7 Variational Bayes

Variational Bayes (VB) was developed in the machine learning community in the 1990s (Attias 1999, Jordan, Ghahramani, Jaakkola *et al*. 1999) and has now become a standard technique to approximated Bayesian inference for latent models, based on the EM-like algorithm. In Chapter 4, we have also dealt with latent models based on the maximum a-posteriori (MAP) EM algorithm. However, the MAP approximation uses the point estimation of model parameters instead of the distribution estimation, which is far from a true Bayesian manner of regarding all the variables introduced in our problem as probabilistic random variables. Another approximation based on the asymptotic approximation in Chapter 6 assumes a complex posterior distribution as a single Gaussian distribution without latent variables, which is not a true assumption for many of our applications. The evidence approximation in Chapter 5 also does not explicitly deal with latent models (can be obtained by combining MAP, VB, or MCMC). Instead of considering the MAP, evidence, and asymptotic approximations, VB can efficiently approximate complicated integrals and expectations over model parameters, based on variational method within a specific family of distribution types (exponential family, as discussed in Section 2.1.3). The key idea of the variational technique is to find the lower bound of the marginal log likelihood, similar to the EM algorithm in Section 3.4, and obtain the posterior distributions directly based on the variational method.

This chapter starts to explain the general framework of VB in Section 7.1, and more specific pattern recognition problems in Section 7.2. Then this chapter goes on to provide a VB version of the EM algorithm for statistical models and model selection in speech and language processing, including speech recognition in Sections 7.3 and 7.4 and speaker verification in Section 7.5. Sections 7.6 and 7.7 also deal with latent topic models and their extensions; these try to capture long-range topic information from (spoken) documents, based on VB solutions.

7.1 Variational inference in general

This section starts by describing a general latent model with observation data $\mathbf{X} = \{\mathbf{x}_n | n = 1, \cdots, N\}$, and the set of all variables introduced in our model including latent variables, parameters, hyperparameters, and model structure Z. The latter sections specify Z with more specific variables. The goal of Bayesian inference is to obtain posterior distributions of any variables introduced in the problem, that is:

$$p(Z|\mathbf{X}). \tag{7.1}$$

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

As discussed in Section 2.1.2, once we obtain $p(Z|\mathbf{X})$, we can estimate various information by the MAP or expectation procedure. In VB, we consider an arbitrary posterior distribution $q(Z|\mathbf{X})$. We use the approximated posterior distribution denoted by $q(\cdot)$ to distinguish it from the true posterior distribution $p(\cdot)$. Then the problem is how to obtain a $q(Z|\mathbf{X})$ that is close to $p(Z|\mathbf{X})$, so as to obtain a well-approximated posterior distribution.

7.1.1 Joint posterior distribution

As a measure of evaluating the difference between two distributions, we use the Kullback–Leibler divergence (Kullback & Leibler 1951), as introduced in the ML–EM algorithm (Section 3.4). The Kullback–Leibler divergence between $q(Z|\mathbf{X})$ and $p(Z|\mathbf{X})$ is defined as follows:

$$\mathrm{KL}(q(Z|\mathbf{X})\|p(Z|\mathbf{X})) \triangleq \int q(Z|\mathbf{X}) \log \frac{q(Z|\mathbf{X})}{p(Z|\mathbf{X})} dZ. \tag{7.2}$$

Z can be a set of discrete variables or a set of both continuous and discrete variables, and, strictly speaking, we should use the summation sum_Z for discrete variables and $\int dZ$ for continuous variables in such a case. However, for simplicity we use $\int dZ$ instead of mixing integrals and summations in the following formulation.

The KL divergence (Eq. (7.2)) is represented as

$$\begin{aligned}\mathrm{KL}(q(Z|\mathbf{X})\|p(Z|\mathbf{X})) &= \int q(Z|\mathbf{X}) \log \frac{q(Z|\mathbf{X})}{\frac{p(\mathbf{X},Z)}{p(\mathbf{X})}} dZ \\ &= \log p(\mathbf{X}) - \underbrace{\int q(Z|\mathbf{X}) \log \frac{p(\mathbf{X},Z)}{q(Z|\mathbf{X})} dZ}_{\triangleq \mathcal{F}[q(Z|\mathbf{X})]},\end{aligned} \tag{7.3}$$

where

$$\mathcal{F}[q(Z|\mathbf{X})] \triangleq \int q(Z|\mathbf{X}) \log \frac{p(\mathbf{X},Z)}{q(Z|\mathbf{X})} dZ \tag{7.4}$$

is called the *variational lower bound*. The reason we call it the lower bound is that $\mathcal{F}[q(Z|\mathbf{X})]$ is a lower bound of the evidence (marginal log likelihood) $\log p(\mathbf{X})$ because of the non-negativity of the KL divergence, i.e.,

$$\mathrm{KL}(q(Z|\mathbf{X})\|p(Z|\mathbf{X})) \geq 0 \iff \log p(\mathbf{X}) \geq \mathcal{F}[q(Z|\mathbf{X})]. \tag{7.5}$$

Equation (7.3) means that we can obtain the optimal $q(Z|\mathbf{X})$ by maximizing the variational lower bound $\mathcal{F}[q(Z|\mathbf{X})]$ that corresponds to minimizing the KL divergence since the log evidence $\log p(\mathbf{X})$ does not depend on Z. That is

$$\tilde{q}(Z|\mathbf{X}) = \arg\max_{q(Z|\mathbf{X})} \mathcal{F}[q(Z|\mathbf{X})] \iff \arg\max_{q(Z|\mathbf{X})} \mathrm{KL}(q(Z|\mathbf{X})\|p(Z|\mathbf{X})). \tag{7.6}$$

To obtain the optimal $\tilde{q}(Z|\mathbf{X})$, we use a variational method for a *functional* $\mathcal{F}[q(Z|\mathbf{X})]$, which achieves mapping a function to a real (or complex) value, i.e., $f \mapsto a \in \mathbb{R}$. Thus, the approach is called *variational Bayes*. The variational method is discussed in Section 7.1.3.

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

7.1.2 Factorized posterior distribution

In practical applications, we need to consider the factorized form of the joint distribution $q(Z|\mathbf{X})$ to make the calculation simple. To do this, we consider the jth element of Z and make the following conditional independence assumption:

$$q(Z|\mathbf{X}) = \prod_{j=1}^{J} q(Z_j|\mathbf{X}). \tag{7.7}$$

J is the total number of elements. This is an essential approximation requirement of VB to make the problem practical. Z_j would be a subset of model parameters (e.g., Gaussian mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ in a particular component of GMM/CDHMM), or an HMM state indicator s_t at frame t, for instance. Note that we do not assume the factorization form for the true posterior $p(Z|\mathbf{X})$. Instead, the true posterior $p(Z_i|\mathbf{X})$ can be represented as the following marginalized distribution of $p(Z|\mathbf{X})$ over all Z_j except Z_i:

$$p(Z_i|\mathbf{X}) = \int \cdots \int p(Z|\mathbf{X}) \prod_{j\neq i}^{J} dZ_j \triangleq \int p(Z|\mathbf{X}) dZ_{\backslash i}, \tag{7.8}$$

where $Z_{\backslash i}$ denotes the complementary set of Z_i.

