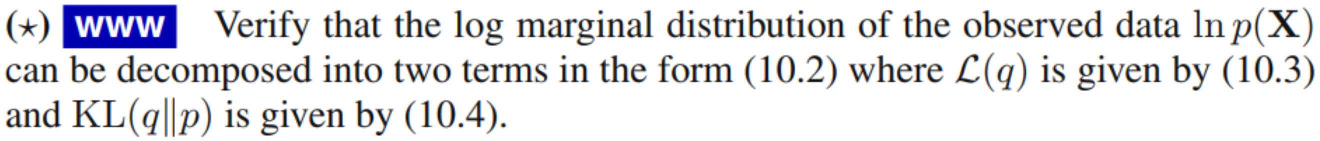
# ECE1513
# Tutorial 8:
# Selected Exercises from Chapters 10 and 11

TA: Faye Pourghasem, fateme.pourghasem@mail.utoronto.ca

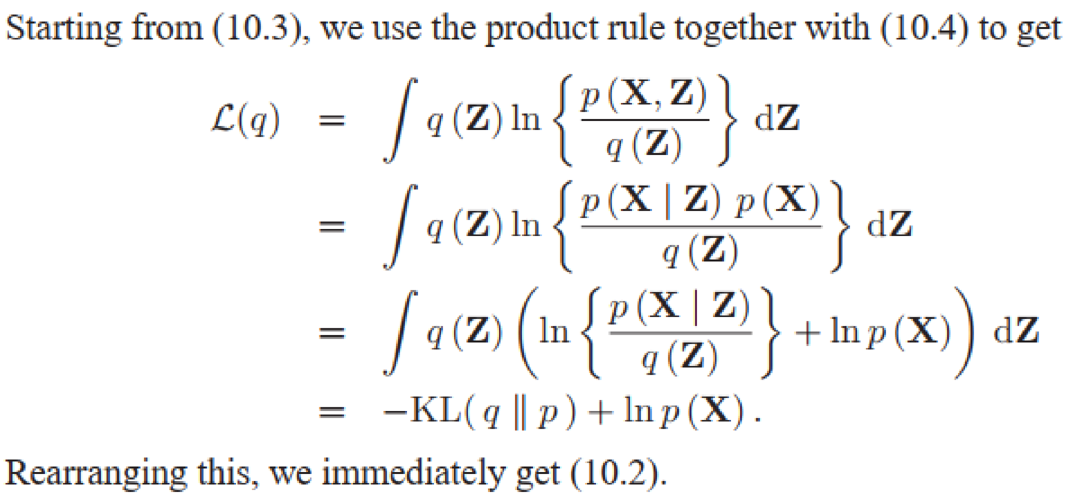
November 6, 2023

## 1 Example 10.1



($\star$) www Verify that the log marginal distribution of the observed data $\ln p(\mathbf{X})$ can be decomposed into two terms in the form (10.2) where $\mathcal{L}(q)$ is given by (10.3) and $\mathrm{KL}(q\|p)$ is given by (10.4).

### Solution



Starting from (10.3), we use the product rule together with (10.4) to get

$$\begin{aligned}\mathcal{L}(q) &= \int q(\mathbf{Z}) \ln\left\{\frac{p(\mathbf{X},\mathbf{Z})}{q(\mathbf{Z})}\right\} \mathrm{d}\mathbf{Z} \\ &= \int q(\mathbf{Z}) \ln\left\{\frac{p(\mathbf{X}\mid\mathbf{Z})\, p(\mathbf{X})}{q(\mathbf{Z})}\right\} \mathrm{d}\mathbf{Z} \\ &= \int q(\mathbf{Z}) \left(\ln\left\{\frac{p(\mathbf{X}\mid\mathbf{Z})}{q(\mathbf{Z})}\right\} + \ln p(\mathbf{X})\right) \mathrm{d}\mathbf{Z} \\ &= -\mathrm{KL}(\,q \,\|\, p\,) + \ln p(\mathbf{X})\,.\end{aligned}$$

Rearranging this, we immediately get (10.2).

