigg[(-\mathbf{u}_i^T)(\mathbf{x}_n - \sum_{i=1}^{M} z_{ni}\mathbf{u}_i - \sum_{i=M+1}^{D} b_i\mathbf{u}_i) \\
&\qquad + (-\mathbf{u}_i)(\mathbf{x}_n^T - \sum_{i=1}^{M} z_{ni}\mathbf{u}_i^T - \sum_{i=M+1}^{D} b_i\mathbf{u}_i)\Bigg] \\
\Rightarrow &\quad \mathbf{x}_n^T - \sum_{i=1}^{M} z_{ni}\mathbf{u}_i^T - \sum_{i=M+1}^{D} b_i\mathbf{u}_i = 0 \qquad (1)
\end{aligned}$$

Multiplying Eq (1) by $\mathbf{u}_j$ (j = 1 $\sim$ M)

$$\mathbf{x}_n^T\mathbf{u}_j - \sum_{i=1}^{M} z_{ni}\mathbf{u}_i^T\mathbf{u}_j - \sum_{i=M+1}^{D} b_i\mathbf{u}_i\mathbf{u}_j = 0$$

Since $\mathbf{u}_i^T\mathbf{u}_j = \delta_{ij}$ in the second term and $\mathbf{u}_i\mathbf{u}_j = 0$ in the last term,

$$\therefore \;\; z_{nj} = \mathbf{x}_n^T\mathbf{u}_j$$

To obtain Eq (12.13), multiply Eq (1) by $\mathbf{u}_j$, where j = M+1 $\sim$ D.

$$\frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n^T\mathbf{u}_j - \sum_{i=1}^{M} z_{ni}\mathbf{u}_i^T\mathbf{u}_j - \sum_{i=M+1}^{D} b_i\mathbf{u}_i\mathbf{u}_j) = 0$$

Since $\mathbf{u}_i^T\mathbf{u}_j = 0$ in the second term and $\mathbf{u}_i\mathbf{u}_j = \delta_{ij}$ in the last term,

$$\overline{\mathbf{x}}^T\mathbf{u}_j - \frac{1}{N}\sum_{n=1}^{N} b_j = 0$$

$$\therefore \;\; b_j = \overline{\mathbf{x}}^T\mathbf{u}_j$$

Eq 12.23: (PRML p.567)

$$\mathbf{SU} = \mathbf{UL}$$

Proof :

$$\mathbf{S} \;=\; \begin{pmatrix} \mathbf{s}_1 \\ \mathbf{s}_2 \\ \vdots \\ \mathbf{s}_D \end{pmatrix} \qquad\qquad \mathbf{U} \;=\; \Big(\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_D\Big)$$

where $\mathbf{s}$ is a row vector, and $\mathbf{u}$ a column vector as usual.

$$\mathbf{L} = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_D \end{pmatrix}$$

We know that

$$\mathbf{S}\mathbf{u}_1 = \lambda_1 \mathbf{u}_1$$

$$\mathbf{S}\mathbf{u}_2 = \lambda_2 \mathbf{u}_2$$

$$\vdots$$

$$\mathbf{SU} \;=\; \begin{pmatrix} \mathbf{s}_1 \\ \mathbf{s}_2 \\ \vdots \\ \mathbf{s}_D \end{pmatrix} \Big(\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_D\Big) \;=\; \begin{pmatrix} \mathbf{s}_1\mathbf{u}_1 & \mathbf{s}_1\mathbf{u}_2 & \cdots & \mathbf{s}_1\mathbf{u}_D \\ \mathbf{s}_2\mathbf{u}_1 & \mathbf{s}_2\mathbf{u}_2 & \cdots & \mathbf{s}_2\mathbf{u}_D \\ \vdots & & & \\ \mathbf{s}_D\mathbf{u}_1 & \mathbf{s}_D\mathbf{u}_2 & \cdots & \mathbf{s}_D\mathbf{u}_D \end{pmatrix}$$

$$= \Big(\mathbf{S}\mathbf{u}_1, \mathbf{S}\mathbf{u}_2, \cdots, \mathbf{S}\mathbf{u}_D\Big)$$

$$\mathbf{UL} \;=\; \Big(\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_D\Big) \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_D \end{pmatrix} \;=\; \begin{pmatrix} \lambda_1 u_{11} & \lambda_2 u_{21} & \cdots & \lambda_D u_{D1} \\ \lambda_1 u_{12} & \lambda_2 u_{22} & \cdots & \lambda_D u_{D2} \\ \vdots & & & \\ \lambda_1 u_{1D} & \lambda_2 u_{2D} & \cdots & \lambda_D u_{DD} \end{pmatrix}$$

$$= \Big(\lambda_1 \mathbf{u}_1, \lambda_2 \mathbf{u}_2, \cdots, \lambda_D \mathbf{u}_D\Big)$$

$$\therefore \quad \mathbf{SU} = \mathbf{UL}$$

Eq 12.30: (PRML p.570)

$$\mathbf{u}_i = \frac{1}{N\lambda_i}\mathbf{X}^T\mathbf{v}_i$$

There must be a typo: $(N\lambda_i)^{1/2}$ should be $N\lambda_i$.

Proof :

We know that $\|\mathbf{u}_i\| = 1$ and $\mathbf{v}_i$ is not normalized.

Using $\mathbf{v}_i = \mathbf{X}\mathbf{u}_i$,

$$\begin{aligned}
\mathbf{v}_i^T\mathbf{v}_i &= \mathbf{u}_i^T\mathbf{X}^T\mathbf{X}\mathbf{u}_i \\
&\text{(using Eq (12.26))} \\
&= \mathbf{u}_i^T(N\lambda_i\mathbf{u}_i) \\
&= N\lambda_i\mathbf{u}_i^T\mathbf{u}_i \\
&= N\lambda_i
\end{aligned}$$

Since $\mathbf{u}_i \propto \mathbf{X}^T\mathbf{v}_i$, and $\mathbf{u}_i = c\mathbf{X}^T\mathbf{v}_i$.

Let's determine c.

$$\begin{aligned}
\mathbf{u}_i^T\mathbf{u}_i &= c\mathbf{v}_i^T\mathbf{X}(c\mathbf{X}^T\mathbf{v}_i) \\
&= c^2\mathbf{v}_i^T(\mathbf{X}\mathbf{X}^T\mathbf{v}_i) \\
&\text{(using Eq (12.28))} \\
&= c^2\mathbf{v}_i^T(N\lambda_i\mathbf{v}_i) \\
&= c^2N\lambda_i\|\mathbf{v}_i\|^2
\end{aligned}$$

$$\Rightarrow \quad c^2 = \frac{1}{(N\lambda_i)^2}$$

$$\therefore \quad \mathbf{u}_i = \frac{1}{N\lambda_i}\mathbf{X}^T\mathbf{v}_i$$

Eq 12.40: (PRML p.573)

$$\mathbf{C}^{-1} = \sigma^{-2}\mathbf{I} - \sigma^{-2}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^T$$
$$\text{where } \mathbf{M} = \mathbf{W}^T\mathbf{W} + \sigma^2\mathbf{I}$$

Proof :

$$\mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I} \tag{12.36}$$

Woodbury identity:

$$(\mathbf{A} + \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{D} + \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1}$$

Comparing to Eq (12.36),

$$\mathbf{A} \Leftrightarrow \sigma^2\mathbf{I}, \quad \mathbf{B} \Leftrightarrow \mathbf{W}, \quad \mathbf{D}^{-1} \Leftrightarrow \mathbf{I}, \quad \mathbf{C} \Leftrightarrow \mathbf{W}^T$$

$$\begin{aligned}
\mathbf{C}^{-1} &= \sigma^{-2}\mathbf{I} - \sigma^{-2}\mathbf{I}\mathbf{W}(\mathbf{I} + \mathbf{W}^T\sigma^{-2}\mathbf{I}\mathbf{W})^{-1}\mathbf{W}^T\sigma^{-2}\mathbf{I} \\
&= \sigma^{-2}\mathbf{I} - \sigma^{-2}\mathbf{W}(\mathbf{I} + \mathbf{W}^T\mathbf{W}\sigma^{-2})^{-1}\mathbf{W}^T\sigma^{-2} \\
&= \sigma^{-2}\mathbf{I} - \sigma^{-2}\mathbf{W}(\sigma^2\mathbf{I} + \mathbf{W}^T\mathbf{W})^{-1}\sigma^2\mathbf{W}^T\sigma^{-2} \\
&\text{(since } \sigma^2\mathbf{I} + \mathbf{W}^T\mathbf{W} = \mathbf{M}) \\
&= \sigma^{-2} - \sigma^{-2}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^T
\end{aligned}$$

Eq 12.44: (PRML p.574)

$$\ln p(\mathbf{X}|\mathbf{W}, \boldsymbol{\mu}, \sigma^2) = -\frac{N}{2}\{D\ln(2\pi) + \ln|\mathbf{C}| + \mathrm{Tr}(\mathbf{C}^{-1}\mathbf{S})\}$$

Proof :

Starting from Eq (12.43),

$$\ln p(\mathbf{X}|\mathbf{W}, \boldsymbol{\mu}, \sigma^2) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\mathbf{C}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^T\mathbf{C}^{-1}(\mathbf{x}_n - \boldsymbol{\mu})$$

To prove Eq (12.44), all we have to do is to show

$$\sum_{n=1}^{N}(\mathbf{x}_n - \overline{\mathbf{x}})^T\mathbf{C}^{-1}(\mathbf{x}_n - \overline{\mathbf{x}}) = N\,\mathrm{Tr}(\mathbf{C}^{-1}\mathbf{S})$$
$$\text{where } \mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \overline{\mathbf{x}})(\mathbf{x}_n - \overline{\mathbf{x}})^T$$

To simplify the proof, let's show

$$\sum_{n=1}^{N} \mathbf{x}_n^T \mathbf{A} \mathbf{x}_n = N \operatorname{Tr}(\mathbf{AT})$$
$$\text{where } \mathbf{T} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^T$$

$$\sum_{n=1}^{N} \mathbf{x}_n^T \mathbf{A} \mathbf{x}_n \text{ has } 1 \times 1 \text{ dimension.}$$

$$\because \quad \mathbf{x}_n^T : 1 \times N, \quad \mathbf{A}\mathbf{x}_n : N \times 1 \quad \longrightarrow (1 \times N) \cdot (N \times 1) = 1 \times 1$$

$$\mathbf{AT} = \mathbf{A} \sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^T = \sum_{n=1}^{N} \mathbf{A} \mathbf{x}_n \mathbf{x}_n^T \quad \Rightarrow N \times N$$

Let's show n = 1 case,

$$\mathbf{A}\mathbf{x}_1\mathbf{x}_1^T = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1N} \\ A_{21} & A_{22} & \cdots & A_{2N} \\ \vdots & & & \\ A_{N1} & A_{N2} & \cdots & A_{NN} \end{pmatrix} \begin{pmatrix} x_{11} \\ x_{12} \\ \vdots \\ x_{1N} \end{pmatrix} \begin{pmatrix} x_{11}, x_{12}, \cdots, x_{1N} \end{pmatrix}$$

$$= \begin{pmatrix} A_{11}x_{11} + A_{12}x_{12} + \cdots + A_{1N}x_{1N} \\ A_{21}x_{11} + A_{22}x_{12} + \cdots + A_{2N}x_{1N} \\ \vdots \\ A_{N1}x_{11} + A_{N2}x_{12} + \cdots + A_{NN}x_{1N} \end{pmatrix} \begin{pmatrix} x_{11}, x_{12}, \cdots, x_{1N} \end{pmatrix} \tag{1}$$