By using Eq. (7.8), the KL divergence between $q(Z_i|\mathbf{X})$ and $p(Z_i|\mathbf{X})$ (not the KL divergence between the joint distributions in Eq. (7.2)) is represented as follows:

$$\begin{aligned}
\mathrm{KL}(q(Z_i|\mathbf{X})\|p(Z_i|\mathbf{X})) &= \int q(Z_i|\mathbf{X}) \log \frac{q(Z_i|\mathbf{X})}{\int p(Z|\mathbf{X}) dZ_{\backslash i}} dZ_i \\
&= \int q(Z_i|\mathbf{X}) \log \frac{q(Z_i|\mathbf{X})}{\int \frac{p(\mathbf{X},Z)}{p(\mathbf{X})} dZ_{\backslash i}} dZ_i \\
&= \log p(\mathbf{X}) - \int q(Z_i|\mathbf{X}) \log \frac{\int p(\mathbf{X},Z) dZ_{\backslash i}}{q(Z_i|\mathbf{X})} dZ_i.
\end{aligned} \tag{7.9}$$

Since Eq. (7.9) has integrals in the logarithmic function, it is difficult to deal with. Therefore, we use the following Jensen's inequality for a concave function f, distribution function $p(Y)$ ($\int p(Y)dY = 1$), and an arbitrary function $g(Y)$ introduced in Section 3.4.1:

$$f\left(\int p(Y)g(Y)dY\right) \geq \int p(Y)f(g(Y))dY. \tag{7.10}$$

In the special case of $f(\cdot) = \log(\cdot)$, $Y = Z_{\backslash i}$, $p(Y) = q(Z_{\backslash i}|\mathbf{X})$, and $g(Y) = \frac{p(\mathbf{X},Z)}{q(Z|\mathbf{X})}$, Eq. (7.10) can be rewritten as follows:

$$\begin{aligned}
&\log\left(\int q(Z_{\backslash i}|\mathbf{X}) \frac{p(\mathbf{X},Z)}{q(Z|\mathbf{X})} dZ_{\backslash i}\right) \\
&\quad = \log\left(\frac{\int p(\mathbf{X},Z) dZ_{\backslash i}}{q(Z_i|\mathbf{X})}\right) \geq \int q(Z_{\backslash i}|\mathbf{X}) \log\left(\frac{p(\mathbf{X},Z)}{q(Z|\mathbf{X})}\right) dZ_{\backslash i},
\end{aligned} \tag{7.11}$$

where we use Eq. (7.7) to cancel $q(Z_{\backslash i}|\mathbf{X})$ in the fraction of the first line. We also use the following relationship, which is true when $p(Y) \geq 0$:

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

$$a(Y) \geq b(Y)\ \forall Y \Rightarrow \int p(Y)a(Y)dY \geq \int p(Y)b(Y)dY. \tag{7.12}$$

In the special case of

$$\begin{aligned} f(\cdot) &= \log(\cdot), \\ Y &= Z_i, \\ p(Y) &= q(Z_i|\mathbf{X}), \\ a(Y) &= \log\left(\frac{\int p(\mathbf{X}, Z)dZ_{\backslash i}}{q(Z_i|\mathbf{X})}\right), \\ b(Y) &= \int q(Z_{\backslash i}|\mathbf{X}) \log\left(\frac{p(\mathbf{X}, Z)}{q(Z|\mathbf{X})}\right) dZ_{\backslash i}, \end{aligned} \tag{7.13}$$

Eq. (7.12) is represented as

$$\begin{aligned} &\int q(Z_i|\mathbf{X}) \log\left(\frac{\int p(\mathbf{X}, Z)dZ_{\backslash i}}{q(Z_i|\mathbf{X})}\right) dZ_i \\ &\geq \int q(Z_i|\mathbf{X}) \int q(Z_{\backslash i}|\mathbf{X}) \log\left(\frac{p(\mathbf{X}, Z)}{q(Z|\mathbf{X})}\right) dZ_{\backslash i}dZ_i. \\ &= \int q(Z|\mathbf{X}) \log\left(\frac{p(\mathbf{X}, Z)}{q(Z|\mathbf{X})}\right) dZ = \mathcal{F}[q(Z|\mathbf{X})], \end{aligned} \tag{7.14}$$

where we use Eq. (7.7) to obtain $q(Z|\mathbf{X})$, and use the definition of the variational lower bound in Eq. (7.4). Therefore, by substituting Eq. (7.14) into the KL divergence Eq. (7.9), Eq. (7.9) is finally represented as follows:

$$\mathrm{KL}(q(Z_i|\mathbf{X})\|p(Z_i|\mathbf{X})) \leq \log p(\mathbf{X}) - \mathcal{F}[q(Z|\mathbf{X})]. \tag{7.15}$$

Compared with Eq. (7.3) that is the equality relationship, Eq. (7.15) is the inequality relationship. Equation (7.15) still has the nice property that the maximization of the variational lower bound corresponds to reducing the KL divergence that then causes the approximated posterior $q(Z_i|\mathbf{X})$ to approach the true posterior $p(Z_i|\mathbf{X})$. That is, if we obtain the following posterior distribution:

$$\tilde{q}(Z_i|\mathbf{X}) = \arg\max_{q(Z_i|\mathbf{X})} \mathcal{F}[q(Z|\mathbf{X})], \tag{7.16}$$

$\tilde{q}(Z_i|\mathbf{X})$ could be a well-approximated posterior distribution in terms of reducing the KL divergence between the true posterior $p(Z_i|\mathbf{X})$ and approximated posterior $\tilde{q}(Z_i|\mathbf{X})$. In this section, a tilde $\tilde{}$ is added to indicate variationally optimized values or functions. However, the maximization of the posterior distribution in terms of the lower bound $\mathcal{F}[q(Z|\mathbf{X})]$ does not directly correspond to minimization of the KL divergence, and we cannot globally optimize $q(Z_i|\mathbf{X})$ in terms of the KL divergence when we use the lower bound as the objective functional. This is a shortcoming of the factorization approximation in Eq. (7.7), but it enables us to obtain the posterior distribution of each variable $q(Z_i|\mathbf{X})$, unlike the joint distribution $q(Z|\mathbf{X})$, which is more practical. The next section discusses how to optimize the approximated posterior by using the variational method.

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

7.1.3 Variational method

The variational method is based on functional differentiation, which is a technique for obtaining an optimal function based on a variational calculation, and is defined as follows:

Continuous function case

$$\frac{\delta}{\delta g(y)}\mathcal{H}[g(x)] = \lim_{\epsilon\to 0}\frac{\mathcal{H}[g(x)+\epsilon\delta(x-y)]-\mathcal{H}[g(x)]}{\varepsilon}, \tag{7.17}$$

where $g(x)$ is a continuous function to be optimized, $\mathcal{H}[g(x)]$ is a functional of $g(x)$ and $\delta(x-y)$ is a Dirac delta function.

Discrete function case

$$\frac{\delta}{\delta g_l}\mathcal{H}[g_n] = \lim_{\epsilon\to 0}\frac{\mathcal{H}[g_n+\epsilon\delta(n,l)]-\mathcal{H}[g_n]}{\varepsilon}. \tag{7.18}$$

Similarly, g_n is a discrete function to be optimized, and $\delta(n,l)$ is a Kronecker delta function. This section aims to obtain the following optimized posterior distribution based on the above variational method:

$$\begin{aligned}\tilde{q}(Z_i|\mathbf{X}) &= \arg\max_{q(Z_i|\mathbf{X})}\mathcal{F}[q(Z|\mathbf{X})]\\ &= \arg\max_{q(Z_i|\mathbf{X})}\int q(Z|\mathbf{X})\log\left(\frac{p(\mathbf{X},Z)}{q(Z|\mathbf{X})}\right)dZ. \end{aligned}\tag{7.19}$$

For simplicity of the calculation, we simplify $q(Z_i|\mathbf{X})$ to $q(Z_i)$ in this section.