## 2 Example 10.5

($\star\star$) www Consider a model in which the set of all hidden stochastic variables, denoted collectively by $\mathbf{Z}$, comprises some latent variables $\mathbf{z}$ together with some model parameters $\boldsymbol{\theta}$. Suppose we use a variational distribution that factorizes between latent variables and parameters so that $q(\mathbf{z}, \boldsymbol{\theta}) = q_{\mathbf{z}}(\mathbf{z})q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$, in which the distribution $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ is approximated by a point estimate of the form $q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$ where $\boldsymbol{\theta}_0$ is a vector of free parameters. Show that variational optimization of this factorized distribution is equivalent to an EM algorithm, in which the E step optimizes $q_{\mathbf{z}}(\mathbf{z})$, and the M step maximizes the expected complete-data log posterior distribution of $\boldsymbol{\theta}$ with respect to $\boldsymbol{\theta}_0$.

## Solution

We assume that $q(\mathbf{Z}) = q(\mathbf{z})q(\boldsymbol{\theta})$ and so we can optimize w.r.t. $q(\mathbf{z})$ and $q(\boldsymbol{\theta})$ independently.

For $q(\mathbf{z})$, this is equivalent to minimizing the Kullback-Leibler divergence, (10.4), which here becomes

$$\mathrm{KL}(\,q\,\|\,p\,) = -\iint q\left(\boldsymbol{\theta}\right) q\left(\mathbf{z}\right) \ln \frac{p\left(\mathbf{z}, \boldsymbol{\theta} \mid \mathbf{X}\right)}{q\left(\mathbf{z}\right) q\left(\boldsymbol{\theta}\right)} \,\mathrm{d}\mathbf{z}\,\mathrm{d}\boldsymbol{\theta}.$$

For the particular chosen form of $q(\boldsymbol{\theta})$, this is equivalent to

$$\begin{aligned}
\mathrm{KL}(\,q\,\|\,p\,) &= -\int q\left(\mathbf{z}\right) \ln \frac{p\left(\mathbf{z}, \boldsymbol{\theta}_0 \mid \mathbf{X}\right)}{q\left(\mathbf{z}\right)} \,\mathrm{d}\mathbf{z} + \text{const.} \\
&= -\int q\left(\mathbf{z}\right) \ln \frac{p\left(\mathbf{z} \mid \boldsymbol{\theta}_0, \mathbf{X}\right) p\left(\boldsymbol{\theta}_0 \mid \mathbf{X}\right)}{q\left(\mathbf{z}\right)} \,\mathrm{d}\mathbf{z} + \text{const.} \\
&= -\int q\left(\mathbf{z}\right) \ln \frac{p\left(\mathbf{z} \mid \boldsymbol{\theta}_0, \mathbf{X}\right)}{q\left(\mathbf{z}\right)} \,\mathrm{d}\mathbf{z} + \text{const.},
\end{aligned}$$

where const accumulates all terms independent of $q(\mathbf{z})$. This KL divergence is minimized when $q(\mathbf{z}) = p(\mathbf{z}|\boldsymbol{\theta}_0, \mathbf{X})$, which corresponds exactly to the E-step of the EM algorithm.

To determine $q(\boldsymbol{\theta})$, we consider

$$\begin{aligned}
&\int q\left(\boldsymbol{\theta}\right) \int q\left(\mathbf{z}\right) \ln \frac{p\left(\mathbf{X}, \boldsymbol{\theta}, \mathbf{z}\right)}{q\left(\boldsymbol{\theta}\right) q\left(\mathbf{z}\right)} \,\mathrm{d}\mathbf{z} \,\mathrm{d}\boldsymbol{\theta} \\
&= \int q\left(\boldsymbol{\theta}\right) \mathbb{E}_{q(\mathbf{z})}\left[\ln p\left(\mathbf{X}, \boldsymbol{\theta}, \mathbf{z}\right)\right] \,\mathrm{d}\boldsymbol{\theta} - \int q\left(\boldsymbol{\theta}\right) \ln q\left(\boldsymbol{\theta}\right) \,\mathrm{d}\boldsymbol{\theta} + \text{const.}
\end{aligned}$$

where the last term summarizes terms independent of $q\left(\boldsymbol{\theta}\right)$. Since $q(\boldsymbol{\theta})$ is constrained to be a point density, the contribution from the entropy term (which formally diverges) will be constant and independent of $\boldsymbol{\theta}_0$. Thus, the optimization problem is reduced to maximizing expected complete log posterior distribution

$$\mathbb{E}_{q(\mathbf{z})}\left[\ln p\left(\mathbf{X}, \boldsymbol{\theta}_0, \mathbf{z}\right)\right],$$

w.r.t. $\boldsymbol{\theta}_0$, which is equivalent to the M-step of the EM algorithm.