Diagonal elements are

$$(1,1) = A_{11}x_{11}^2 + A_{12}x_{11}x_{12} + \cdots + A_{1N}x_{11}x_{1N}$$
$$(2,2) = A_{21}x_{11}x_{12} + A_{22}x_{12}^2 + \cdots + A_{2N}x_{12}x_{1N}$$
$$\vdots$$

$$\mathbf{x}_1^T \mathbf{A} \mathbf{x}_1 = \begin{pmatrix} x_{11}, x_{12}, \cdots, x_{1N} \end{pmatrix} \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1N} \\ A_{21} & A_{22} & \cdots & A_{2N} \\ \vdots & & & \\ A_{N1} & A_{N2} & \cdots & A_{NN} \end{pmatrix} \begin{pmatrix} x_{11} \\ x_{12} \\ \vdots \\ x_{1N} \end{pmatrix}$$

$$= \begin{pmatrix} x_{11}, x_{12}, \cdots, x_{1N} \end{pmatrix} \begin{pmatrix} A_{11}x_{11} + A_{12}x_{12} + \cdots + A_{1N}x_{1N} \\ A_{21}x_{11} + A_{22}x_{12} + \cdots + A_{2N}x_{1N} \\ \vdots \\ A_{N1}x_{11} + A_{N2}x_{12} + \cdots + A_{NN}x_{1N} \end{pmatrix}$$

Compared to Eq (1), this is just the diagonal terms in Eq (1).

$$\Rightarrow \quad \mathbf{x}_1^T \mathbf{A} \mathbf{x}_1 = \mathrm{Tr}(\mathbf{A} \mathbf{x}_1 \mathbf{x}_1^T)$$

For n = 2 case,

$$\mathbf{x}_2^T \mathbf{A} \mathbf{x}_2 = \mathrm{Tr}(\mathbf{A} \mathbf{x}_2 \mathbf{x}_2^T)$$
$$\vdots$$

For n = N case,

$$\mathbf{x}_N^T \mathbf{A} \mathbf{x}_N = \mathrm{Tr}(\mathbf{A} \mathbf{x}_N \mathbf{x}_N^T)$$

$$\begin{aligned} \Rightarrow \quad \sum_{n=1}^{N} \mathbf{x}_n \mathbf{A} \mathbf{x}_n &= \sum_{n=1}^{N} \mathrm{Tr}(\mathbf{A} \mathbf{x}_n \mathbf{x}_n^T) \\ &= \mathrm{Tr}\left[\sum_{n=1}^{N} \mathbf{A} \mathbf{x}_n \mathbf{x}_n^T \right] \\ &= \mathrm{Tr}\left[\mathbf{A} \sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^T \right] \\ &= \mathrm{Tr}\left[\mathbf{A} N \mathbf{T} \right] \\ &= N \mathrm{Tr}\left[\mathbf{A} \mathbf{T} \right] \end{aligned}$$

Eq 12.53: (PRML p.578)

$$\mathbb{E}[\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\mu},\mathbf{W},\sigma^2)] = -\sum_{n=1}^{N}\left\{\frac{D}{2}\ln(2\pi\sigma^2) + \frac{1}{2}\mathrm{Tr}(\mathbb{E}[\mathbf{z}_n\mathbf{z}_n^T]) + \frac{1}{2\sigma^2}\|\mathbf{x}_n-\boldsymbol{\mu}\|^2 \right.$$
$$\left. -\frac{1}{\sigma^2}\mathbb{E}[\mathbf{z}_n]^T\cdot\mathbf{W}^T(\mathbf{x}_n-\boldsymbol{\mu}) + \frac{M}{2}\ln(2\pi) + \frac{1}{2\sigma^2}\mathrm{Tr}(\mathbb{E}\left[\mathbf{z}_n\mathbf{z}_n^T\right]\mathbf{W}^T\mathbf{W})\right\}$$

Proof :

$$\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\mu},\mathbf{W},\sigma^2) = \sum_n\{\ln p(\mathbf{x}_n|\mathbf{z}_n) + \ln p(\mathbf{z}_n)\} \tag{12.52}$$

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0},\mathbf{I}) \tag{12.31}$$

$$p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|\mathbf{W}\mathbf{z}+\boldsymbol{\mu},\sigma^2\mathbf{I}) \tag{12.32}$$

$$\mathcal{N}(\mathbf{x}_n|\mathbf{W}\mathbf{z}_n+\boldsymbol{\mu},\sigma^2\mathbf{I}) = \frac{1}{(2\pi)^{D/2}\|\sigma^2\mathbf{I}\|^{1/2}}\exp\left\{-\frac{1}{2\sigma^2\mathbf{I}}(\mathbf{x}_n-\mathbf{W}\mathbf{z}_n-\boldsymbol{\mu})^T\right.$$
$$\left.\cdot(\mathbf{x}_n-\mathbf{W}\mathbf{z}_n-\boldsymbol{\mu})\right\} \tag{1}$$

$$\mathcal{N}(\mathbf{z}_n|\mathbf{0},\mathbf{I}) = \frac{1}{(2\pi)^{M/2}}\cdot\frac{1}{\|\mathbf{I}\|^{1/2}}\cdot\exp\left\{-\frac{1}{2}\mathbf{z}_n^T\mathbf{z}_n\right\} \tag{2}$$

$$\begin{aligned}\{\cdots\}\text{ in Eq (1) } &= -\frac{1}{2\sigma^2}(\mathbf{x}_n-\mathbf{W}\mathbf{z}_n-\boldsymbol{\mu})^T\cdot(\mathbf{x}_n-\mathbf{W}\mathbf{z}_n-\boldsymbol{\mu})\\ &= -\frac{1}{2\sigma^2}(\mathbf{x}_n-\boldsymbol{\mu})^2 - \frac{1}{2\sigma^2}\mathbf{z}_n^T\mathbf{W}^T\mathbf{W}\mathbf{z}_n + 2\frac{1}{2\sigma^2}\mathbf{z}_n^T\mathbf{W}^T(\mathbf{x}_n-\boldsymbol{\mu})\end{aligned} \tag{3}$$

Putting Eqs (1), (2), and (3) together,

$$\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\mu},\mathbf{W},\sigma^2) = -\sum_n\left\{\frac{D}{2}\ln(2\pi\sigma^2) + \frac{M}{2}\ln(2\pi) + \frac{1}{2\sigma^2}\|\mathbf{x}_n-\boldsymbol{\mu}\|^2\right.$$
$$\left.+\frac{1}{2\sigma^2}\mathbf{z}_n^T\mathbf{W}^T\mathbf{W}\mathbf{z}_n - \frac{1}{\sigma^2}\mathbf{z}_n^T\mathbf{W}^T(\mathbf{x}_n-\boldsymbol{\mu}) + \frac{1}{2}\mathbf{z}_n^T\mathbf{z}_n\right\}$$

$$\mathbb{E}[\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\mu},\mathbf{W},\sigma^2)] = -\sum_n\left\{\frac{D}{2}\ln(2\pi\sigma^2) + \frac{M}{2}\ln(2\pi) + \frac{1}{2\sigma^2}\|\mathbf{x}_n-\boldsymbol{\mu}\|^2\right.$$
$$\left.+\frac{1}{2\sigma^2}\mathbb{E}[\mathbf{z}_n^T\mathbf{W}^T\mathbf{W}\mathbf{z}_n] - \frac{1}{\sigma^2}\mathbb{E}[\mathbf{z}_n^T]\mathbf{W}^T(\mathbf{x}_n-\boldsymbol{\mu}) + \frac{1}{2}\mathbb{E}[\mathbf{z}_n^T\mathbf{z}_n]\right\} \tag{4}$$

The expectation is done over the posterior distribution, which is $p(\mathbf{Z}|\mathbf{X},\boldsymbol{\mu},\mathbf{W},\sigma^2)$. It is a function of $\mathbf{Z}$, so the above $\mathbb{E}$ is over $\mathbf{z}_n$ only.

Now let s take a look at $\mathbb{E}[\mathbf{z}_n^T\mathbf{z}_n]$ and $\mathbb{E}[\mathbf{z}_n^T\mathbf{W}^T\mathbf{W}\mathbf{z}_n]$.

$$\mathbb{E}[\mathbf{z}_n^T\mathbf{z}_n] = \mathrm{Tr}(\mathbb{E}[\mathbf{z}_n\mathbf{z}_n^T]) \tag{5}$$

For example,

$$\begin{pmatrix} 1 & 2 \end{pmatrix} \begin{pmatrix} 3 \\ 4 \end{pmatrix} = 3 + 8 = 11$$

$$\begin{pmatrix} 1 \\ 2 \end{pmatrix} \begin{pmatrix} 3 & 4 \end{pmatrix} = \begin{pmatrix} 3 & 4 \\ 6 & 8 \end{pmatrix} \Rightarrow \mathrm{Tr} \begin{pmatrix} 3 & 4 \\ 6 & 8 \end{pmatrix} = 11$$

Utilizing the following relationship,

$$\mathbb{E}[\mathbf{x}^T(\mathbf{A}\mathbf{x})] = \mathrm{Tr}(\mathbb{E}[\mathbf{x}(\mathbf{A}\mathbf{x})^T]) = \mathrm{Tr}(\mathbb{E}[\mathbf{x}\mathbf{x}^T\mathbf{A}^T])$$

$$\begin{aligned} \mathbb{E}[\mathbf{z}_n^T(\mathbf{W}^T\mathbf{W})\mathbf{z}_n] &= \mathrm{Tr}(\mathbb{E}[\mathbf{z}_n(\mathbf{W}^T\mathbf{W}\mathbf{z}_n)^T]) \\ &= \mathrm{Tr}(\mathbb{E}[\mathbf{z}_n\mathbf{z}_n^T\mathbf{W}^T\mathbf{W}]) \\ &= \mathrm{Tr}(\mathbb{E}[\mathbf{z}_n\mathbf{z}_n^T] \cdot \mathbf{W}^T\mathbf{W}) \end{aligned} \tag{6}$$

Plugging Eqs (5) and (6) into Eq (4) gives Eq (12.53).

Eq 12.63: (PRML p.585)

$$\mathbf{W}_{new} = \left[\sum_{n=1}^{N}(\mathbf{x}_n - \overline{\mathbf{x}})\mathbb{E}[\mathbf{z}_n]^T\right]\left[\sum_{n=1}^{N}\mathbb{E}[\mathbf{z}_n\mathbf{z}_n^T] + \sigma^2\mathbf{A}\right]^{-1}$$

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

Proof :

According to Eq (8.8)

$$p(\widehat{t}, \overline{\mathbf{t}}, \mathbf{w}|\widehat{x}, \overline{\mathbf{x}}, \alpha, \sigma^2) = \left[\prod_{n=1}^{N} p(t_n|\mathbf{x}_n, \mathbf{w}, \sigma^2\right] p(\mathbf{w}|\alpha) \cdot p(\widehat{t}|\widehat{x}, \mathbf{w}, \sigma^2)$$

where $\widehat{x}$ is a new input and $\widehat{t}$ is the corresponding target.