If we consider $\int q(Z_i)dZ_i = 1$ constraint, the functional differentiation is represented by substituting $\mathcal{F}[q(Z)]$ and $q(Z_i)$ into $\mathcal{H}$ and $g(y)$ in Eq. (7.17), respectively, as follows:

$$\begin{aligned}&\frac{\delta}{\delta q(Z_i')}\left(\mathcal{F}[q(Z)]+K\left(\int q(Z_i)dZ_i-1\right)\right)\\ &= \lim_{\epsilon\to 0}\frac{1}{\epsilon}\left(\int\left(q(Z_i)+\epsilon\delta(Z_i-Z_i')\right)\mathbb{E}_{(Z_{\backslash i})}\left[\log\frac{p(\mathbf{X},Z)}{\left(q(Z_i)+\epsilon\delta(Z_i-Z_i')\right)q(Z_{\backslash i})}\right]dZ_i\right.\\ &\qquad -\mathcal{F}[q(Z)]+K\left(\int\left(q(Z_i)+\epsilon\delta(Z_i-Z_i')\right)dZ_i-1\right)\\ &\qquad \left.-K\left(\int q(Z_i)dZ_i-1\right)\right),\end{aligned}\tag{7.20}$$

where K is a Lagrange multiplier, as introduced in Section 3.4.3 for the function derivative. We focus on the first term in the brackets in the second line of Eq. (7.20). The first term is rewritten as follows:

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

$$\begin{aligned}
&\int \big(q(Z_i)+\epsilon\delta(Z_i-Z_i')\big)\,\mathbb{E}_{(Z_{\backslash i})}\left[\log\frac{p(\mathbf{X},Z)}{\big(q(Z_i)+\epsilon\delta(Z_i-Z_i')\big)\,q(Z_{\backslash i})}\right]dZ_i\\
&=\int \big(q(Z_i)+\epsilon\delta(Z_i-Z_i')\big)\,\mathbb{E}_{(Z_{\backslash i})}\left[\log\frac{p(\mathbf{X},Z)}{q(Z)+\frac{q(Z)}{q(Z_i)}\epsilon\delta(Z_i-Z_i')}\right]dZ_i\\
&=\int \big(q(Z_i)+\epsilon\delta(Z_i-Z_i')\big)\,\mathbb{E}_{(Z_{\backslash i})}\left[\log\frac{p(\mathbf{X},Z)}{q(Z)}-\log\left(1+\epsilon\frac{\delta(Z_i-Z_i')}{q(Z_i)}\right)\right]dZ_i.
\end{aligned} \tag{7.21}$$

By expanding the logarithmic term in Eq. (7.21) with respect to ϵ, Eq. (7.21) can be represented thus:

Equation (7.21)

$$\begin{aligned}
&=\int \big(q(Z_i)+\epsilon\delta(Z_i-Z_i')\big)\left(\mathbb{E}_{(Z_{\backslash i})}\left[\log\frac{p(\mathbf{X},Z)}{q(Z)}\right]-\epsilon\frac{\delta(Z_i-Z_i')}{q(Z_i)}\right)dZ_i+\mathbf{o}(\epsilon^2)\\
&=\int q(Z_i)\mathbb{E}_{(Z_{\backslash i})}\left[\log\frac{p(\mathbf{X},Z)}{q(Z)}\right]dZ_i-\epsilon\int\delta(Z_i-Z_i')dZ_i\\
&\quad+\epsilon\int\delta(Z_i-Z_i')\mathbb{E}_{(Z_{\backslash i})}\left[\log\frac{p(\mathbf{X},Z)}{q(Z)}\right]dZ_i+\mathbf{o}(\epsilon^2)\\
&=\mathcal{F}[q(Z)]-\epsilon+\epsilon\mathbb{E}_{(Z_{\backslash i})}\left[\log\frac{p(\mathbf{X},Z')}{q(Z')}\right]+\mathbf{o}(\epsilon^2)\\
&=\mathcal{F}[q(Z)]+\epsilon\left(-1+\mathbb{E}_{(Z_{\backslash i})}\left[\log\frac{p(\mathbf{X},Z')}{q(Z')}\right]\right)+\mathbf{o}(\epsilon^2),
\end{aligned} \tag{7.22}$$

where $\mathbf{o}(\epsilon^2)$ denotes a set of terms of no less than the second power of ϵ. $Z' \triangleq \{Z_i', Z_{\backslash i}\}$, but in the following equations, we simply use Z instead of Z', as we do not have to distinguish them. Therefore, by substituting Eq. (7.22) into Eq. (7.20), it can be represented as:

Equation (7.20)

$$\begin{aligned}
&=\lim_{\epsilon\to 0}\frac{1}{\epsilon}\left(\epsilon\left(-1+\mathbb{E}_{(Z_{\backslash i})}\left[\log\frac{p(\mathbf{X},Z)}{q(Z)}\right]+K\right)+\mathbf{o}(\epsilon^2)\right)\\
&=-1+\mathbb{E}_{(Z_{\backslash i})}\left[\log\frac{p(\mathbf{X},Z)}{q(Z)}\right]+K\\
&=-1+\mathbb{E}_{(Z_{\backslash i})}\big[\log p(\mathbf{X},Z)\big]-\mathbb{E}_{(Z_{\backslash i})}\big[\log q(Z)\big]+K\\
&=-1+\mathbb{E}_{(Z_{\backslash i})}\big[\log p(\mathbf{X},Z)\big]-\log q(Z_i)-\mathbb{E}_{(Z_{\backslash i})}\big[\log q(Z_{\backslash i})\big]+K.
\end{aligned} \tag{7.23}$$

We use Eq. (7.7) to factorize $q(Z_i)$ and $q(Z_{\backslash i})$. Therefore, the optimal posterior (VB posterior) $\widetilde{q}(Z_i)$ satisfies the relation whereby Eq. (7.23) $= 0$, and is obtained as:

$$\log\widetilde{q}(Z_i)=-1+\mathbb{E}_{(Z_{\backslash i})}\big[\log p(\mathbf{X},Z)\big]-\mathbb{E}_{(Z_{\backslash i})}\big[\log q(Z_{\backslash i})\big]+K. \tag{7.24}$$

Since only the second term in the right-hand-side depends on Z_i, the optimal VB posterior is finally derived as:

$$\widetilde{q}(Z_i|\mathbf{X})\propto\exp\left(\mathbb{E}_{(Z_{\backslash i}|\mathbf{X})}\big[\log p(\mathbf{X},Z)\big]\right), \tag{7.25}$$

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

or by considering the normalization constant, it is derived as

$$\widetilde{q}(Z_i|\mathbf{X}) = \frac{\exp\left(\mathbb{E}_{(Z_{\setminus i}|\mathbf{X})}\left[\log p(\mathbf{X}, Z)\right]\right)}{\int \exp\left(\mathbb{E}_{(Z_{\setminus i}|\mathbf{X})}\left[\log p(\mathbf{X}, Z)\right]\right) dZ_i}, \tag{7.26}$$

where the omitted notations are recovered ($q(Z_i) \to q(Z_i|\mathbf{X})$). Thus, we obtain the general form of the VB posterior distribution $\widetilde{q}(Z_i|\mathbf{X})$ by using the variational method. Equation (7.25) tells us that if we want to infer some probabilistic variables, we first need to prepare the joint distribution of the observation $\mathbf{X}$ and target variables. Note that $\widetilde{q}(Z_i|\mathbf{X})$ and the other posterior distributions $\widetilde{q}(Z_{\setminus i}|\mathbf{X}) = \prod_{j\neq i}^{J} q(Z_j|\mathbf{X})$ depend on each other due to the expectation in Eq. (7.25). Therefore, this optimization can be performed iteratively from the initial posterior distributions for all $\widetilde{q}(Z_i|\mathbf{X})$. The following sections provide more practical forms of VB posteriors.