## 3 Example 10.10

($\star$) **www** Derive the decomposition given by (10.34) that is used to find approximate posterior distributions over models using variational inference.

### Solution

**NOTE**: In the 1st printing of PRML, there are errors that affect this exercise. $\mathcal{L}_m$ used in (10.34) and (10.35) should really be $\mathcal{L}$, whereas $\mathcal{L}_m$ used in (10.36) is given in Solution 10.11 below.

This completely analogous to Solution 10.1. Starting from (10.35), we can use the product rule to get,

$$\begin{aligned}
\mathcal{L} &= \sum_m \sum_{\mathbf{Z}} q(\mathbf{Z}|m)q(m) \ln \left\{ \frac{p(\mathbf{Z}, \mathbf{X}, m)}{q(\mathbf{Z}|m)\, q(m)} \right\} \\
&= \sum_m \sum_{\mathbf{Z}} q(\mathbf{Z}|m)q(m) \ln \left\{ \frac{p(\mathbf{Z}, m|\mathbf{X})\, p(\mathbf{X})}{q(\mathbf{Z}|m)\, q(m)} \right\} \\
&= \sum_m \sum_{\mathbf{Z}} q(\mathbf{Z}|m)q(m) \ln \left\{ \frac{p(\mathbf{Z}, m|\mathbf{X})}{q(\mathbf{Z}|m)\, q(m)} \right\} + \ln p(\mathbf{X}).
\end{aligned}$$

Rearranging this, we obtain (10.34).

## 4 Example 10.16

(⋆⋆) **www** Verify the results (10.71) and (10.72) for the first two terms in the lower bound for the variational Gaussian mixture model given by (10.70).

Solution

To derive (10.71) we make use of (10.38) to give

$$
\begin{aligned}
&\mathbb{E}[\ln p(D|\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] \\
&\quad = \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{E}[z_{nk}]\left\{\mathbb{E}[\ln|\boldsymbol{\Lambda}_k|] - \mathbb{E}[(\mathbf{x}_n - \boldsymbol{\mu}_k)\boldsymbol{\Lambda}_k(\mathbf{x}_n - \boldsymbol{\mu}_k)] - D\ln(2\pi)\right\}.
\end{aligned}
$$

We now use $\mathbb{E}[z_{nk}] = r_{nk}$ together with (10.64) and the definition of $\widetilde{\Lambda}_k$ given by (10.65) to give

$$
\begin{aligned}
\mathbb{E}[\ln p(D|\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] = \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\big\{\ln \widetilde{\Lambda}_k \\
-D\beta_k^{-1} - \nu_k(\mathbf{x}_n - \mathbf{m}_k)^{\mathrm{T}}\mathbf{W}_k(\mathbf{x}_n - \mathbf{m}_k) - D\ln(2\pi)\big\}.
\end{aligned}
$$