With reference to the relationship shown in Fig (8.6) between the graph and the joint probability, Fig (12.13) gives

$$p(\mathbf{X}, \mathbf{Z}, \mathbf{W}|\boldsymbol{\mu}, \sigma^2, \alpha) = \prod_n p(\mathbf{x}_n|\mathbf{z}_n) \cdot \prod_n p(\mathbf{z}_n) \cdot p(\mathbf{W}|\alpha)$$

Using the equations derived in Eq (12.53),

$$\ln p(\mathbf{X}, \mathbf{Z}, \mathbf{W}|\boldsymbol{\mu}, \sigma^2, \alpha) = \sum_n \{\ln p(\mathbf{x}_n|\mathbf{z}_n) + \ln p(\mathbf{z}_n)\} + \ln p(\mathbf{W}|\alpha)$$

where

$$
\begin{aligned}
p(\mathbf{W}|\boldsymbol{\alpha}) &= \mathcal{N}(\mathbf{W}|\mathbf{0},\boldsymbol{\alpha}) \\
&= \frac{\|\mathbf{A}\|^{1/2}}{(2\pi)^{M/2}} \exp\{-\frac{1}{2}\mathbf{W}^T\mathbf{A}\mathbf{W}\} \\
\mathbf{A} &= \mathrm{diag}(\alpha_i)
\end{aligned}
$$

With the additional term of $p(\mathbf{W}|\boldsymbol{\alpha})$, Eq (12.53) becomes

$$
\begin{aligned}
\mathbb{E}[\ln p(\mathbf{X},\mathbf{Z},\mathbf{W}|\boldsymbol{\mu},\sigma^2,\alpha)] = -\sum_{n=1}^{N} \Bigg\{ & \frac{D}{2}\ln(2\pi\sigma^2) + \frac{1}{2}\mathrm{Tr}(\mathbb{E}[\mathbf{z}_n\mathbf{z}_n^T]) + \frac{1}{2\sigma^2}\|\mathbf{x}_n - \boldsymbol{\mu}\|^2 \\
& -\frac{1}{\sigma^2}\mathbb{E}[\mathbf{z}_n]^T \cdot \mathbf{W}^T(\mathbf{x}_n - \boldsymbol{\mu}) + \frac{M}{2}\ln(2\pi) + \frac{1}{2\sigma^2}\mathrm{Tr}(\mathbb{E}[\mathbf{z}_n\mathbf{z}_n^T] \cdot \mathbf{W}^T\mathbf{W}) \Bigg\} \\
& + \frac{1}{2}\ln\|\mathbf{A}\| - \frac{M}{2}\ln(2\pi) - \frac{1}{2}\mathbf{W}^T\mathbf{A}\mathbf{W}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial}{\partial \mathbf{W}}\mathbb{E}[\ln p(\mathbf{X},\mathbf{Z},\mathbf{W}|\boldsymbol{\mu},\sigma^2,\alpha)] &= \sum_{n=1}^{N}\left\{\frac{1}{\sigma^2}(\mathbf{x}_n - \overline{\mathbf{x}})\mathbb{E}[\mathbf{z}_n]^T - \frac{1}{\sigma^2}\mathbf{W}\mathbb{E}[\mathbf{z}_n\mathbf{z}_n^T]\right\} - \mathbf{W}\mathbf{A} \\
&= 0
\end{aligned}
$$

$$
\begin{aligned}
\Rightarrow \quad \mathbf{W}\left[\sum_n \frac{1}{\sigma^2}\mathbb{E}[\mathbf{z}_n\mathbf{z}_n^T] + \mathbf{A}\right] &= \sum_n \left\{\frac{1}{\sigma^2}(\mathbf{x}_n - \overline{\mathbf{x}})\mathbb{E}[\mathbf{z}_n]^T\right\} \\
\therefore \quad \mathbf{W} &= \left[\sum_{n=1}^{N}(\mathbf{x}_n - \overline{\mathbf{x}})\mathbb{E}[\mathbf{z}_n]^T\right]\left[\sum_{n=1}^{N}\mathbb{E}[\mathbf{z}_n\mathbf{z}_n^T] + \sigma^2\mathbf{A}\right]^{-1}
\end{aligned}
$$

Eq 12.65: (PRML p.585)

$$
p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\mathbf{C})
$$
$$
\text{where} \quad \mathbf{C} = \mathbf{W}\mathbf{W}^T + \boldsymbol{\Psi}
$$

Proof :

Making use of Marginal / Conditional Gaussians,

Current:

$$
p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0},\mathbf{I}) \tag{12.31}
$$

$$
p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|\mathbf{W}\mathbf{z} + \boldsymbol{\mu},\boldsymbol{\Psi}) \tag{12.64}
$$

Marginal/Conditional Gaussians	Current
x	z
y	x
μ	0
Λ^{-1}	I
A	W
b	μ
L^{-1}	Ψ

By making substitution in Eq (2.115),

$$\begin{aligned}\Rightarrow \quad p(\mathbf{x}) &= \mathcal{N}(\mathbf{x}|\mathbf{W}\cdot\mathbf{0}+\boldsymbol{\mu}, \boldsymbol{\Psi}+\mathbf{W}\mathbf{I}\mathbf{W}^T) \\ &= \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Psi}+\mathbf{W}\mathbf{W}^T)\end{aligned}$$

Eq 12.79: (PRML p.588)

$$\mathbf{K}^2\mathbf{a}_i = \lambda_i N \mathbf{K}\mathbf{a}_i$$

Proof :

$$\frac{1}{N}\sum_{n=1}^{N} k(\mathbf{x}_l, \mathbf{x}_n)\sum_{m=1}^{N} a_{im} k(\mathbf{x}_n, \mathbf{x}_m) = \lambda_i \sum_{n=1}^{N} a_{in} k(\mathbf{x}_l, \mathbf{x}_n) \tag{12.78}$$

$$\begin{aligned}&\sum_n k(\mathbf{x}_l, \mathbf{x}_n)\sum_m a_{im} k(\mathbf{x}_n, \mathbf{x}_m) \\ &= \sum_n k(x_l, x_n)[a_{i1}k(x_n, x_1) + a_{i2}k(x_n, x_2) + \cdots + a_{iN}k(x_n, x_N)] \\ &= k(x_l, x_1)[a_{i1}k(x_1, x_1) + a_{i2}k(x_1, x_2) + \cdots + a_{iN}k(x_1, x_N)] \\ &\quad + k(x_l, x_2)[a_{i1}k(x_2, x_1) + a_{i2}k(x_2, x_2) + \cdots + a_{iN}k(x_2, x_N)] \\ &\quad + \cdots \\ &\quad + k(x_l, x_N)[a_{i1}k(x_N, x_1) + a_{i2}k(x_N, x_2) + \cdots + a_{iN}k(x_N, x_N)] \\ &= a_{i1}[k(x_l, x_1)\cdot k(x_1, x_1) + k(x_l, x_2)\cdot k(x_2, x_1) + \cdots + k(x_l, x_N)\cdot k(x_N, x_1)] \\ &\quad + a_{i2}[k(x_l, x_1)\cdot k(x_1, x_2) + k(x_l, x_2)\cdot k(x_2, x_2) + \cdots + k(x_l, x_N)\cdot k(x_N, x_2)] \\ &\quad + \cdots \\ &\quad + a_{iN}[k(x_l, x_1)\cdot k(x_1, x_N) + k(x_l, x_2)\cdot k(x_2, x_N) + \cdots + k(x_l, x_N)\cdot k(x_N, x_N)] \quad (1)\end{aligned}$$

Now let's calculate $\mathbf{K} \cdot \mathbf{K}$,

$$\mathbf{K} = \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & \cdots & k(x_1, x_N) \\ k(x_2, x_1) & k(x_2, x_2) & \cdots & k(x_2, x_N) \\ \vdots & & & \\ k(x_N, x_1) & k(x_N, x_2) & \cdots & k(x_N, x_N) \end{bmatrix}$$

$$K^2_{11} = k(x_1, x_1) \cdot k(x_1, x_1) + k(x_1, x_2) \cdot k(x_2, x_1) + \cdots + k(x_1, x_N) \cdot k(x_N, x_1)$$

$$K^2_{12} = k(x_1, x_1) \cdot k(x_1, x_2) + k(x_1, x_2) \cdot k(x_2, x_2) + \cdots + k(x_1, x_N) \cdot k(x_N, x_2)$$

$$\vdots$$

$$K^2_{21} = k(x_2, x_1) \cdot k(x_1, x_1) + k(x_2, x_2) \cdot k(x_2, x_1) + \cdots + k(x_2, x_N) \cdot k(x_N, x_1)$$

$$K^2_{22} = k(x_2, x_1) \cdot k(x_1, x_2) + k(x_2, x_2) \cdot k(x_2, x_2) + \cdots + k(x_2, x_N) \cdot k(x_N, x_2)$$

$$\vdots$$

$$K^2_{N1} = k(x_N, x_1) \cdot k(x_1, x_1) + k(x_N, x_2) \cdot k(x_2, x_1) + \cdots + k(x_N, x_N) \cdot k(x_N, x_1)$$

$$K^2_{N2} = k(x_N, x_1) \cdot k(x_1, x_2) + k(x_N, x_2) \cdot k(x_2, x_2) + \cdots + k(x_N, x_N) \cdot k(x_N, x_2)$$

$$\vdots$$

$$K^2_{NN} = k(x_N, x_1) \cdot k(x_1, x_N) + k(x_N, x_2) \cdot k(x_2, x_N) + \cdots + k(x_N, x_N) \cdot k(x_N, x_N)$$

$$\mathbf{K}^2 \cdot \mathbf{a}_i = \begin{bmatrix} (K^2)_{11} & (K^2)_{12} & \cdots & (K^2)_{1N} \\ (K^2)_{21} & (K^2)_{22} & \cdots & (K^2)_{2N} \\ \vdots & & & \\ (K^2)_{N1} & (K^2)_{N2} & \cdots & (K^2)_{NN} \end{bmatrix} \begin{bmatrix} a_{i1} \\ a_{i2} \\ \vdots \\ a_{iN} \end{bmatrix}$$

$$= \begin{bmatrix} K^2_{11} a_{i1} + K^2_{12} a_{i2} + \cdots + K^2_{1N} a_{iN} \\ K^2_{21} a_{i1} + K^2_{22} a_{i2} + \cdots + K^2_{2N} a_{iN} \\ \vdots \\ K^2_{N1} a_{i1} + K^2_{N2} a_{i2} + \cdots + K^2_{NN} a_{iN} \end{bmatrix}$$

Let's calculate the first row of $\mathbf{K}^2\mathbf{a}_i$,

$$
\begin{aligned}
&K_{11}^2 a_{i1} + K_{12}^2 a_{i2} + \cdots + K_{1N}^2 a_{iN} \\
&= a_{i1}[k(x_1,x_1)\cdot k(x_1,x_1) + k(x_1,x_2)\cdot k(x_2,x_1) + \cdots + k(x_1,x_N)\cdot k(x_N,x_1)] \\
&\quad + a_{i2}[k(x_1,x_1)\cdot k(x_1,x_2) + k(x_1,x_2)\cdot k(x_2,x_2) + \cdots + k(x_1,x_N)\cdot k(x_N,x_2)] \\
&\quad \vdots \\
&\quad + a_{iN}[k(x_1,x_1)\cdot k(x_1,x_N) + k(x_1,x_2)\cdot k(x_2,x_N) + \cdots + k(x_1,x_N)\cdot k(x_N,x_N)] \quad (2)
\end{aligned}
$$

Comparing Eq (2) with Eq (1), we can see that Eq (2) is just the case when l = 1 in Eq (1).