7.2 Variational inference for classification problems

This section provides more specific formulations for our speech and language processing issues which focus more on pattern classification problems. Let $\mathbf{O}$ be a training data set of feature vectors, and Z be a set of discrete latent variables. Then, with a fixed model structure M, posterior distributions for model parameters $p(\Theta^{(c)}|\mathbf{O}, M)$ and $p(Z^{(c)}|\mathbf{O}, M)$ given category c are expressed as follows:[1]

$$p(\Theta^{(c)}|\mathbf{O}, M) = \sum_{Z} \int \frac{p(\mathbf{O}, Z|\Theta, M)p(\Theta|M)}{p(\mathbf{O}|M)} d\Theta^{(\setminus c)} \tag{7.27}$$

and

$$p(Z^{(c)}|\mathbf{O}, M) = \sum_{Z^{(\setminus c)}} \int \frac{p(\mathbf{O}, Z|\Theta, M)p(\Theta|M)}{p(\mathbf{O}|M)} d\Theta, \tag{7.28}$$

where $p(\Theta|M)$ is a prior distribution for Θ. Here, $\setminus c$ represents the set of all categories without c. In this section, we can also regard the prior hyperparameter setting as the model structure setting, and include its variations in index M. The posterior distributions for the model structure $p(M|\mathbf{O})$ are expressed as follows:

$$p(M|\mathbf{O}) = \sum_{Z} \int \frac{p(\mathbf{O}, Z|\Theta, M)p(\Theta|M)p(M)}{p(\mathbf{O})} d\Theta, \tag{7.29}$$

where $p(M)$ denotes a prior distribution for model structure M.

These equations cannot be solved analytically, because the acoustic model for speech recognition includes latent variables in HMMs and GMMs, as discussed in Section 3.2, and the total number of model parameters amounts to more than *one million*. In addition, these parameters depend on each other hierarchically. Solving all integrals and expectations numerically requires huge amounts of computation time. Therefore, when applying

[1] It is reasonable to deal with the prior distribution $p(\Theta|M)$ of model parameters given model M instead of $p(\Theta)$, since the actual functional form of model parameters is determined by model M. Conversely, it is very difficult to consider the prior distribution of model parameters $p(\Theta)$ without the model setting.

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

the Bayesian approach to acoustic modeling for speech recognition, an effective approximation technique is necessary. Therefore, this section focuses on the VB approach and derives general solutions for VB posterior distributions $q(\Theta|\mathbf{O},M)$, $q(Z|\mathbf{O},M)$, and $q(M|\mathbf{O})$ to approximate the corresponding true posteriors. To begin with, by following the general VB formulation in Section 7.1.2, we assume that

$$\begin{aligned} q(\Theta, Z|\mathbf{O}, M) &= \prod_c q(\Theta^{(c)}|\mathbf{O}^{(c)}, M) q(Z^{(c)}|\mathbf{O}^{(c)}, M), \\ p(\Theta, Z|\mathbf{O}, M) &= \prod_c p(\Theta^{(c)}|\mathbf{O}^{(c)}, M) p(Z^{(c)}|\mathbf{O}^{(c)}, M). \end{aligned} \tag{7.30}$$

This assumption means that probabilistic variables associated with each category are statistically independent from other categories. In addition, these posterior distributions depend on the model variable M, which is not marginalized. The speech data used are assumed to be well transcribed and the label information is assumed to be reliable. In addition, the frequently used feature extraction (e.g., MFCC) from the speech is good enough for the statistical independence assumption of the observation data to be guaranteed. Therefore, the assumption of class independence is reasonable.

7.2.1 VB posterior distributions for model parameters

This subsection discusses VB posterior distributions for model parameters with fixed model structure M. Initially, arbitrary posterior distribution $q(\Theta^{(c)}|\mathbf{O},M)$ is introduced, and the Kullback–Leibler (KL) divergence (Kullback & Leibler 1951) between $q(\Theta^{(c)}|\mathbf{O},M)$ and true posterior distribution $p(\Theta^{(c)}|\mathbf{O},M)$ is considered:

$$\mathrm{KL}(q(\Theta^{(c)}|\mathbf{O},M)\|p(\Theta^{(c)}|\mathbf{O},M)) = \int q(\Theta^{(c)}|\mathbf{O},M) \log \frac{q(\Theta^{(c)}|\mathbf{O},M)}{p(\Theta^{(c)}|\mathbf{O},M)} d\Theta^{(c)}. \tag{7.31}$$

Substituting Eq. (7.27) into Eq. (7.31), Eq. (7.31) is rewritten as follows:

$$\begin{aligned} &\mathrm{KL}(q(\Theta^{(c)}|\mathbf{O},M)\|p(\Theta^{(c)}|\mathbf{O},M)) \\ &\quad = \int q(\Theta^{(c)}|\mathbf{O},M) \log \frac{q(\Theta^{(c)}|\mathbf{O},M)}{\sum_Z \int \frac{p(\mathbf{O},Z|\Theta,M)p(\Theta|M)}{p(\mathbf{O}|M)} d\Theta^{(\backslash c)}} d\Theta^{(c)} \\ &\quad = \log p(\mathbf{O}|M) - \int q(\Theta^{(c)}|\mathbf{O},M) \\ &\qquad\qquad \times \log \frac{\sum_Z \int p(\mathbf{O},Z|\Theta,M)p(\Theta|M) d\Theta^{(\backslash c)}}{q(\Theta^{(c)}|\mathbf{O},M)} d\Theta^{(c)}. \end{aligned} \tag{7.32}$$

Then applying the continuous Jensen's inequality Eq. (7.10) to Eq. (7.32), the following inequality is obtained:

$$\begin{aligned} &\mathrm{KL}(q(\Theta^{(c)}|\mathbf{O},M)\|p(\Theta^{(c)}|\mathbf{O},M)) \\ &\quad \leq \log p(\mathbf{O}|M) - \sum_Z \int q(\Theta^{(c)}|\mathbf{O},M) q(\Theta^{(\backslash c)}|\mathbf{O},M) q(Z|\mathbf{O},M) \\ &\qquad\qquad \times \log \frac{p(\mathbf{O},Z|\Theta,M)p(\Theta|M)}{q(\Theta^{(c)}|\mathbf{O},M) q(\Theta^{(\backslash c)}|\mathbf{O},M) q(Z)} d\Theta^{(c)} d\Theta^{(\backslash c)} \end{aligned}$$

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

$$= \log p(\mathbf{O}|M) - \sum_{Z} \int q(\Theta|\mathbf{O}, M)q(Z|\mathbf{O}, M) \\ \times \log \frac{p(\mathbf{O}, Z|\Theta, M)p(\Theta|M)}{q(\Theta|\mathbf{O}, M)q(Z|\mathbf{O}, M)p(\mathbf{O}|M)} d\Theta. \tag{7.33}$$