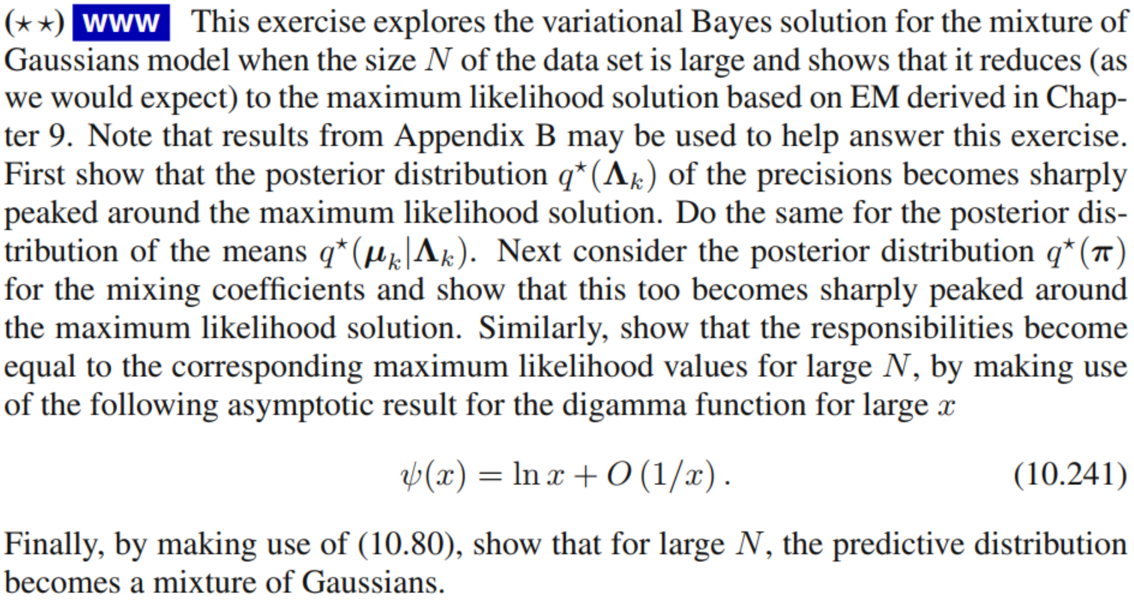
Now we use the definitions (10.51) to (10.53) together with the result (268) to give (10.71).

We can derive (10.72) simply by taking the logarithm of $p(\mathbf{z}|\boldsymbol{\pi})$ given by (10.37)

$$
\mathbb{E}[\ln p(\mathbf{z}|\boldsymbol{\pi})] = \sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{E}[z_{nk}]\mathbb{E}[\ln \pi_k]
$$

and then making use of $\mathbb{E}[z_{nk}] = r_{nk}$ together with the definition of $\widetilde{\pi}_k$ given by (10.65).

## 5 Example 10.20



($\star\star$) www This exercise explores the variational Bayes solution for the mixture of Gaussians model when the size $N$ of the data set is large and shows that it reduces (as we would expect) to the maximum likelihood solution based on EM derived in Chapter 9. Note that results from Appendix B may be used to help answer this exercise. First show that the posterior distribution $q^\star(\mathbf{\Lambda}_k)$ of the precisions becomes sharply peaked around the maximum likelihood solution. Do the same for the posterior distribution of the means $q^\star(\boldsymbol{\mu}_k|\mathbf{\Lambda}_k)$. Next consider the posterior distribution $q^\star(\boldsymbol{\pi})$ for the mixing coefficients and show that this too becomes sharply peaked around the maximum likelihood solution. Similarly, show that the responsibilities become equal to the corresponding maximum likelihood values for large $N$, by making use of the following asymptotic result for the digamma function for large $x$

$$\psi(x) = \ln x + O\left(1/x\right). \qquad (10.241)$$

Finally, by making use of (10.80), show that for large $N$, the predictive distribution becomes a mixture of Gaussians.

# Solution

Consider first the posterior distribution over the precision of component $k$ given by

$$q^{\star}(\mathbf{\Lambda}_k) = \mathcal{W}(\mathbf{\Lambda}_k|\mathbf{W}_k, \nu_k).$$

From (10.63) we see that for large $N$ we have $\nu_k \rightarrow N_k$, and similarly from (10.62) we see that $\mathbf{W}_k \rightarrow N_k^{-1}\mathbf{S}_k^{-1}$. Thus the mean of the distribution over $\mathbf{\Lambda}_k$, given by $\mathrm{E}[\mathbf{\Lambda}_k] = \nu_k \mathbf{W}_k \rightarrow \mathbf{S}_k^{-1}$ which is the maximum likelihood value (this assumes that the quantities $r_{nk}$ reduce to the corresponding EM values, which is indeed the case as we shall show shortly). In order to show that this posterior is also sharply peaked, we consider the differential entropy, $\mathrm{H}[\mathbf{\Lambda}_k]$ given by (B.82), and show that, as $N_k \rightarrow \infty$, $\mathrm{H}[\mathbf{\Lambda}_k] \rightarrow 0$, corresponding to the density collapsing to a spike. First consider the normalizing constant $B(\mathbf{W}_k, \nu_k)$ given by (B.79). Since $\mathbf{W}_k \rightarrow N_k^{-1}\mathbf{S}_k^{-1}$ and $\nu_k \rightarrow N_k$,

$$-\ln B(\mathbf{W}_k, \nu_k) \rightarrow -\frac{N_k}{2}\left(D \ln N_k + \ln |\mathbf{S}_k| - D \ln 2\right) + \sum_{i=1}^{D} \ln \Gamma\left(\frac{N_k + 1 - i}{2}\right).$$