Therefore we get

$$\sum_{n=1}^{N} k(\mathbf{x}_l,\mathbf{x}_n) \sum_{m=1}^{N} a_{im} k(\mathbf{x}_n,\mathbf{x}_m) = \mathbf{K}^2\mathbf{a}_i$$

Likewise we can show

$$\sum_{n=1}^{N} a_{in} k(\mathbf{x}_l,\mathbf{x}_n) = \mathbf{K}\mathbf{a}_i$$

Therefore Eq (12.78) becomes

$$\mathbf{K}^2\mathbf{a}_i = \lambda_i N \mathbf{K}\mathbf{a}_i$$

Chapter 13. Sequential Data

Eq 13.10: (PRML p.612)

$$p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta}) = p(\mathbf{z}_1|\boldsymbol{\pi}) \left[\prod_{n=2}^{N} p(\mathbf{z}_n|\mathbf{z}_{n-1},\mathbf{A})\right] \cdot \left[\prod_{m=1}^{N} p(\mathbf{x}_m|\mathbf{z}_m,\boldsymbol{\phi})\right]$$

Proof :

$$
\begin{aligned}
p(\mathbf{X},\mathbf{Z}|\boldsymbol{\pi},\mathbf{A},\boldsymbol{\phi}) &= \frac{p(\mathbf{X},\mathbf{Z},\boldsymbol{\pi},\mathbf{A},\boldsymbol{\phi})}{p(\boldsymbol{\pi},\mathbf{A},\boldsymbol{\phi})} \\
&= \frac{p(\mathbf{X}|\mathbf{Z},\boldsymbol{\pi},\mathbf{A},\boldsymbol{\phi})\cdot p(\mathbf{Z},\boldsymbol{\pi},\mathbf{A},\boldsymbol{\phi})}{p(\boldsymbol{\pi},\mathbf{A},\boldsymbol{\phi})} \\
&= p(\mathbf{X}|\mathbf{Z},\boldsymbol{\pi},\mathbf{A},\boldsymbol{\phi})\cdot p(\mathbf{Z}|\boldsymbol{\pi},\mathbf{A},\boldsymbol{\phi}) \quad (1)
\end{aligned}
$$

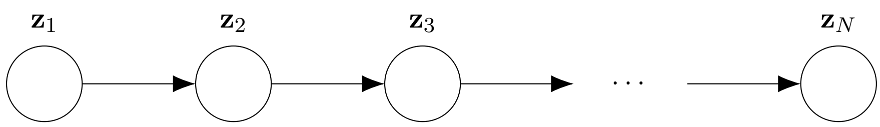
Since the information on $\boldsymbol{\pi}$ and $\mathbf{A}$ is included in $\mathbf{Z}$,

$$p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\pi}, \mathbf{A}, \boldsymbol{\phi}) = p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\phi})$$

Since $\boldsymbol{\phi}$ governs the $\mathbf{x}$ distribution only,

$$p(\mathbf{Z}|\boldsymbol{\pi}, \mathbf{A}, \boldsymbol{\phi}) = p(\mathbf{Z}|\boldsymbol{\pi}, \mathbf{A})$$

$p(\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_N|\boldsymbol{\pi}, \mathbf{A})$: joint distribution of $\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_N$



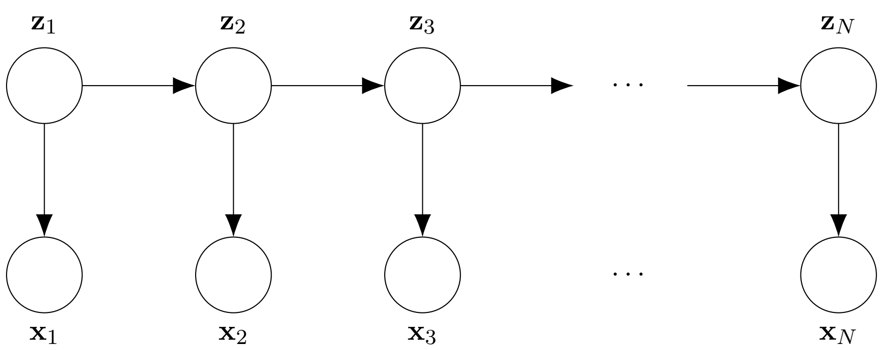
The joint distribution can be expressed with conditional distributions based on the above Markov chain.

As stated in PRML p.610, $\mathbf{z}_n$ distribution depends on $p(\mathbf{z}_n|\mathbf{z}_{n-1})$.

$$\Rightarrow \quad p(\mathbf{Z}|\boldsymbol{\pi}, \mathbf{A}) = p(\mathbf{z}_1|\boldsymbol{\pi}) \cdot \prod_{n=2}^{N} p(\mathbf{z}_n|\mathbf{z}_{n-1}, \mathbf{A}) \tag{2}$$

(See Eq (8.26))

$$p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\phi}) = p(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N|\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_N, \boldsymbol{\phi})$$



Since each observation ($\mathbf{x}_n$) is independent and $\mathbf{x}_n$ only depends on $\mathbf{z}_n$,

$$p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\phi}) = p(\mathbf{x}_1|\mathbf{z}_1, \boldsymbol{\phi}) \cdot p(\mathbf{x}_2|\mathbf{z}_2, \boldsymbol{\phi}) \cdots p(\mathbf{x}_N|\mathbf{z}_N, \boldsymbol{\phi}) \tag{3}$$

Eqs (1), (2), and (3) prove Eq (13.10).

Eq 13.17: (PRML p.617)

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = \sum_{k=1}^{K} \gamma(z_{1k}) \ln \pi_k + \sum_{n=2}^{N} \sum_{j=1}^{K} \sum_{k=1}^{K} \xi(z_{n-1}, z_{nk}) \ln A_{jk}$$
$$+ \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \ln p(\mathbf{x}_n | \boldsymbol{\phi}_k)$$

Proof :

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{old}) \cdot \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \tag{13.12}$$

$$p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) = p(\mathbf{z}_1|\boldsymbol{\pi}) \left[\prod_{n=2}^{N} p(\mathbf{z}_n|\mathbf{z}_{n-1}, \mathbf{A}) \right] \cdot \left[\prod_{m=1}^{N} p(\mathbf{x}_m|\mathbf{z}_m, \boldsymbol{\phi}) \right] \tag{13.10}$$

Plugging Eq (13.10) into (13.12),

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{old}) \cdot \left\{ \ln [(\mathbf{z}_1|\boldsymbol{\pi}) + \sum_{n=2}^{N} \ln p(\mathbf{z}_n|\mathbf{z}_{n-1}, \mathbf{A}) + \sum_{m=1}^{N} \ln p(\mathbf{x}_m|\mathbf{z}_m, \boldsymbol{\phi}) \right\}$$
$$\text{where } p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{old}) = p(\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_N|\mathbf{X}, \boldsymbol{\theta}^{old})$$

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{old}) \ln [(\mathbf{z}_1|\boldsymbol{\pi}) \tag{1}$$

$$+ \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{old}) \sum_{n=2}^{N} \ln p(\mathbf{z}_n|\mathbf{z}_{n-1}, \mathbf{A}) \tag{2}$$

$$+ \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{old}) \sum_{m=1}^{N} \ln p(\mathbf{x}_m|\mathbf{z}_m, \boldsymbol{\phi}) \tag{3}$$

$$(1) = \sum_{\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_N} p(\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_N | \mathbf{X}, \boldsymbol{\theta}^{old}) \ln p(\mathbf{z}_1 | \boldsymbol{\pi})$$

$$\text{(using the product rule } \sum_A p(A, B) = p(B))$$

$$= \sum_{\mathbf{z}_1} p(\mathbf{z}_1 | \mathbf{X}, \boldsymbol{\theta}^{old}) \ln [(\mathbf{z}_1 | \boldsymbol{\pi})$$

$$(2) = \sum_{\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_N} p(\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_N | \mathbf{X}, \boldsymbol{\theta}^{old}) \sum_{n=2}^{N} \ln p(\mathbf{z}_n | \mathbf{z}_{n-1}, \mathbf{A})$$

$$= \sum_{n=2}^{N} \sum_{\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_N} p(\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_N | \mathbf{X}, \boldsymbol{\theta}^{old}) \ln p(\mathbf{z}_n | \mathbf{z}_{n-1}, \mathbf{A})$$

$$= \sum_{\mathbf{z}_1, \mathbf{z}_2} \ln p(\mathbf{z}_2 | \mathbf{z}_1, \mathbf{A}) p(\mathbf{z}_1, \mathbf{z}_2 | \mathbf{X}, \boldsymbol{\theta}^{old}) \qquad \text{(n=2 case)}$$

$$+ \sum_{\mathbf{z}_2, \mathbf{z}_3} \ln p(\mathbf{z}_3 | \mathbf{z}_2, \mathbf{A}) p(\mathbf{z}_2, \mathbf{z}_3 | \mathbf{X}, \boldsymbol{\theta}^{old}) \qquad \text{(n=3 case)}$$

$$\vdots$$

$$+ \sum_{\mathbf{z}_{N-1}, \mathbf{z}_N} \ln p(\mathbf{z}_N | \mathbf{z}_{N-1}, \mathbf{A}) p(\mathbf{z}_{N-1}, \mathbf{z}_N | \mathbf{X}, \boldsymbol{\theta}^{old}) \qquad \text{(n=N case)}$$

Using Eq (13.14): $\xi(\mathbf{z}_{n-1}, \mathbf{z}_n) = p(\mathbf{z}_{n-1}, \mathbf{z}_n | \mathbf{X}, \boldsymbol{\theta}^{old})$,

$$(2) = \sum_{\mathbf{z}_1, \mathbf{z}_2} \ln p(\mathbf{z}_2 | \mathbf{z}_1, \mathbf{A}) \cdot \xi(\mathbf{z}_1, \mathbf{z}_2) + \sum_{\mathbf{z}_2, \mathbf{z}_3} \ln p(\mathbf{z}_3 | \mathbf{z}_2, \mathbf{A}) \cdot \xi(\mathbf{z}_2, \mathbf{z}_3) + \cdots$$
$$+ \sum_{\mathbf{z}_{N-1}, \mathbf{z}_N} \ln p(\mathbf{z}_N | \mathbf{z}_{N-1}, \mathbf{A}) \cdot \xi(\mathbf{z}_{N-1}, \mathbf{z}_N)$$

Using Eq (13.7): $\ln p(\mathbf{z}_2 | \mathbf{z}_1, \mathbf{A}) = \sum_{k=1}^{K} \sum_{j=1}^{K} (\ln A_{jk}) \mathbf{z}_{1j} \mathbf{z}_{2k}$

$$(2) = \left[\sum_{\mathbf{z}_1, \mathbf{z}_2} \sum_{j,k} (\ln A_{jk}) z_{1j} z_{2k} \xi(\mathbf{z}_1, \mathbf{z}_2) \right] + \left[\sum_{\mathbf{z}_2, \mathbf{z}_3} \sum_{j,k} (\ln A_{jk}) z_{2j} z_{3k} \xi(\mathbf{z}_2, \mathbf{z}_3) \right] + \cdots$$
$$+ \left[\sum_{\mathbf{z}_{N-1}, \mathbf{z}_N} \sum_{j,k} (\ln A_{jk}) z_{N-1,j} z_{Nk} \xi(\mathbf{z}_{N-1}, \mathbf{z}_N) \right]$$
$$= \sum_{j,k} (\ln A_{jk}) \left[\sum_{\mathbf{z}_1 . \mathbf{z}_2} \xi(\mathbf{z}_1, \mathbf{z}_2) z_{1j} z_{2k} + \sum_{\mathbf{z}_2, \mathbf{z}_3} \xi(\mathbf{z}_2, \mathbf{z}_3) z_{2j} z_{3k} + \cdots \right.$$
$$\left. + \sum_{\mathbf{z}_{N-1}, \mathbf{z}_N} \xi(\mathbf{z}_{N-1}, \mathbf{z}_N) z_{N-1,j} z_{Nk} \right]$$