From the third to the fourth line, we use the definition $d\Theta^{(c)}d\Theta^{(\backslash c)} \equiv d\Theta$ and the relation $q(\Theta^{(c)}|\mathbf{O}, M)q(\Theta^{(\backslash c)}|\mathbf{O}, M) = q(\Theta|\mathbf{O}, M)$, which is derived from Eq. (7.30). Thus, we finally obtain the following inequality:

$$\mathrm{KL}(q(\Theta^{(c)}|\mathbf{O}, M)\|p(\Theta^{(c)}|\mathbf{O}, M)) \\ \leq \log p(\mathbf{O}|M) - \mathcal{F}^M[q(\Theta|\mathbf{O}, M), q(Z|\mathbf{O}, M)], \tag{7.34}$$

where

$$\mathcal{F}^M[q(\Theta|\mathbf{O}, M), q(Z|\mathbf{O}, M)] \\ \triangleq \sum_{Z} \int q(\Theta|\mathbf{O}, M)q(Z|\mathbf{O}, M) \log \frac{p(\mathbf{O}, Z|\Theta, M)p(\Theta|M)}{q(\Theta|\mathbf{O}, M)q(Z|\mathbf{O}, M)} d\Theta \\ = \mathbb{E}_{(\Theta, Z)}\left[\log \frac{p(\mathbf{O}, Z|\Theta, M)p(\Theta|M)}{q(\Theta|\mathbf{O}, M)q(Z|\mathbf{O}, M)}\right]. \tag{7.35}$$

This corresponds to the variational lower bound, as discussed in Eq. (7.3). The inequality (7.34) is strict unless $q(\Theta|\mathbf{O}, M) = p(\Theta|\mathbf{O}, M)$ and $q(Z|\mathbf{O}, M) = p(Z|\mathbf{O}, M)$ (i.e., the arbitrary posterior distribution q is equivalent to the true posterior distribution p). From the assumption Eq. (7.30), $\mathcal{F}^M$ is decomposed into each category as follows:

$$\mathcal{F}^M[q(\Theta|\mathbf{O}, M), q(Z|\mathbf{O}, M)] \\ = \sum_{c} \mathbb{E}_{(\Theta^{(c)}, Z^{(c)})}\left[\log \frac{p(\mathbf{O}^{(c)}, Z^{(c)}|\Theta^{(c)}, M)p(\Theta^{(c)}|M)}{q(\Theta^{(c)}|\mathbf{O}^{(c)}, M)q(Z^{(c)}|\mathbf{O}^{(c)}, M)}\right] \\ = \sum_{c} \mathcal{F}^{M,(c)}[q(\Theta^{(c)}|\mathbf{O}^{(c)}, M), q(Z^{(c)}|\mathbf{O}^{(c)}, M)]. \tag{7.36}$$

This indicates that the total objective function is calculated by summing all objective functions for each category.

From inequality Eq. (7.34), $q(\Theta^{(c)}|\mathbf{O}, M)$ approaches $p(\Theta^{(c)}|\mathbf{O}, M)$ as the right-hand-side decreases. Therefore, the optimal posterior distribution can be obtained by a variational method. Since the term $\log p(\mathbf{O}|M)$ can be disregarded, minimization is changed to maximization of $\mathcal{F}^M$ with respect to $q(\Theta^{(c)}|\mathbf{O}, M)$, and is given by the following variational equation:

$$\frac{\delta}{\delta q(\Theta^{(c)}|\mathbf{O}, M)} \mathcal{F}^M[q(\Theta|\mathbf{O}, M), q(Z|\mathbf{O}, M)] \\ = \frac{\delta}{\delta q(\Theta^{(c)}|\mathbf{O}, M)} \mathcal{F}^{M,(c)}[q(\Theta^{(c)}|\mathbf{O}^{(c)}, M), q(Z^{(c)}|\mathbf{O}^{(c)}, M)] = 0. \tag{7.37}$$

From this equation, the optimal VB posterior distribution $\tilde{q}(\Theta^{(c)}|\mathbf{O}, M)$ is obtained by using the variational method as follows:

$$\tilde{q}(\Theta^{(c)}|\mathbf{O}, M) \propto p(\Theta^{(c)}|M) \exp\left(\mathbb{E}_{(Z^{(c)})}\left[\log p(\mathbf{O}^{(c)}, Z^{(c)}|\Theta^{(c)}, M)\right]\right). \tag{7.38}$$

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

This result can also be obtained by using the general formula of the variational posterior in Eq. (7.25). By using the replacement $Z_i \to \Theta^{(c)}$, Eq. (7.25) can be rewritten as follows:

$$\begin{aligned}
\tilde{q}(\Theta^{(c)}|\mathbf{O}^{(c)}, M) &\propto \exp\left(\mathbb{E}_{(\Theta^{(\backslash c)},Z)}\left[\log p(\mathbf{O}, \Theta, Z|M)\right]\right) \\
&= \exp\left(\mathbb{E}_{(\Theta^{(\backslash c)},Z)}\left[\log p(\mathbf{O}, Z|\Theta, M) p(\Theta|M)\right]\right) \\
&= \exp\left(\sum_{c' \neq c} \mathbb{E}_{(\Theta^{(c')},Z^{(c')})}\left[\log p(\mathbf{O}^{(c')}, Z^{(c')}|\Theta^{(c')}, M) p(\Theta^{(c')}|M)\right]\right) \\
&\quad \times \exp\left(\mathbb{E}_{(Z^{(c)})}\left[\log p(\mathbf{O}^{(c)}, Z^{(c)}|\Theta^{(c)}, M) p(\Theta^{(c)}|M)\right]\right) \\
&\propto p(\Theta^{(c)}|M) \exp\left(\mathbb{E}_{(Z^{(c)})}\left[\log p(\mathbf{O}^{(c)}, Z^{(c)}|\Theta^{(c)}, M)\right]\right).
\end{aligned} \tag{7.39}$$

Here, we use the factorization property of the posterior distributions in Eq. (7.30). This result means that the optimal posterior distribution of model parameters $\tilde{q}(\Theta^{(c)}|\mathbf{O}^{(c)}, M)$ is obtained by its prior distribution $p(\Theta^{(c)}|M)$ and the expected complete data likelihood $p(\mathbf{O}^{(c)}, Z^{(c)}|\Theta^{(c)}, M)$.

7.2.2 VB posterior distributions for latent variables

A similar method is used for the optimal VB posterior distribution $\tilde{q}(Z^{(c)}|\mathbf{O}, M)$. An inequality similar to Eq. (7.37) is obtained by considering the KL divergence between the arbitrary posterior distribution $q(Z^{(c)}|\mathbf{O}, M)$ and the true posterior distribution $p(Z^{(c)}|\mathbf{O}, M)$ as follows:

$$\begin{aligned}
&\mathrm{KL}(q(Z^{(c)}|\mathbf{O}, M) \| p(Z^{(c)}|\mathbf{O}, M)) \\
&\quad \leq \log p(\mathbf{O}|M) - \mathcal{F}^M[q(\Theta|\mathbf{O}, M), q(Z|\mathbf{O}, M)].
\end{aligned} \tag{7.40}$$

The optimal VB posterior distribution $\tilde{q}(Z^{(c)}|\mathbf{O}, M)$ is also obtained by maximizing $\mathcal{F}^M$ with respect to $q(Z^{(c)}|\mathbf{O}, M)$ with the variational method as follows:

$$\tilde{q}(Z^{(c)}|\mathbf{O}, M) \propto \exp\left(\mathbb{E}_{(\Theta^{(c)})}\left[\log p(\mathbf{O}^{(c)}, Z^{(c)}|\Theta^{(c)}, M)\right]\right). \tag{7.41}$$

This result is also obtained by using the general formula of the variational posterior in Eq. (7.25). Compared with the result for $\tilde{q}(\Theta^{(c)}|\mathbf{O}^{(c)}, M)$ in Eq. (7.38), Eq. (7.41) does not need to prepare the prior distribution for Z.