We then make use of Stirling's approximation (1.146) to obtain

$$\ln \Gamma\left(\frac{N_k + 1 - i}{2}\right) \simeq \frac{N_k}{2}\left(\ln N_k - \ln 2 - 1\right)$$

which leads to the approximate limit

$$\begin{aligned} -\ln B(\mathbf{W}_k, \nu_k) &\rightarrow -\frac{N_k D}{2}\left(\ln N_k - \ln 2 - \ln N_k + \ln 2 + 1\right) - \frac{N_k}{2} \ln |\mathbf{S}_k| \\ &= -\frac{N_k}{2}\left(\ln |\mathbf{S}_k| + D\right). \end{aligned} \tag{280}$$

Next, we use (10.241) and (B.81) in combination with $\mathbf{W}_k \rightarrow N_k^{-1}\mathbf{S}_k^{-1}$ and $\nu_k \rightarrow N_k$ to obtain the limit

$$\begin{aligned} \mathbb{E}\left[\ln |\mathbf{\Lambda}|\right] &\rightarrow D \ln \frac{N_k}{2} + D \ln 2 - D \ln N_k - \ln |\mathbf{S}_k| \\ &= -\ln |\mathbf{S}_k|, \end{aligned}$$

where we approximated the argument to the digamma function by $N_k/2$. Substituting this and (280) into (B.82), we get

$$\mathrm{H}[\mathbf{\Lambda}] \rightarrow 0$$

when $N_k \rightarrow \infty$.

Next consider the posterior distribution over the mean $\boldsymbol{\mu}_k$ of the $k^{\text{th}}$ component given by

$$q^{\star}(\boldsymbol{\mu}_k|\mathbf{\Lambda}_k) = \mathcal{N}(\boldsymbol{\mu}_k|\mathbf{m}_k, \beta_k \mathbf{\Lambda}_k).$$

From (10.61) we see that for large $N$ the mean $\mathbf{m}_k$ of this distribution reduces to $\overline{\mathbf{x}}_k$ which is the corresponding maximum likelihood value. From (10.60) we see that

$\beta_k \to N_k$ and Thus the precision $\beta_k \mathbf{\Lambda}_k \to \beta_k \nu_k \mathbf{W}_k \to N_k \mathbf{S}_k^{-1}$ which is large for large $N$ and hence this distribution is sharply peaked around its mean.

Now consider the posterior distribution $q^\star(\boldsymbol{\pi})$ given by (10.57). For large $N$ we have $\alpha_k \to N_k$ and so from (B.17) and (B.19) we see that the posterior distribution becomes sharply peaked around its mean $\mathrm{E}[\pi_k] = \alpha_k/\overline{\alpha} \to N_k/N$ which is the maximum likelihood solution.

For the distribution $q^\star(\mathbf{z})$ we consider the responsibilities given by (10.67). Using (10.65) and (10.66), together with the asymptotic result for the digamma function, we again obtain the maximum likelihood expression for the responsibilities for large $N$.

Finally, for the predictive distribution we first perform the integration over $\boldsymbol{\pi}$, as in the solution to Exercise 10.19, to give

$$p(\widehat{\mathbf{x}}|D) = \sum_{k=1}^{K} \frac{\alpha_k}{\overline{\alpha}} \iint \mathcal{N}(\widehat{\mathbf{x}}|\boldsymbol{\mu}_k, \mathbf{\Lambda}_k) q(\boldsymbol{\mu}_k, \mathbf{\Lambda}_k) \, \mathrm{d}\boldsymbol{\mu}_k \, \mathrm{d}\mathbf{\Lambda}_k.$$

The integrations over $\boldsymbol{\mu}_k$ and $\mathbf{\Lambda}_k$ are then trivial for large $N$ since these are sharply peaked and hence approximate delta functions. We therefore obtain