Using Eq (13.16): $\xi(z_{n-1,j}, z_{nk}) = \sum_{\mathbf{z}_{n-1},\mathbf{z}_n} \xi(\mathbf{z}_{n-1}, \mathbf{z}_n) z_{n-1,j} z_{nk}$

$$\begin{aligned}(2) &= \sum_{j,k} \ln A_{jk}[\xi(\mathbf{z}_1, \mathbf{z}_2) + \xi(\mathbf{z}_2, \mathbf{z}_3) + \cdots + \xi(\mathbf{z}_{N-1}, \mathbf{z}_N)] \\ &= \sum_{n=2}^{N} \sum_{jk} \ln A_{jk} \xi(\mathbf{z}_{N-1}, \mathbf{z}_N)\end{aligned}$$

$$\begin{aligned}(3) &= \sum_{\mathbf{z}_1,\mathbf{z}_2,\cdots,\mathbf{z}_N} p(\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_N|\mathbf{X}, \boldsymbol{\theta}^{old}) \sum_{m=1}^{N} \ln p(\mathbf{x}_m|\mathbf{z}_m, \boldsymbol{\phi}) \\ &= \sum_{m=1}^{N} \sum_{\mathbf{z}_1,\mathbf{z}_2,\cdots,\mathbf{z}_N} p(\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_N|\mathbf{X}, \boldsymbol{\theta}^{old}) \ln p(\mathbf{x}_m|\mathbf{z}_m, \boldsymbol{\phi}) \\ &= \sum_{\mathbf{z}_1} p(\mathbf{z}_1|\mathbf{X}, \boldsymbol{\theta}^{old}) \ln p(\mathbf{x}_1|\mathbf{z}_1, \boldsymbol{\phi}) && \text{(m=1 case)} \\ &\quad + \sum_{\mathbf{z}_2} p(\mathbf{z}_2|\mathbf{X}, \boldsymbol{\theta}^{old}) \ln p(\mathbf{x}_2|\mathbf{z}_2, \boldsymbol{\phi}) && \text{(m=2 case)} \\ &\quad \vdots \\ &\quad + \sum_{\mathbf{z}_N} p(\mathbf{z}_N|\mathbf{X}, \boldsymbol{\theta}^{old}) \ln p(\mathbf{x}_N|\mathbf{z}_N, \boldsymbol{\phi}) && \text{(m=N case)}\end{aligned}$$

Using Eq (13.9): $p(\mathbf{x}_n|\mathbf{z}_n, \boldsymbol{\phi}) = \prod_{k=1}^{K} p(\mathbf{x}_n|\boldsymbol{\phi}_k)^{\mathbf{z}_{nk}}$

$$\longrightarrow \quad \ln p(\mathbf{x}_n|\mathbf{z}_n, \boldsymbol{\phi}) = \sum_{k=1}^{K} \mathbf{z}_{nk} \ln p(\mathbf{x}_n|\boldsymbol{\phi}_k)$$

$$\begin{aligned}(3) &= \sum_{\mathbf{z}_1} p(\mathbf{z}_1|\mathbf{X}, \boldsymbol{\theta}^{old}) \sum_{k=1}^{K} \mathbf{z}_{nk} \ln p(\mathbf{x}_1|\phi_k) + \sum_{\mathbf{z}_2} p(\mathbf{z}_2|\mathbf{X}, \boldsymbol{\theta}^{old}) \sum_{k=1}^{K} \mathbf{z}_{2k} \ln p(\mathbf{x}_2|\boldsymbol{\phi}_k) + \cdots \\ &\quad + \sum_{\mathbf{z}_N} p(\mathbf{z}_N|\mathbf{X}, \boldsymbol{\theta}^{old}) \sum_{k=1}^{K} \mathbf{z}_{Nk} \ln p(\mathbf{x}_N|\boldsymbol{\phi}_k)\end{aligned}$$

Using $\sum_{\mathbf{z}_1} \mathbf{z}_{1k} p(\mathbf{z}_1|\mathbf{X}, \boldsymbol{\theta}^{old}) = \sum_{\mathbf{z}_1} \gamma(\mathbf{z}_1)\mathbf{z}_{1k} = \gamma(\mathbf{z}_{1k})$

where Eqs (13.13) and (13.16) are used.

$$(3) = \sum_k \left[\gamma(\mathbf{z}_{1k}) \ln p(\mathbf{x}_1|\boldsymbol{\phi}_k) + \gamma(\mathbf{z}_{2k}) \ln p(\mathbf{x}_2|\boldsymbol{\phi}_k) + \cdots + \gamma(\mathbf{z}_{Nk}) \ln p(\mathbf{x}_N|\boldsymbol{\phi}_k)\right]$$
$$= \sum_{m=1}^{N} \sum_{k=1}^{K} \gamma(\mathbf{z}_{mk}) \ln p(\mathbf{x}_m|\boldsymbol{\phi}_k)$$

Eqs 13.20 & 13.21: (PRML p.618)

$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})\mathbf{x}_n}{\sum_{n=1}^{N} \gamma(z_{nk})}$$
$$\boldsymbol{\Sigma}_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T}{\sum_{n=1}^{N} \gamma(z_{nk})}$$

Proof :

The third term in Eq (13.17):

$$\sum_n \sum_k \gamma(z_{nk}) \ln p(\mathbf{x}_n|\boldsymbol{\phi}_k) = \sum_n \sum_k \gamma(z_{nk}) \ln \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$
$$\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi)^{N/2}\|\boldsymbol{\Sigma}_k\|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)\right]$$
$$\ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = -\frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln\|\boldsymbol{\Sigma}_k\| - \frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)$$

To find $\boldsymbol{\mu}_k$ max,

$$\frac{\partial}{\partial \boldsymbol{\mu}_k} \sum_n \sum_k \gamma(z_{nk}) \ln \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \sum_n \gamma(z_{nk}) \cdot (\mathbf{x}_n - \boldsymbol{\mu}_k)\boldsymbol{\Sigma}_k^{-1} = 0$$
$$\sum_n \gamma(z_{nk})\mathbf{x}_n = \sum_n \gamma(z_{nk})\boldsymbol{\mu}_k$$
$$\therefore \quad \boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})\mathbf{x}_n}{\sum_{n=1}^{N} \gamma(z_{nk})}$$

To find $\boldsymbol{\Sigma}_k$ max, (using Matrix Cookbook Eq (70))

$$\frac{\partial}{\partial \boldsymbol{\Sigma}_k} \sum_n \sum_k \gamma(z_{nk}) \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$
$$= \sum_n \gamma(z_{nk}) \left[-\frac{1}{2}\left(\frac{-1}{\boldsymbol{\Sigma}_k^2}\right)(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T - \frac{1}{2}\frac{1}{\|\Sigma_k\|}\left(\frac{\partial\|\boldsymbol{\Sigma}_k\|}{\partial \boldsymbol{\Sigma}_k}\right)\right]$$

Using Matrix Cookbook Eq (49), $\dfrac{\partial\|\mathbf{X}\|}{\partial \mathbf{X}} = \|\mathbf{X}\|(\mathbf{X}^{-1})^T$

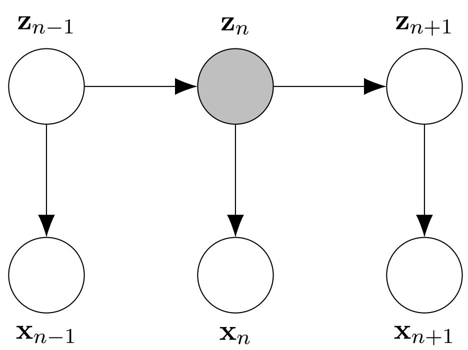
$$
\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\Sigma}_k} \sum_n \sum_k \gamma(z_{nk}) \ln \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\
&\quad = \sum_n \gamma(z_{nk}) \left[\frac{1}{2\boldsymbol{\Sigma}_k^2} (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T - \frac{1}{2} \frac{1}{\|\boldsymbol{\Sigma}_k\|} \|\boldsymbol{\Sigma}_k\| (\boldsymbol{\Sigma}_k^{-1})^T \right] = 0 \\
&\qquad \Rightarrow \quad \sum_n \gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T = \boldsymbol{\Sigma}_k \sum_n \gamma(z_{nk}) \\
&\qquad\quad \therefore \quad \boldsymbol{\Sigma}_k = \frac{\sum_{n=1}^N \gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T}{\sum_{n=1}^N \gamma(z_{nk})}
\end{aligned}
$$

Eq 13.43: (PRML p.623)

$$
p(\mathbf{x}_1, \cdots, \mathbf{x}_N | \mathbf{z}_{n-1}, \mathbf{z}_n) = p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1} | \mathbf{z}_{n-1}) \, p(\mathbf{x}_n | \mathbf{z}_n) \, p(\mathbf{x}_{n+1}, \cdots, \mathbf{x}_N | \mathbf{z}_n)
$$

Proof :

$$
\begin{aligned}
& p(\mathbf{x}_1, \cdots, \mathbf{x}_N | \mathbf{z}_{n-1}, \mathbf{z}_n) \\
& = \frac{1}{p(\mathbf{z}_{n-1}, \mathbf{z}_n)} \, p(\mathbf{x}_1, \cdots, \mathbf{x}_N, \mathbf{z}_{n-1}, \mathbf{z}_n) \\
& = \frac{1}{p(\mathbf{z}_{n-1}, \mathbf{z}_n)} \, p(\mathbf{x}_1, \cdots, \mathbf{x}_{n-1} | \mathbf{x}_n, \cdots, \mathbf{x}_N, \mathbf{z}_{n-1}, \mathbf{z}_n) \, p(\mathbf{x}_n, \cdots, \mathbf{x}_N, \mathbf{z}_{n-1}, \mathbf{z}_n) \\
& (\text{when } \mathbf{z}_{n-1} \text{ is observed, } \{\mathbf{x}_1, \cdots, \mathbf{x}_{n-1}\} \perp \{\mathbf{x}_n, \cdots, \mathbf{x}_N, \mathbf{z}_n\}) \\
& = \frac{1}{p(\mathbf{z}_{n-1}, \mathbf{z}_n)} \, p(\mathbf{x}_1, \cdots, \mathbf{x}_{n-1} | \mathbf{z}_{n-1}) \, p(\mathbf{x}_n, \cdots, \mathbf{x}_N, \mathbf{z}_{n-1}, \mathbf{z}_n) \\
& = \frac{1}{p(\mathbf{z}_{n-1}, \mathbf{z}_n)} \, p(\mathbf{x}_1, \cdots, \mathbf{x}_{n-1} | \mathbf{z}_{n-1}) \, p(\mathbf{x}_n | \mathbf{x}_{n+1}, \cdots, \mathbf{x}_N, \mathbf{z}_{n-1}, \mathbf{z}_n) \\
& \quad \cdot p(\mathbf{x}_{n+1}, \cdots, \mathbf{x}_N, \mathbf{z}_{n-1}, \mathbf{z}_n)
\end{aligned}
$$



When $\mathbf{z}_n$ is observed, $\mathbf{x}_n \perp \{\mathbf{x}_{n+1}, \cdots, \mathbf{x}_N, \mathbf{z}_{n-1}\}$.