7.2.3 VB–EM algorithm

Equations (7.38) and (7.41) are closed-form expressions, and these optimizations can be effectively performed by iterative calculations analogous to the expectation and maximization (EM) algorithm (Dempster *et al*. 1976), as discussed in Sections 3.4 and 4.2, which increases $\mathcal{F}^M$ at every iteration up to a converged value. Then, Eqs. (7.38) and (7.41), respectively, correspond to the maximization step (M-step) and the expectation

Table 7.1 Training specifications for ML and VB.

	Training	Min-max optimization	Objective function
ML	ML–EM	differential method	Q function
VB	VB–EM	variational method	$\mathcal{F}^M$ functional

step (E-step) in the VB approach. We call this algorithm the variational Bayes expectation and maximization (VB–EM) algorithm. Therefore, by substituting q into $\tilde{q}$, these equations can be represented as follows:

$$\begin{cases} \tilde{q}(\Theta^{(c)}|\mathbf{O}, M) \propto p(\Theta^{(c)}|M) \exp\left(\mathbb{E}_{(Z^{(c)})}\left[\log p(\mathbf{O}^{(c)}, Z^{(c)}|\Theta^{(c)}, M)\right]\right), \\ \tilde{q}(Z^{(c)}|\mathbf{O}, M) \propto \exp\left(\mathbb{E}_{(\Theta^{(c)})}\left[\log p(\mathbf{O}^{(c)}, Z^{(c)}|\Theta^{(c)}, M)\right]\right). \end{cases} \tag{7.42}$$

Note that optimal posterior distributions for a particular category can be obtained simply by using the category's variables, i.e., we are not concerned with the other categories in the calculation, since Eq. (7.42) only depends on category c, which is based on the assumption given by Eq. (7.30).

Finally, to compare the VB approach with the conventional ML approach for training latent variable models, the training specifications for ML and VB are summarized in Table 7.1.

7.2.4 VB posterior distribution for model structure

The VB posterior distributions for a model structure are derived in the same way as in Section 7.2.1, and model selection is carried out employing the posterior distribution. Arbitrary posterior distribution $q(M|\mathbf{O})$ is introduced and the KL divergence between $q(M|\mathbf{O})$ and the true posterior distribution $p(M|\mathbf{O})$ is considered:

$$\mathrm{KL}(q(M|\mathbf{O})\|p(M|\mathbf{O})) = \sum_M q(M|\mathbf{O}) \log \frac{q(M|\mathbf{O})}{p(M|\mathbf{O})}. \tag{7.43}$$

Substituting Eq. (7.29) into Eq. (7.43) and using Jensen's inequality, the inequality of Eq. (7.43) can be obtained as follows:

$$\begin{aligned} &\mathrm{KL}\left(q(M|\mathbf{O})\|p(M|\mathbf{O})\right) \\ &\quad \leq \log p(\mathbf{O}) + \mathbb{E}_{(M)}\left[\log \frac{q(M|\mathbf{O})}{p(M)} - \mathcal{F}^M[q(\Theta|\mathbf{O}, M), q(Z|\mathbf{O}, M)]\right]. \end{aligned} \tag{7.44}$$

Similarly to the discussion in Section 7.2.1, from the inequality Eq. (7.44), $q(M|\mathbf{O})$ approaches $p(M|\mathbf{O})$ as the right-hand-side decreases.

Compared with the posterior distributions of model parameters and latent variables, we cannot use the formula Eq. (7.25), since it is not practical to marginalize all possible model structures M. Therefore, the optimal posterior distribution for a model structure can again be obtained by a variational method, as explained in Section 7.1.3. If we consider the constraint $\sum_M q(M|\mathbf{O}) = 1$, the functional differentiation is represented by substituting respectively $\mathcal{F}^M$ and $q(M|\mathbf{O})$ into $\mathcal{H}$ and g_n in Eq. (7.18) as follows:

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

$$
\begin{aligned}
&\frac{\delta}{\delta q(M'|\mathbf{O})}\left(\mathbb{E}_{(M)}\left[\log\frac{q(M|\mathbf{O})}{p(M)}-\mathcal{F}^M\right]+K\left(\sum_M q(M|\mathbf{O})-1\right)\right)\\
&=\lim_{\epsilon\to 0}\frac{1}{\epsilon}\left(\sum_M (q(M|\mathbf{O})+\epsilon\delta_{MM'})\left(\log\frac{q(M|\mathbf{O})+\epsilon\delta_{MM'}}{p(M)}-\mathcal{F}^M\right)\right.\\
&\qquad-\mathbb{E}_{(M)}\left[\log\frac{q(M|\mathbf{O})}{p(M)}-\mathcal{F}^M\right]\\
&\qquad\left.+K\left(\sum_M q(M|\mathbf{O})+\epsilon\delta_{MM'}-1\right)-K\left(\sum_M q(M|\mathbf{O})-1\right)\right),
\end{aligned}
\tag{7.45}
$$

where K is a Lagrange multiplier. We focus on the first term in the brackets in the 2nd line of Eq. (7.45). This term can be rewritten as follows:

$$
\begin{aligned}
&\sum_M (q(M|\mathbf{O})+\epsilon\delta_{MM'})\left(\log\frac{q(M|\mathbf{O})+\epsilon\delta_{MM'}}{p(M)}-\mathcal{F}^M\right)\\
&=\sum_M (q(M|\mathbf{O})+\epsilon\delta_{MM'})\left(\log\frac{q(M|\mathbf{O})}{p(M)}+\log\left(1+\epsilon\frac{\delta_{MM'}}{q(M|\mathbf{O})}\right)-\mathcal{F}^M\right).
\end{aligned}
\tag{7.46}
$$

By expanding the logarithmic term in Eq. (7.46) with respect to ϵ, Eq. (7.46) is represented as:

$$
\begin{aligned}
&\text{Equation (7.46)}\\
&=\mathbb{E}_{(M)}\left[\log\frac{q(M|\mathbf{O})}{p(M)}-\mathcal{F}^M\right]+\epsilon\left(\log\frac{q(M'|\mathbf{O})}{p(M')}-\mathcal{F}^{M'}+1\right)+\mathbf{o}(\epsilon^2).
\end{aligned}
\tag{7.47}
$$

Therefore, by substituting Eq. (7.47) into Eq. (7.45), Eq. (7.45) is represented as:

$$
\text{Equation (7.45)}=\log\frac{q(M'|\mathbf{O})}{p(M')}-\mathcal{F}^{M'}+1+K.
\tag{7.48}
$$

Therefore, the optimal posterior (VB posterior) $\tilde{q}(M|\mathbf{O})$ satisfies the relation whereby Eq. (7.48) $= 0$, and is obtained as:

$$
\log\frac{\tilde{q}(M|\mathbf{O})}{p(M)}-\mathcal{F}^M+1+K=0.
\tag{7.49}
$$

By disregarding the normalization constant, the optimal VB posterior is finally derived as:

$$
\tilde{q}(M|\mathbf{O})\propto p(M)\exp\left(\mathcal{F}^M[q(\Theta|\mathbf{O},M),q(Z|\mathbf{O},M)]\right).
\tag{7.50}
$$

Compared with Eqs. (7.38) and (7.41), the posterior obtained is represented by the total variational lower bound.