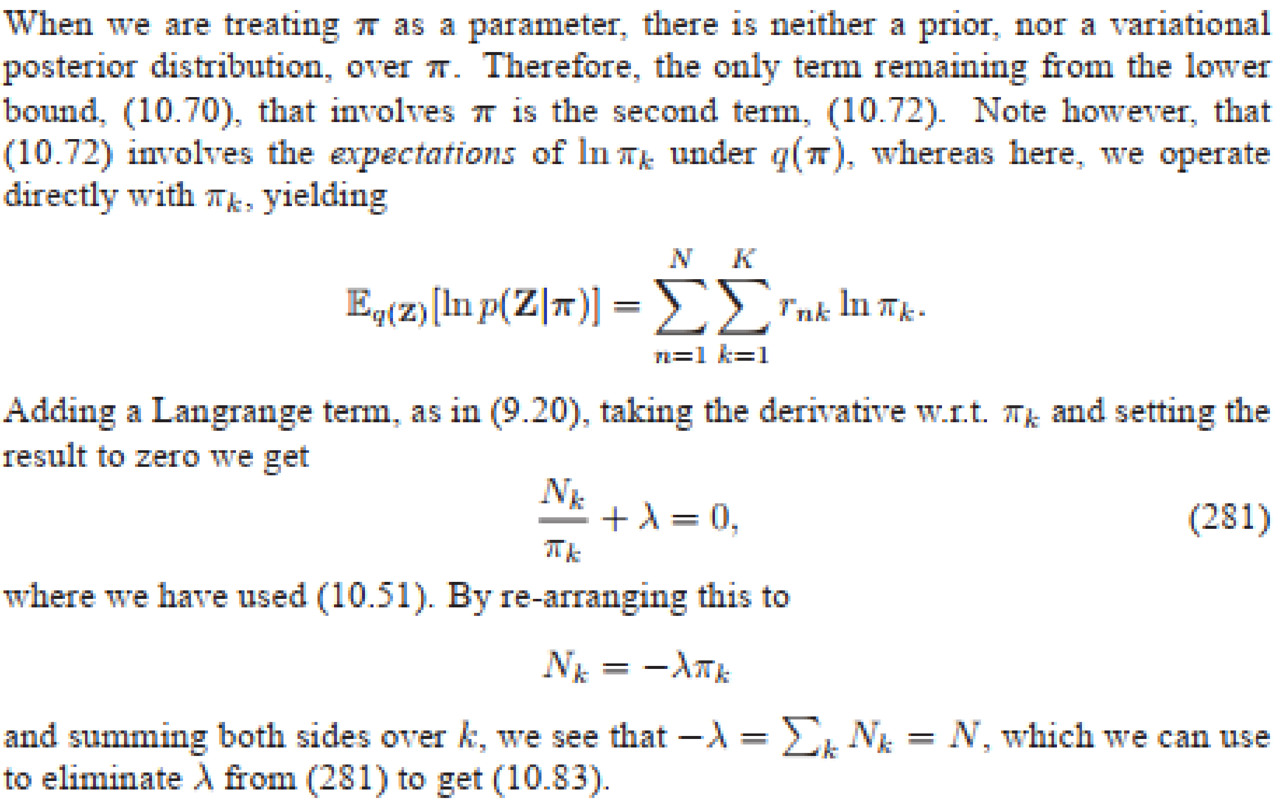
$$p(\widehat{\mathbf{x}}|D) = \sum_{k=1}^{K} \frac{N_k}{N} \mathcal{N}(\widehat{\mathbf{x}}|\overline{\mathbf{x}}_k, \mathbf{W}_k)$$

which is a mixture of Gaussians, with mixing coefficients given by $N_k/N$.

## 6 Example 10.23

(⋆⋆) **www** Consider a variational Gaussian mixture model in which there is no prior distribution over mixing coefficients $\{\pi_k\}$. Instead, the mixing coefficients are treated as parameters, whose values are to be found by maximizing the variational lower bound on the log marginal likelihood. Show that maximizing this lower bound with respect to the mixing coefficients, using a Lagrange multiplier to enforce the constraint that the mixing coefficients sum to one, leads to the re-estimation result (10.83). Note that there is no need to consider all of the terms in the lower bound but only the dependence of the bound on the $\{\pi_k\}$.