$$\begin{aligned}
&p(\mathbf{x}_1, \cdots, \mathbf{x}_N | \mathbf{z}_{n-1}, \mathbf{z}_n) \\
&= \frac{1}{p(\mathbf{z}_{n-1}, \mathbf{z}_n)}\, p(\mathbf{x}_1, \cdots, \mathbf{x}_{n-1} | \mathbf{z}_{n-1})\, p(\mathbf{x}_n | \mathbf{z}_n)\, p(\mathbf{x}_{n+1}, \cdots, \mathbf{x}_N, \mathbf{z}_{n-1}, \mathbf{z}_n) \\
&= p(\mathbf{x}_1, \cdots, \mathbf{x}_{n-1} | \mathbf{z}_{n-1})\, p(\mathbf{x}_n | \mathbf{z}_n)\, p(\mathbf{x}_{n+1}, \cdots, \mathbf{x}_N | \mathbf{z}_{n-1}, \mathbf{z}_n)
\end{aligned}$$

Eqs 13.45 & 13.46: (PRML p.625)

$$h(\mathbf{z}_1) = p(\mathbf{z}_1) p(\mathbf{x}_1 | \mathbf{z}_1)$$

$$f_n(\mathbf{z}_{n-1}, \mathbf{z}_n) = p(\mathbf{z}_n | \mathbf{z}_{n-1})\, p(\mathbf{x}_n | \mathbf{z}_n)$$

Proof :

From Fig 13.5 (Markov chain), the joint distribution of the graph is

$$p(\mathbf{X}, \mathbf{Z}) = p(\mathbf{z}_1)\, p(\mathbf{x}_1 | \mathbf{z}_1)\, p(\mathbf{z}_2 | \mathbf{z}_1)\, p(\mathbf{x}_2 | \mathbf{z}_2), \cdots, p(\mathbf{z}_N | \mathbf{z}_{N-1})\, p(\mathbf{x}_N | \mathbf{z}_N) \tag{1}$$

To transform tis to a factor graph as in Fig 13.14,

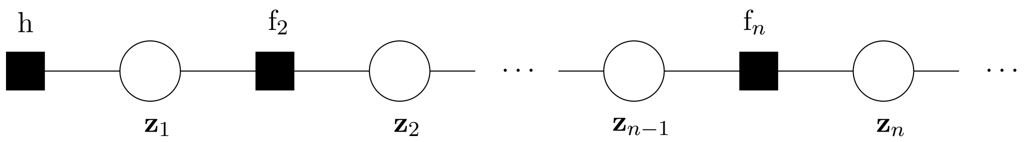
$$\begin{aligned}
&\chi = p(\mathbf{z}_1) \\
&g_1(\mathbf{z}_1, \mathbf{x}_1) = p(\mathbf{x}_1 | \mathbf{z}_1) \\
&\psi_1(\mathbf{z}_1, \mathbf{z}_2) = p(\mathbf{z}_2 | \mathbf{z}_1) \\
&g_2(\mathbf{z}_2, \mathbf{x}_2) = p(\mathbf{x}_2 | \mathbf{z}_2) \\
&\psi_2(\mathbf{z}_2, \mathbf{z}_3) = p(\mathbf{z}_3 | \mathbf{z}_2) \\
&\vdots \\
&g_N(\mathbf{z}_N, \mathbf{x}_N) = p(\mathbf{x}_N | \mathbf{z}_N) \\
&\psi_{N-1}(\mathbf{z}_{N-1}, \mathbf{z}_N) = p(\mathbf{z}_N | \mathbf{z}_{N-1})
\end{aligned}$$

$$\Rightarrow \quad p(\mathbf{X}, \mathbf{Z}) = \chi\, g_1(\mathbf{z}_1, \mathbf{x}_1) \prod_{n=1}^{N-1} g_{n+1}(\mathbf{z}_{n+1}, \mathbf{x}_{n+1})\, \Psi_n(\mathbf{z}_n, \mathbf{z}_{n-1})$$

We can simplify this factor graph as follows,

$$
\begin{aligned}
h(\mathbf{z}_1) &= p(\mathbf{z}_1)p(\mathbf{x}_1|\mathbf{z}_1) \\
f_2 &= p(\mathbf{z}_2|\mathbf{z}_1)\, p(\mathbf{x}_2|\mathbf{z}_2) \\
&\vdots \\
f_n &= p(\mathbf{z}_n|\mathbf{z}_{n-1})\, p(\mathbf{x}_n|\mathbf{z}_n) \\
&\vdots \\
f_N &= p(\mathbf{z}_N|\mathbf{z}_{N-1})\, p(\mathbf{x}_N|\mathbf{z}_N) \\
\Rightarrow \quad p(\mathbf{X}, \mathbf{Z}) &= h \cdot \prod_{n=2}^{N} f_n(\mathbf{z}_{n-1}, \mathbf{z}_n)
\end{aligned}
$$

The corresponding factor graph diagram will be



Eq 13.61: (PRML p.628)

$$
\widehat{\beta}(\mathbf{z}_n) = \frac{p(\mathbf{x}_{n+1}, \cdots, \mathbf{x}_N|\mathbf{z}_n)}{p(\mathbf{x}_{n+1}, \cdots, \mathbf{x}_N|\mathbf{x}_1, \cdots, \mathbf{x}_n)}
$$

Proof :

$$
\beta(\mathbf{z}_n) = \left(\prod_{m=n+1}^{N} c_m \right) \widehat{\beta}(\mathbf{z}_n) \tag{13.60}
$$

$$
\beta(\mathbf{z}_n) = p(\mathbf{x}_{n+1}, \cdots, \mathbf{x}_N|\mathbf{z}_n) \tag{13.35}
$$

$$
p(\mathbf{x}_1, \cdots, \mathbf{x}_n) = \prod_{m=1}^{n} c_m \tag{13.57}
$$

$$
\begin{aligned}
\prod_{m=n+1}^{N} c_m = \frac{\prod_{m=1}^{N} c_m}{\prod_{m=1}^{n} c_m} &= \frac{p(\mathbf{x}_1, \cdots, \mathbf{x}_N)}{p(x_1, \cdots, \mathbf{x}_n)} \\
&= p(\mathbf{x}_1, \cdots, \mathbf{x}_N|\mathbf{x}_1, \cdots, \mathbf{x}_n) \\
&\text{(since } \mathbf{x}_1, \cdots, \mathbf{x}_n \text{ are conditioning)} \\
&= p(\mathbf{x}_{n+1}, \cdots, \mathbf{x}_N|\mathbf{x}_1, \cdots, \mathbf{x}_n)
\end{aligned}
$$

$$\therefore \quad \widehat{\beta}(\mathbf{z}_n) = \frac{\beta(\mathbf{z}_n)}{\prod_{m=n+1}^{N} c_m} = \frac{p(\mathbf{x}_{n+1}, \cdots, \mathbf{x}_N|\mathbf{z}_n)}{p(\mathbf{x}_{n+1}, \cdots, \mathbf{x}_N|\mathbf{x}_1, \cdots, \mathbf{x}_n)}$$

Eq 13.87: (PRML p.638)

$$\int \mathcal{N}(\mathbf{z}_n|\mathbf{A}\mathbf{z}_{n-1}, \Gamma) \cdot \mathcal{N}(\mathbf{z}_{n-1}|\boldsymbol{\mu}_{n-1}, \mathbf{V}_{n-1}) d\mathbf{z}_{n-1} = \mathcal{N}(\mathbf{x}_n|\mathbf{A}\boldsymbol{\mu}_{n-1}, \mathbf{P}_{n-1})$$

$$\text{where } \mathbf{P}_{n-1} = \mathbf{A}\mathbf{V}_{n-1}\mathbf{A}^T + \mathbf{V}$$

Proof :

$$\int \mathcal{N}(\mathbf{z}_n|\mathbf{A}\mathbf{z}_{n-1}, \boldsymbol{\Gamma}) \cdot \mathcal{N}(\mathbf{z}_{n-1}|\boldsymbol{\mu}_{n-1}, \mathbf{V}_{n-1}) d\mathbf{z}_{n-1} \tag{1}$$

If we make a comparison the above integration with the following equation,

$$\begin{aligned} p(\mathbf{y}) &= \int p(\mathbf{x}, \mathbf{y}) dx \\ &= \int p(\mathbf{y}|\mathbf{x}) \cdot p(\mathbf{x}) d\mathbf{x} \end{aligned}$$

We can see that the first integrand in Eq (1) corresponds to $p(\mathbf{y}|\mathbf{x})$ and the second to $p(\mathbf{x})$. If we utilize the Marginal / Conditional Gaussians, we can calculate $p(z_n)$ as follows,

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \tag{2.113}$$

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + b, \mathbf{L}^{-1}) \tag{2.114}$$

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + b, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T) \tag{2.115}$$

Marginal/Conditional Gaussians	Current
x	z_{n-1}
y	z_n
p(x)	$\mathcal{N}(z_{n-1}\|\mu_{n-1}, V_{n-1})$
$p(y\|x)$	$\mathcal{N}(z_n\|Az_{n-1}, \Gamma)$
μ	μ_{n-1}
Λ^{-1}	V_{n-1}
A	A
b	0
L^{-1}	Γ

By making substitutions, we have

$$\Rightarrow \quad p(\mathbf{z}_n) = \mathcal{N}(\mathbf{z}_n|\mathbf{A}\boldsymbol{\mu}_{n-1} + 0, \boldsymbol{\Gamma} + \mathbf{A}\mathbf{V}_{n-1}\mathbf{A}^T)$$
$$= \mathcal{N}(\mathbf{z}_n|\mathbf{A}\boldsymbol{\mu}_{n-1}, \boldsymbol{\Gamma} + \mathbf{A}\mathbf{V}_{n-1}\mathbf{A}^T)$$

Eqs 13.89 ∼ 13.91 : (PRML p.639)

$$\boldsymbol{\mu}_n = \mathbf{A}\boldsymbol{\mu}_{n-1} + \mathbf{K}_n(\mathbf{x}_n - \mathbf{C}\mathbf{A}\boldsymbol{\mu}_{n-1})$$
$$\mathbf{V}_n = (\mathbf{I} - \mathbf{K}_n\mathbf{C})\mathbf{P}_{n-1}$$
$$c_n = \mathcal{N}(\mathbf{x}_n|\mathbf{C}\mathbf{A}\boldsymbol{\mu}_{n-1}, \mathbf{C}\mathbf{P}_{n-1}\mathbf{C}^T + \boldsymbol{\Sigma})$$

Proof :

From Eqs (13.86) and (13.87),

$$c_n\mathcal{N}(\mathbf{z}_n|\boldsymbol{\mu}_n, \mathbf{V}_n) = \mathcal{N}(\mathbf{x}_n|\mathbf{C}\mathbf{z}_n, \boldsymbol{\Sigma}) \cdot \mathcal{N}(\mathbf{z}_n|\mathbf{A}\boldsymbol{\mu}_{n-1}, \mathbf{P}_{n-1})$$

where $\mathcal{N}(\mathbf{x}_n|\mathbf{C}\mathbf{z}_n, \boldsymbol{\Sigma})$ corresponds to Eq (2.114): $p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{X} + b, \mathbf{L}^{-1})$, and $\mathcal{N}(\mathbf{z}_n|\mathbf{A}\boldsymbol{\mu}_{n-1}, \mathbf{P}_{n-1})$ to Eq (2.113): $p(\mathbf{x}) = (\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$.