Assuming that $p(M)$ is a uniform distribution,[2] the proportional relation between $\tilde{q}(M|\mathbf{O})$ and $\mathcal{F}^M$ is obtained as follows, based on the convexity of the logarithmic function:

$$
\mathcal{F}^{M'}\geq\mathcal{F}^M\Leftrightarrow\tilde{q}(M'|\mathbf{O})\geq\tilde{q}(M|\mathbf{O}).
\tag{7.51}
$$

[2] We can set an informative prior distribution for $p(M)$ instead of the uniform distribution. Several prior distributions for model structure are considered in Chapter 8.

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

Therefore, an optimal model structure in the sense of maximum posterior probability estimation can be selected as follows:

$$\widetilde{M} = \arg\max_{M} \widetilde{q}(M|\mathbf{O}) = \arg\max_{M} \mathcal{F}^{M}. \tag{7.52}$$

This indicates that by maximizing total $\mathcal{F}^M$ with respect to both $q(\Theta|\mathbf{O}, M)$, $q(Z|\mathbf{O}, M)$, and M, we can obtain the optimal parameter distributions and select the optimal model structure simultaneously (Attias 1999, Ueda & Ghahramani 2002).

Thus, we analytically derive the variational posterior distributions of general latent models. The next section applies these solutions to the continuous density hidden Markov model (CDHMM), as we apply ML–EM and MAP–EM to CDHMMs in Sections 3.4 and 4.3, respectively.

7.3 Continuous density hidden Markov model

This section reformulates the CDHMM training for speech processing based on the VB framework (Valente & Wellekens 2003, Somervuo 2004, Watanabe *et al*. 2004). The four formulations are obtained by using the VB framework to perform acoustic model construction (model setting, training, and selection) and speech classification consistently, based on the Bayesian approach. Consequently, the conventional formulations based on the ML and MAP approaches in Sections 3.4 and 4.3 are replaced by formulations based on the Bayesian approach as follows:

- Set generative model distributions
  → *Set generative model distributions and prior distributions* (Section 7.3.1 and 7.3.2);

- ML/MAP Baum–Welch algorithm
  → *VB Baum–Welch algorithm* (Section 7.3.3);

- Log likelihood
  → *VB objective function* (Section 7.3.4);

- ML/MAP-based classification
  → *VB–BPC* (Section 7.3.5).

These four formulations are explained in the following four subsections, by applying the acoustic model for speech recognition to the general solution in Section 7.2.

7.3.1 Generative model

Similarly to the MAP–EM algorithm in Section 4.3, setting of the emission and prior distributions is required when calculating the VB posterior distributions. This section provides these distributions for CDHMM again, to provide the VB-based analytical solutions.

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

Let $\mathbf{O} = \{\mathbf{o}_t \in \mathbb{R}^D | t = 1, \ldots, T\}$ be a sequential speech data set for a speech segment of a phoneme category. Since the formulations for the posterior distributions are common to all phoneme categories, the phoneme category index c is omitted from this section to simplify the equation forms. D is used to denote the dimension number of the feature vector and T to denote the frame number. The complete data likelihood function with speech, HMM state, and GMM component sequences ($\mathbf{O}$, S, and V), which is introduced in Eqs. (3.44) and (4.23), is expressed by

$$p(\mathbf{O}, S, V|\Theta, M) = \prod_{t=1}^{T} a_{s_{t-1}s_t}\omega_{s_t v_t} p(\mathbf{o}_t|\Theta_{s_t v_t}, M), \tag{7.53}$$

where $a_{s_0 s_1} = \pi_{s_1}$.[3] Although we have many segments for each phoneme category and the generative model distribution must consider the product of each segment in Eq. (7.53), this is also omitted in this book. Here, S and V are sets of discrete latent variables, which are the concrete forms of Z in Section 7.2. The parameter a_{ij} denotes the state transition probability from state i to state j, and ω_{jk} is the kth weight factor of the Gaussian mixture for state j. In addition, $p(\mathbf{o}_t|\Theta_{jk})(= \mathcal{N}(\mathbf{o}_t|\boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk}))$ denotes the Gaussian with mean vector $\boldsymbol{\mu}_{jk}$ and covariance matrix $\boldsymbol{\Sigma}_{jk}$ defined as:

$$\mathcal{N}(\mathbf{o}_t|\boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk}) \triangleq (2\pi)^{-\frac{D}{2}} |\boldsymbol{\Sigma}_{jk}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{o}_t - \boldsymbol{\mu}_{jk})^{\mathsf{T}} \boldsymbol{\Sigma}_{jk}^{-1}(\mathbf{o}_t - \boldsymbol{\mu}_{jk})\right). \tag{7.54}$$

$\Theta = \{a_{ij}, \omega_{jk}, \boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk}^{-1} | i, j = 1, \ldots, J, k = 1, \ldots, K\}$ is a set of model parameters. Here, J denotes the number of states in an HMM sequence and K denotes the number of Gaussian components in a state. This section only considers the diagonal covariance matrix case.

7.3.2 Prior distribution

Conjugate distributions, which are based on the exponential function, are as easy to use as prior distributions since the function forms of prior and posterior distributions become the same (Berger 1985, Gauvain & Lee 1994, Bernardo & Smith 2009), as discussed in Sections 2.1.3 and 4.3.3. Then a distribution is selected where the probabilistic variable constraint is the same as that of the model parameter. The state transition probability a_{ij} and the mixture weight factor ω_{jk} have the constraint that $\sum_j a_{ij} = 1$ and $\sum_k \omega_{jk} = 1$. Therefore, the Dirichlet distributions for π_j, a_{ij}, and ω_{jk} are used, where the variables of the Dirichlet distribution satisfy the above constraint. Similarly, the diagonal elements of the inverse covariance matrix $\boldsymbol{\Sigma}_{jk}^{-1}$ are always positive, and the gamma distribution is used. The range of the mean vector $\boldsymbol{\mu}_{jk}$ is from $-\infty$ to ∞, and the multivariate Gaussian distribution is used. Thus, as introduced in Eqs. (4.29) and (4.32), the prior distribution for a CDHMM with a diagonal covariance matrix is expressed as follows:

[3] This section does not explicitly provide the posterior solution of the initial weight, as it is trivial.

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

$$
\begin{aligned}
p(\Theta|M) &\triangleq \prod_{i=1}^{J} p(\{a_{ij'}\}_{j'=1}^{J}|M) \prod_{j=1}^{J} p(\{\omega_{jk'}\}_{k'=1}^{K}|M) \prod_{k=1}^{K} p(\boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk}|M) \\
&\triangleq \prod_{i=1}^{J} \mathrm{Dir}(\{a_{ij'}\}_{j'=1}^{J}|\{\phi_{ij'}^{a}\}_{j'=1}^{J}) \prod_{j=1}^{J} \mathrm{Dir}(\{\omega_{jk'}\}_{k'=1}^{K}|\{\phi_{jk'}^{\omega}\}_{k'=1}^{K}) \\
&\quad \times \prod_{k=1}^{K} \prod_{d=1}^{D} \mathcal{N}(\mu_{jkd}|\mu_{jkd}^{0}, (\phi_{jk}^{\mu} r_{jkd})^{-1}) \mathrm{Gam}_2(r_{jkd}|\phi_{jk}^{r}, r_{jkd}^{0}).
\end{aligned}
\tag{7.55}
$$