Solution



When we are treating $\boldsymbol{\pi}$ as a parameter, there is neither a prior, nor a variational posterior distribution, over $\boldsymbol{\pi}$. Therefore, the only term remaining from the lower bound, (10.70), that involves $\boldsymbol{\pi}$ is the second term, (10.72). Note however, that (10.72) involves the *expectations* of $\ln \pi_k$ under $q(\boldsymbol{\pi})$, whereas here, we operate directly with $\pi_k$, yielding

$$\mathbb{E}_{q(\mathbf{Z})}[\ln p(\mathbf{Z}|\boldsymbol{\pi})] = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk} \ln \pi_k.$$

Adding a Langrange term, as in (9.20), taking the derivative w.r.t. $\pi_k$ and setting the result to zero we get

$$\frac{N_k}{\pi_k} + \lambda = 0, \qquad (281)$$

where we have used (10.51). By re-arranging this to

$$N_k = -\lambda \pi_k$$

and summing both sides over $k$, we see that $-\lambda = \sum_k N_k = N$, which we can use to eliminate $\lambda$ from (281) to get (10.83).

## 7 Example 11.1

($\star$) **www** Show that the finite sample estimator $\widehat{f}$ defined by (11.2) has mean equal to $\mathbb{E}[f]$ and variance given by (11.3).

Solution

Since the samples are independent, for the mean, we have

$$\mathbb{E}\left[\widehat{f}\right] = \frac{1}{L}\sum_{l=1}^{L}\int f(z^{(l)})p(z^{(l)})\,\mathrm{d}z^{(l)} = \frac{1}{L}\sum_{l=1}^{L}\mathbb{E}\left[f\right] = \mathbb{E}\left[f\right].$$

Using this together with (1.38) and (1.39), for the variance, we have

$$\begin{aligned}\operatorname{var}\left[\widehat{f}\right] &= \mathbb{E}\left[\left(\widehat{f} - \mathbb{E}\left[\widehat{f}\right]\right)^2\right] \\ &= \mathbb{E}\left[\widehat{f}^2\right] - \mathbb{E}\left[f\right]^2.\end{aligned}$$

Now note

$$\begin{aligned}\mathbb{E}\left[f(z^{(k)}), f(z^{(m)})\right] &= \begin{cases}\operatorname{var}[f] + \mathbb{E}[f^2] & \text{if } n = k, \\ \mathbb{E}[f^2] & \text{otherwise,}\end{cases} \\ &= \mathbb{E}[f^2] + \delta_{mk}\operatorname{var}[f],\end{aligned}$$

where we again exploited the fact that the samples are independent.

Hence

$$\begin{aligned}\operatorname{var}\left[\widehat{f}\right] &= \mathbb{E}\left[\frac{1}{L}\sum_{m=1}^{L} f(z^{(m)})\frac{1}{L}\sum_{k=1}^{L} f(z^{(k)})\right] - \mathbb{E}[f]^2 \\ &= \frac{1}{L^2}\sum_{m=1}^{L}\sum_{k=1}^{L}\left\{\mathbb{E}[f^2] + \delta_{mk}\operatorname{var}[f]\right\} - \mathbb{E}[f]^2 \\ &= \frac{1}{L}\operatorname{var}[f] \\ &= \frac{1}{L}\mathbb{E}\left[(f - \mathbb{E}\left[f\right])^2\right].\end{aligned}$$

## 8 Example 11.5

($\star$) www Let $\mathbf{z}$ be a $D$-dimensional random variable having a Gaussian distribution with zero mean and unit covariance matrix, and suppose that the positive definite symmetric matrix $\mathbf{\Sigma}$ has the Cholesky decomposition $\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$ where $\mathbf{L}$ is a lower-triangular matrix (i.e., one with zeros above the leading diagonal). Show that the variable $\mathbf{y} = \boldsymbol{\mu} + \mathbf{L}\mathbf{z}$ has a Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\mathbf{\Sigma}$. This provides a technique for generating samples from a general multivariate Gaussian using samples from a univariate Gaussian having zero mean and unit variance.