To calculate p(y) equivalent,

Marginal/Conditional Gaussians	Current
$\mathbf{x}$	$\mathbf{z}_n$
$\mathbf{y}$	$\mathbf{x}_n$
$\mathbf{A}$	$\mathbf{C}$
b	0
$\mathbf{L}^{-1}$	$\boldsymbol{\Sigma}$
$\boldsymbol{\mu}$	$\mathbf{A}\boldsymbol{\mu}_{n-1}$
$\boldsymbol{\Lambda}^{-1}$	$\mathbf{P}_{n-1}$

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + b, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T) \tag{2.115}$$

by making substitutions

$$\Rightarrow \quad \mathcal{N}(\mathbf{x}_n|\mathbf{C}\mathbf{A}\boldsymbol{\mu}_{n-1} + 0, \boldsymbol{\Sigma} + \mathbf{C}\mathbf{P}_{n-1}\mathbf{C}^T)$$

We can identify this as c_n in Eq (13.86).

$$\therefore \ \ c_n = \mathcal{N}(\mathbf{x}_n|\mathbf{CA}\boldsymbol{\mu}_{n-1}, \boldsymbol{\Sigma} + \mathbf{CP}_{n-1}\mathbf{C}^T)$$

Next, let's find $p(\mathbf{x}|\mathbf{y})$ equivalent,

$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}\{A^T L(\mathbf{y}\text{-}b) + \boldsymbol{\Lambda\mu}\}, \boldsymbol{\Sigma}) \tag{2.116}$$
$$\text{where } \boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^T\mathbf{LA})^{-1}$$

$$\Rightarrow \ \mathcal{N}\Big(\mathbf{z}_n|(\mathbf{P}_{n-1}^{-1} + \mathbf{C}^T\boldsymbol{\Sigma}^{-1}\mathbf{C})^{-1}\{\mathbf{C}^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - 0) + \mathbf{P}_{n-1}^{-1}\mathbf{A}\boldsymbol{\mu}_{n-1}\}, (\mathbf{P}_{n-1}^{-1} + \mathbf{C}^T\boldsymbol{\Sigma}^{-1}\mathbf{C})^{-1}\Big)$$
$$= \mathcal{N}\Big(\mathbf{z}_n|(\mathbf{P}_{n-1}^{-1} + \mathbf{C}^T\boldsymbol{\Sigma}^{-1}\mathbf{C})^{-1}\mathbf{C}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_n + (\mathbf{P}_{n-1}^{-1} + \mathbf{C}^T\boldsymbol{\Sigma}^{-1}\mathbf{C})^{-1}\mathbf{P}_{n-1}^{-1}\mathbf{A}\boldsymbol{\mu}_{n-1},$$
$$(\mathbf{P}_{n-1}^{-1} + \mathbf{C}^T\boldsymbol{\Sigma}^{-1}\mathbf{C})^{-1}\Big)$$

Let's calculate term by term.

1st term:

$$\begin{aligned}(\mathbf{P}_{n-1}^{-1} + \mathbf{C}^T\boldsymbol{\Sigma}^{-1}\mathbf{C})^{-1}\mathbf{C}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_n &= \mathbf{P}_{n-1}(1 + \mathbf{C}^T\boldsymbol{\Sigma}^{-1}\mathbf{CP}_{n-1})^{-1}\mathbf{C}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_n \\ &= \mathbf{P}_{n-1}\mathbf{C}^T(\boldsymbol{\Sigma} + \mathbf{C}^T\mathbf{CP}_{n-1})^{-1}\mathbf{x}_n \\ &\text{(Here we identify } \mathbf{P}_{n-1}\mathbf{C}^T(\boldsymbol{\Sigma} + \mathbf{C}^T\mathbf{CP}_{n-1})^{-1} = \mathbf{K}_n) \\ &= \mathbf{K}_n\mathbf{x}_n \qquad (1)\end{aligned}$$

This also can be derived from using Eq (C.5),

$$(\mathbf{P}^{-1} + \mathbf{B}^T\mathbf{R}^{-1}\mathbf{B})^{-1}\mathbf{B}^T\mathbf{R}^{-1} = \mathbf{PB}^T(\mathbf{BPB}^T + \mathbf{R})^{-1} \tag{C.5}$$

$$\mathbf{P} \longleftrightarrow \mathbf{P}_{n-1}$$
$$\mathbf{B} \longleftrightarrow \mathbf{C}$$
$$\mathbf{R} \longleftrightarrow \boldsymbol{\Sigma}$$

$$\Rightarrow \quad (\mathbf{P}_{n-1}^{-1} + \mathbf{C}^T\boldsymbol{\Sigma}^{-1}\mathbf{C})^{-1}\boldsymbol{\Sigma}^{-1} = \mathbf{P}_{n-1}\mathbf{C}^T(\mathbf{CP}_{n-1}\mathbf{C}^T + \boldsymbol{\Sigma})^{-1}$$

2nd term:

$$(\mathbf{P}_{n-1}^{-1} + \mathbf{C}^T\boldsymbol{\Sigma}^{-1}\mathbf{C})^{-1}\mathbf{P}_{n-1}^{-1}\mathbf{A}\boldsymbol{\mu}_{n-1}$$

$$(\mathbf{A} + \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{D} + \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1} \tag{C.7}$$

$$\begin{aligned}
\mathbf{A} &\longleftrightarrow \mathbf{P}_{n-1}^{-1} \\
\mathbf{B} &\longleftrightarrow \mathbf{C}^T \\
\mathbf{D} &\longleftrightarrow \mathbf{\Sigma} \\
\mathbf{C} &\longleftrightarrow \mathbf{C}
\end{aligned}$$

By making substitutions using the above correspondence, the second term becomes,

$$\begin{aligned}
\text{2nd term } &= \left[\mathbf{P}_{n-1} - \mathbf{P}_{n-1}\mathbf{C}^T(\mathbf{\Sigma} + \mathbf{C}\mathbf{P}_{n-1}\mathbf{C}^T)^{-1}\mathbf{C}\mathbf{P}_{n-1}\right]\mathbf{P}_{n-1}^{-1}\mathbf{A}\boldsymbol{\mu}_{n-1} \\
&= [1 - \mathbf{P}_{n-1}\mathbf{C}^T(\mathbf{\Sigma} + \mathbf{C}\mathbf{P}_{n-1}\mathbf{C}^T)^{-1}\mathbf{C}]\mathbf{A}\boldsymbol{\mu}_{n-1} \\
&= \mathbf{A}\boldsymbol{\mu}_{n-1} - \mathbf{P}_{n-1}\mathbf{C}^T(\mathbf{\Sigma} + \mathbf{C}\mathbf{P}_{n-1}\mathbf{C}^T)^{-1}\mathbf{C}\mathbf{A}\boldsymbol{\mu}_{n-1} \\
&= \mathbf{A}\boldsymbol{\mu}_{n-1} - \mathbf{K}_n\mathbf{C}\mathbf{A}\boldsymbol{\mu}_{n-1} \qquad (2)
\end{aligned}$$

From Eqs (1) and (2), we have found $\boldsymbol{\mu}_n$ to be Eq (13.86).

$$\therefore \quad \boldsymbol{\mu}_n = \mathbf{A}\boldsymbol{\mu}_{n-1} + \mathbf{K}_n(\mathbf{x}_n - \mathbf{C}\mathbf{A}\boldsymbol{\mu}_{n-1})$$

3rd term:

Using Eq (C.7),

$$\begin{aligned}
(\mathbf{P}_{n-1}^{-1} + \mathbf{C}^T\mathbf{\Sigma}^{-1}\mathbf{C})^{-1} &= \mathbf{P}_{n-1} - \mathbf{P}_{n-1}\mathbf{C}^T(\mathbf{\Sigma} + \mathbf{C}\mathbf{P}_{n-1}\mathbf{C}^T)^{-1}\mathbf{C}\mathbf{P}_{n-1} \\
&= \mathbf{P}_{n-1} - \mathbf{K}_n\mathbf{C}\mathbf{P}_{n-1} \\
\therefore \quad \mathbf{V}_n &= (\mathbf{I} - \mathbf{K}_n\mathbf{C})\mathbf{P}_{n-1}
\end{aligned}$$

Eq 13.117: (PRML p.645)

$$
\begin{aligned}
\mathbb{E}[f(\mathbf{z}_n)] &= \int f(\mathbf{z}_n)\, p(\mathbf{z}_n|\mathbf{X}_n) d\mathbf{z}_n \\
&= \int f(\mathbf{z}_n)\, p(\mathbf{z}_n|\mathbf{x}_n, \mathbf{X}_{n-1}) d\mathbf{z}_n \\
&= \frac{\int f(\mathbf{z}_n)\, p(\mathbf{x}_n|\mathbf{z}_n)\, p(\mathbf{z}_n|\mathbf{X}_{n-1}) d\mathbf{z}_n}{\int p(\mathbf{x}_n|\mathbf{z}_n)\, p(\mathbf{z}_n|\mathbf{X}_{n-1}) d\mathbf{z}_n} \\
&\simeq \sum_{l=1}^{L} \mathbf{w}_n^{(l)} f(\mathbf{z}_n^{(l)}) \\
\text{where } \mathbf{w}_n^{(l)} &= \frac{p(\mathbf{x}_n|\mathbf{z}_n^{(l)})}{\sum_{m=1}^{L} p(\mathbf{x}_n|\mathbf{z}_n^{(m)})}
\end{aligned}
$$

Proof :

The class of distribution considered here is from Fig (13.5).

If z_n is conditioned (same as observed), then the following conditional independence is preserved.

$$
\begin{aligned}
& p(\mathbf{x}_n, \mathbf{X}_{n-1}|\mathbf{z}_n) = p(\mathbf{x}_n|\mathbf{z}_n) \cdot p(\mathbf{X}_{n-1}|\mathbf{z}_n) \\
& p(\mathbf{x}_n|\mathbf{z}_n, \mathbf{X}_{n-1}) = p(\mathbf{x}_n|\mathbf{z}_n) \qquad (1)
\end{aligned}
$$

Let's prove this.

The conditional independence says that

$$p(\mathbf{A}, \mathbf{B}|\mathbf{C}) = p(\mathbf{A}|\mathbf{C}) \cdot p(\mathbf{B}|\mathbf{C})$$

$$
\begin{aligned}
p(\mathbf{x}_n|\mathbf{z}_n, \mathbf{X}_{n-1}) &= \frac{p(\mathbf{x}_n, \mathbf{z}_n, \mathbf{X}_{n-1})}{p(\mathbf{z}_n, \mathbf{X}_{n-1})} \\
&= \frac{p(\mathbf{x}_n, \mathbf{X}_{n-1}|\mathbf{z}_n) \cdot p(\mathbf{z}_n)}{p(\mathbf{z}_n, \mathbf{X}_{n-1})} \\
&= \frac{p(\mathbf{x}_n|\mathbf{z}_n) \cdot p(\mathbf{X}_{n-1}|\mathbf{z}_n) \cdot p(\mathbf{z}_n)}{p(\mathbf{z}_n, \mathbf{X}_{n-1})} \\
&= \frac{p(\mathbf{x}_n|\mathbf{z}_n) \cdot p(\mathbf{X}_{n-1}|\mathbf{z}_n)}{p(\mathbf{X}_{n-1}|\mathbf{z}_n)} \\
&= p(\mathbf{x}_n|\mathbf{z}_n)
\end{aligned}
$$