Solution

Since $\mathbb{E}\left[\mathbf{z}\right] = \mathbf{0}$,

$$\mathbb{E}\left[\mathbf{y}\right] = \mathbb{E}\left[\boldsymbol{\mu} + \mathbf{L}\mathbf{z}\right] = \boldsymbol{\mu}.$$

Similarly, since $\mathbb{E}\left[\mathbf{z}\mathbf{z}^{\mathrm{T}}\right] = \mathbf{I}$,

$$\begin{aligned}
\operatorname{cov}\left[\mathbf{y}\right] &= \mathbb{E}\left[\mathbf{y}\mathbf{y}^{\mathrm{T}}\right] - \mathbb{E}\left[\mathbf{y}\right]\mathbb{E}\left[\mathbf{y}^{\mathrm{T}}\right] \\
&= \mathbb{E}\left[\left(\boldsymbol{\mu} + \mathbf{L}\mathbf{z}\right)\left(\boldsymbol{\mu} + \mathbf{L}\mathbf{z}\right)^{\mathrm{T}}\right] - \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} \\
&= \mathbf{L}\mathbf{L}^{\mathrm{T}} \\
&= \mathbf{\Sigma}.
\end{aligned}$$

# 9 Example 11.6

$(\star\star)$ www In this exercise, we show more carefully that rejection sampling does indeed draw samples from the desired distribution $p(\mathbf{z})$. Suppose the proposal distribution is $q(\mathbf{z})$ and show that the probability of a sample value $\mathbf{z}$ being accepted is given by $\widetilde{p}(\mathbf{z})/kq(\mathbf{z})$ where $\widetilde{p}$ is any unnormalized distribution that is proportional to $p(\mathbf{z})$, and the constant $k$ is set to the smallest value that ensures $kq(\mathbf{z}) \geqslant \widetilde{p}(\mathbf{z})$ for all values of $\mathbf{z}$. Note that the probability of drawing a value $\mathbf{z}$ is given by the probability of drawing that value from $q(\mathbf{z})$ times the probability of accepting that value given that it has been drawn. Make use of this, along with the sum and product rules of probability, to write down the normalized form for the distribution over $\mathbf{z}$, and show that it equals $p(\mathbf{z})$.

## Solution

The probability of acceptance follows directly from the mechanism used to accept or reject the sample. The probability of a sample $\mathbf{z}$ being accepted equals the probability of a sample $u$, drawn uniformly from the interval $[0, kq(\mathbf{z})]$, being less than or equal to a value $\widetilde{p}(\mathbf{z}) \leqslant kq(\mathbf{z})$, and is given by is given by

$$p(\text{acceptance}|\mathbf{z}) = \int_0^{\widetilde{p}(\mathbf{z})} \frac{1}{kq(\mathbf{z})}\,\mathrm{d}u = \frac{\widetilde{p}(\mathbf{z})}{kq(\mathbf{z})}.$$

Therefore, the probability of drawing a sample, $\mathbf{z}$, is

$$q(\mathbf{z})p(\text{acceptance}|\mathbf{z}) = q(\mathbf{z})\frac{\widetilde{p}(\mathbf{z})}{kq(\mathbf{z})} = \frac{\widetilde{p}(\mathbf{z})}{k}. \qquad (308)$$

Integrating both sides w.r.t. $\mathbf{z}$, we see that $kp(\text{acceptance}) = Z_p$, where

$$Z_p = \int \widetilde{p}(\mathbf{z})\,\mathrm{d}\mathbf{z}.$$

Combining this with (308) and (11.13), we obtain

$$\frac{q(\mathbf{z})p(\text{acceptance}|\mathbf{z})}{p(\text{acceptance})} = \frac{1}{Z_p}\widetilde{p}(\mathbf{z}) = p(\mathbf{z})$$

as required.

## 10 Example 11.15

($\star$) www Using (11.56) and (11.57), show that the Hamiltonian equation (11.58) is equivalent to (11.53). Similarly, using (11.57) show that (11.59) is equivalent to (11.55).

### Solution

Using (11.56), we can differentiate (11.57), yielding

$$\frac{\partial H}{\partial r_i} = \frac{\partial K}{\partial r_i} = r_i$$

and thus (11.53) and (11.58) are equivalent.

Similarly, differentiating (11.57) w.r.t. $z_i$ we get

$$\frac{\partial H}{\partial z_i} = \frac{\partial E}{\partial z_i},$$

and from this, it is immediately clear that (11.55) and (11.59) are equivalent.