Now going back to prove Eq (13.117),

$$\mathbb{E}[f] = \int f(\mathbf{z})\, p(\mathbf{z}) d\mathbf{z} \qquad (11.1)$$

Utilizing Eq (11.1),

$$\begin{aligned}\mathbb{E}[f(\mathbf{z}_n)] &= \int f(\mathbf{z}_n)\, p(\mathbf{z}_n|\mathbf{X}_n) d\mathbf{z}_n \\ &= \int f(\mathbf{z}_n)\, p(\mathbf{z}_n|\mathbf{x}_n, \mathbf{X}_{n-1}) d\mathbf{z}_n\end{aligned}$$

$$\begin{aligned}p(\mathbf{z}_n|\mathbf{x}_n, \mathbf{X}_{n-1}) &= \frac{p(\mathbf{z}_n, \mathbf{x}_n, \mathbf{X}_{n-1})}{p(\mathbf{x}_n, \mathbf{X}_{n-1})} \\ &= \frac{p(\mathbf{x}_n|\mathbf{z}_n, \mathbf{X}_{n-1}) \cdot p(\mathbf{z}_n, \mathbf{X}_{n-1})}{p(\mathbf{x}_n, \mathbf{X}_{n-1})} \\ &\text{(using Eq (1))} \\ &= \frac{p(\mathbf{x}_n|\mathbf{z}_n) \cdot p(\mathbf{z}_n, \mathbf{X}_{n-1})}{p(\mathbf{x}_n, \mathbf{X}_{n-1})} \\ &= \frac{p(\mathbf{x}_n|\mathbf{z}_n) \cdot p(\mathbf{z}_n|\mathbf{X}_{n-1}) \cdot p(\mathbf{X}_{n-1})}{p(\mathbf{x}_n, \mathbf{X}_{n-1})}\end{aligned}$$

Since

$$\begin{aligned}p(\mathbf{x}_n, \mathbf{X}_{n-1}) &= \int p(\mathbf{x}_n, \mathbf{X}_{n-1}, \mathbf{z}_n) d\mathbf{z}_n \\ &= \int p(\mathbf{x}_n|\mathbf{z}_n, \mathbf{X}_{n-1}) \cdot p(\mathbf{z}_n, \mathbf{X}_{n-1}) d\mathbf{z}_n\end{aligned}$$

$$\begin{aligned}p(\mathbf{z}_n|\mathbf{x}_n, \mathbf{X}_{n-1}) &= \frac{p(\mathbf{x}_n|\mathbf{z}_n) \cdot p(\mathbf{z}_n|\mathbf{X}_{n-1})}{\int p(\mathbf{x}_n|\mathbf{z}_n) \cdot \frac{p(\mathbf{z}_n, X_{n-1})}{p(\mathbf{X}_{n-1})} d\mathbf{z}_n} \\ &= \frac{p(\mathbf{x}_n|\mathbf{z}_n) \cdot p(\mathbf{z}_n|\mathbf{X}_{n-1})}{\int p(\mathbf{x}_n|\mathbf{z}_n) \cdot p(\mathbf{z}_n|\mathbf{X}_{n-1}) d\mathbf{z}_n}\end{aligned}$$

$$\Rightarrow \quad \mathbb{E}[f(\mathbf{z}_n)] = \frac{\int f(\mathbf{z}_n) \cdot p(\mathbf{x}_n|\mathbf{z}_n) \cdot p(\mathbf{z}_n|\mathbf{X}_{n-1}) d\mathbf{z}_n}{\int p(\mathbf{x}_n|\mathbf{z}_n) \cdot p(\mathbf{z}_n|\mathbf{X}_{n-1}) d\mathbf{z}_n}$$

When a set of samples $\{\mathbf{z}_n^{(l)}\}$ is drawn from $p(\mathbf{z}_n|\mathbf{X}_{n-1})$ distribution, then $f(\mathbf{z}_n) \cdot p(\mathbf{z}_n|\mathbf{X}_{n-1})$ collapses to $f(\mathbf{z}_n(l))$.
Utilizing

$$\mathbf{w}_n^{(l)} = \frac{p(\mathbf{x}_n|\mathbf{z}_n^{(l)})}{\sum_{m=1}^{L} p(\mathbf{x}_n|\mathbf{z}_n^{(m)})}$$

$$\therefore \;\; \mathbb{E}[f(\mathbf{z}_n)] \simeq \sum_{l=1}^{L} \mathbf{w}_n^{(l)} \cdot f(\mathbf{z}_n^{(l)})$$

Eq 13.119: (PRML p.646)

$$p(\mathbf{z}_{n+1}|\mathbf{X}_n) = \int p(\mathbf{z}_{n+1}|\mathbf{z}_n, \mathbf{X}_n) \cdot p(\mathbf{z}_n|\mathbf{X}_n) d\mathbf{z}_n$$
$$\simeq \sum_l \mathbf{w}_n^{(l)} [(\mathbf{z}_{n+1}|\mathbf{z}_n^{(l)}$$

Proof :

$$p(\mathbf{z}_{n+1}|\mathbf{X}_n) = \frac{p(\mathbf{z}_{n+1}, \mathbf{X}_n)}{p(\mathbf{X}_n)} = \frac{\int p(\mathbf{z}_{n+1}, \mathbf{X}_n, \mathbf{z}_n) d\mathbf{z}_n}{p(\mathbf{X}_n)}$$

Since

$$p(\mathbf{z}_{n+1}, \mathbf{X}_n, \mathbf{z}_n) = p(\mathbf{z}_{n+1}|\mathbf{z}_n, \mathbf{X}_n) \cdot p(\mathbf{z}_n, \mathbf{X}_n)$$
$$p(\mathbf{z}_{n+1}|\mathbf{X}_n) = \int p(\mathbf{z}_{n+1}|\mathbf{z}_n, \mathbf{X}_n) \cdot p(\mathbf{z}_n|\mathbf{X}_n) d\mathbf{z}_n$$

Since

$$\begin{aligned}
\frac{p(\mathbf{z}_{n+1}, \mathbf{z}_n, \mathbf{X}_n)}{p(\mathbf{X}_n, \mathbf{z}_n)} &= \frac{p(\mathbf{z}_{n+1}, \mathbf{X}_n|\mathbf{z}_n) \cdot p(\mathbf{z}_n)}{p(\mathbf{z}_n, \mathbf{X}_n)} \\
&= \frac{[p(\mathbf{z}_{N+1}|\mathbf{z}_n) \cdot p(\mathbf{X}_n|\mathbf{z}_n)] \cdot p(\mathbf{z}_n)}{p(\mathbf{z}_n, \mathbf{X}_n)} \\
&= p(\mathbf{z}_{n+1}|\mathbf{z}_n)
\end{aligned}$$

$$\begin{aligned}
p(\mathbf{z}_{n+1}|\mathbf{X}_n) &= \int p(\mathbf{z}_{n+1}|\mathbf{z}_n) \cdot p(\mathbf{z}_n|\mathbf{X}_n) d\mathbf{z}_n \\
&= \int p(\mathbf{z}_{n+1}|\mathbf{z}_n) \cdot p(\mathbf{z}_n|\mathbf{x}_n, \mathbf{X}_{n-1}) d\mathbf{z}_n \\
&\quad \text{(utilizaing the third line in Eq (13.117)),} \\
&= \frac{\int p(\mathbf{z}_{n+1}|\mathbf{z}_n) \cdot p(\mathbf{x}_n|\mathbf{z}_n) \cdot p(\mathbf{z}_n|X_{n-1}) d\mathbf{z}_n}{\int p(\mathbf{x}_n|\mathbf{z}_n) \cdot p(\mathbf{z}_n|\mathbf{X}_{n-1}) d\mathbf{z}_n}
\end{aligned}$$

Since we sampled $\mathbf{z}_n^{(l)}$ from $p(\mathbf{z}_n|\mathbf{X}_{n-1})$, this equation can be written as,

$$\therefore \quad p(\mathbf{z}_{n+1}|\mathbf{X}_n) \simeq \sum_l \mathbf{w}_n^{(l)} p(\mathbf{z}_{n+1}|\mathbf{z}_n^{(l)})$$

Chapter 14. Combining Models

Eq 14.24: (PRML p.661)

$$w_n^{(m+1)} = w_n^{(m)}\exp\left\{-\frac{1}{2}t_n\alpha_m y_m(\mathbf{x}_n)\right\}$$

Proof :

$$\begin{aligned} w_n^{(m)} &= \exp\{-t_n f_{m-1}(\mathbf{x}_n)\} \quad \longleftarrow \text{ defined below Eq (14.22).} \\ w_n^{(m+1)} &= \exp\{-t_n f_m(\mathbf{x}_n)\} \end{aligned}$$

where

$$f_m(\mathbf{x}_n) = \frac{1}{2}\sum_{l=1}^{m}\alpha_l y_l(\mathbf{x}_n) \tag{14.21}$$

$$\begin{aligned} \Rightarrow \quad w_n^{(m+1)} &= \exp\left\{-t_n f_{m-1}(\mathbf{x}_n) - \frac{1}{2}t_n\alpha_m y_m(\mathbf{x}_n)\right\} \\ &= w_n^{(m)}\exp\left\{-\frac{1}{2}t_n\alpha_m y_m(\mathbf{x}_n)\right\} \end{aligned}$$

Eq 14.51: (PRML p.672)

$$\nabla_k Q = \sum_{n=1}^{N}\gamma_{nk}(t_n - y_{nk})\boldsymbol{\phi}_n$$

Proof :

We would like to find w_k that makes Q minimum. To do that we have to solve $\nabla_{w_k}Q = 0$. However, we do not have a closed form of solution for this, so we rely on IRLS algo for the iterative method.

As explained in §4.3.3, using Newton-Raphson method we can find $w_k^{(new)}$ from the following equation (Eq (4.92)).

$$\mathbf{w}_k^{(new)} = \mathbf{w}^{(old)} - \mathbf{H}^{-1}\nabla Q(\mathbf{w}_k)$$

Now let's calculate $\nabla_k Q$,

$$Q = \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}\{\ln\pi_k + t_n\ln y_{nk} + (1-t_n)\ln(1-y_{nk})\} \tag{14.49}$$

To calculate $\dfrac{\partial Q}{\partial \mathbf{w}_k}$, we need to calculate $\dfrac{\partial \ln y_{nk}}{\partial \mathbf{w}_k}$ and $\dfrac{\partial \ln(1-y_{nk})}{\partial \mathbf{w}_k}$.

Since $y_{nk} = \sigma(\mathbf{w}_k^T \boldsymbol{\phi}_n)$,

$$\begin{aligned}
\frac{\partial y_{nk}}{\partial \mathbf{w}_k} &= \frac{\partial \sigma(\mathbf{w}_k^T \boldsymbol{\phi}_n)}{\partial \mathbf{w}_k} \cdot \frac{\partial (\mathbf{w}_k^T \boldsymbol{\phi}_n)}{\partial \mathbf{w}_k} = \sigma(1-\sigma)\boldsymbol{\phi}_n \\
\frac{\partial \ln y_{nk}}{\partial \mathbf{w}_k} &= \frac{1}{y_{nk}} \cdot \frac{\partial y_{nk}}{\partial \mathbf{w}_k} = \frac{1}{y_{nk}} \sigma(1-\sigma)\boldsymbol{\phi}_n \\
\frac{\partial \ln(1-y_{nk})}{\partial \mathbf{w}_k} &= \frac{(-1)\cdot\sigma(1-\sigma)\boldsymbol{\phi}_n}{1-y_{nk}}
\end{aligned}$$

$$\begin{aligned}
\Rightarrow \quad \frac{\partial Q}{\partial \mathbf{w}_k} &= \sum_{n=1}^{N} \gamma_{nk} \left\{ t_n \frac{1}{y_{nk}} \cdot y_{nk}(1-y_{nk})\boldsymbol{\phi}_n - (1-t_n) \cdot \frac{y_{nk}(1-y_{nk})}{1-y_{nk}} \boldsymbol{\phi}_n \right\} \\
&= \sum_{n=1}^{N} \gamma_{nk} \{ t_n(1-y_{nk}) - (1-t_n) y_{nk} \} \boldsymbol{\phi}_n \\
&= \sum_{n=1}^{N} \gamma_{nk} (t_n - y_{nk}) \boldsymbol{\phi}_n
\end{aligned}$$