Here, $\Phi^0 \triangleq \{\phi_{ij}^{a}, \phi_{jk}^{\omega}, \phi_{jk}^{\mu}, \mu_{jkd}^{0}, \phi_{jk}^{r}, r_{jkd}^{0} | i, j = 1, \ldots, J, k = 1, \ldots, K, d = 1, \cdots, D\}$ is a set of prior parameters. In Eq. (7.55), $\mathrm{Dir}(\cdot)$ denotes a Dirichlet distribution and $\mathrm{Gam}_2(\cdot)$ denotes a gamma distribution. (It is different from the conventional definition of the gamma distribution, see Appendix C.11.) If the covariance matrix elements are off the diagonal, a Gaussian–Wishart distribution is used as the prior distribution of $\boldsymbol{\mu}_{jk}$ and $\boldsymbol{\Sigma}_{jk}$. The explicit forms of the distributions are defined as follows (Appendixes C.4, C.5, and C.11):

$$
\begin{cases}
\mathrm{Dir}(\{a_{ij}\}_{j=1}^{J}|\{\phi_{ij}^{a}\}_{j=1}^{J}) & \triangleq C_{\mathrm{Dir}}(\{\phi_{ij}^{a}\}_{j=1}^{J}) \prod_{j=1}^{J} (a_{ij})^{\phi_{ij}^{a}-1}, \\
\mathrm{Dir}(\{\omega_{jk}\}_{k=1}^{K}|\{\phi_{jk}^{\omega}\}_{k=1}^{K}) & \triangleq C_{\mathrm{Dir}}(\{\phi_{jk}^{\omega}\}_{k=1}^{K}) \prod_{k=1}^{K} (\omega_{jk})^{\phi_{jk}^{\omega}-1}, \\
\mathcal{N}(\mu_{jkd}|\mu_{jkd}^{0}, (\phi_{jk}^{\mu} r_{jkd})^{-1}) & \triangleq C_{\mathcal{N}}(\phi_{jk}^{\mu})(r_{jkd})^{\frac{1}{2}} \exp\left(-\frac{\phi_{jk}^{\mu} r_{jkd}(\mu_{jkd}-\mu_{jkd}^{0})^2}{2}\right), \\
\mathrm{Gam}_2(r_{jkd}|\phi_{jk}^{r}, r_{jkd}^{0}) & \triangleq C_{\mathrm{Gam}_2}(\phi_{jk}^{r}, r_{jkd}^{0})(r_{jkd})^{\frac{\phi_{jk}^{r}}{2}-1} \exp\left(-\frac{r_{jkd}^{0} r_{jkd}}{2}\right),
\end{cases}
\tag{7.56}
$$

where the normalization constants are defined as follows:

$$
\begin{cases}
C_{\mathrm{Dir}}(\{\phi_{ij}^{a}\}_{j=1}^{J}) & \triangleq \dfrac{\Gamma(\sum_{j=1}^{J} \phi_{ij}^{a})}{\prod_{j=1}^{J} \Gamma(\phi_{ij}^{a})}, \\
C_{\mathrm{Dir}}(\{\phi_{jk}^{\omega}\}_{k=1}^{K}) & \triangleq \dfrac{\Gamma(\sum_{k=1}^{K} \phi_{jk}^{\omega})}{\prod_{k=1}^{K} \Gamma(\phi_{jk}^{\omega})}, \\
C_{\mathcal{N}}(\phi_{jk}^{\mu}) & \triangleq \left(\dfrac{\phi_{jk}^{\mu}}{2\pi}\right)^{\frac{1}{2}}, \\
C_{\mathrm{Gam}_2}(\phi_{jk}^{r}, r_{jkd}^{0}) & \triangleq \dfrac{\left(\frac{r_{jkd}^{0}}{2} 2\right)^{\frac{\phi_{jk}^{r}}{2}}}{\Gamma\left(\frac{\phi_{jk}^{r}}{2}\right)}.
\end{cases}
\tag{7.57}
$$

In the Bayesian approach, an important problem is how to set the prior parameters. Here, two kinds of prior parameters of μ^0 and r^0 are set using sufficient amounts of data from:

- Statistics of higher hierarchy acoustic models for the acoustic model construction task;
- Statistics of speaker independent models for the speaker adaptation task.

The other parameters ($\phi^a, \phi^\omega, \phi^\mu$, and ϕ^r) have a meaning as regarding tuning the balance between the values obtained from training data and the above statistics. These parameters are set appropriately based on experiments, as discussed in speaker adaptation (Section 4.4).

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

Finally, Algorithm 10 provides a generative process for a CDHMM with prior distribution. For simplicity, the initial weight, the hyperparameters, and the model structure are given in this generative process, but it can also be sampled from some distributions. Compared with Algorithm 3, CDHMM parameters are also sampled from the prior distributions.

Algorithm 10 Generative process for continuous density hidden Markov model with prior distributions

```
Require: $\Psi$, $M$, and $\{\pi_j\}_{j=1}^J$
for $i,j = 1, \cdots, J$ do
    Draw $a_{ij}$ from Dir($\{a_{ij}\}_{j=1}^J | \{\phi^a_{ij}\}_{j=1}^J$)
end for
for $j = 1, \cdots, J$ do
    for $k = 1, \cdots, K$ do
        Draw $\omega_{jk}$ from Dir($\{\omega_{jk}\}_{k=1}^K | \{\phi^\omega_{jk}\}_{k=1}^K$)
        for $d = 1, \cdots, D$ do
            Draw $r_{jkd}$ from Gam$_2$($r_{jkd} | \phi^r_{jk}, r^0_{jkd}$)
            Draw $\mu_{jkd}$ from $\mathcal{N}(\mu_{jkd} | \mu^0_{jkd}, (\phi^\mu_{jk} r_{jkd})^{-1})$
        end for
    end for
end for
Draw $s_1$ from Mult($s_1 | \{\pi_j\}_{j=1}^J$)
Draw $v_1$ from Mult($v_1 | \{\omega_{s_1 k}\}_{k=1}^K$)
Draw $\mathbf{o}_1$ from $\mathcal{N}(\mathbf{o}_1 | \boldsymbol{\mu}_{s_1 v_1}, \boldsymbol{\Sigma}_{s_1 v_1})$
for $t = 2, \cdots, T$ do
    Draw $s_t$ from Mult($s_t | \{a_{s_{t-1} j}\}_{j=1}^J$)
    Draw $v_t$ from Mult($v_t | \{\omega_{s_t k}\}_{k=1}^K$)
    Draw $\mathbf{o}_t$ from $\mathcal{N}(\mathbf{o}_t | \boldsymbol{\mu}_{s_t v_t}, \boldsymbol{\Sigma}_{s_t v_t})$
end for
```

7.3.3 VB Baum–Welch algorithm

This subsection introduces concrete forms of the VB posterior distributions for model parameters $q(\Theta|\mathbf{O}, M)$ and for latent variables $q(Z|\mathbf{O}, M)$ in acoustic modeling, which are efficiently computed by VB iterative calculations within the VB framework. This calculation is effectively carried out by the VB Baum–Welch algorithm (MacKay 1997).

VB M-step

First, the VB M-step for acoustic model training is explained. This is solved by substituting the acoustic model setting in Section 7.3.1 into the general solution for the VB M-step in Section 7.2. From Eq. (7.42), the VB posterior distributions for the model parameters are represented as follows:

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

$$\widetilde{q}(\Theta|\mathbf{O}, M) \propto p(\Theta|M) \exp\left(\mathbb{E}_{(S,V)}\left[\log p(\mathbf{O}, S, V|\Theta, M)\right]\right). \tag{7.58}$$

Taking the logarithmic operation, Eq. (7.58) is represented as:

$$\begin{aligned}\log \widetilde{q}(\Theta|\mathbf{O}, M) &\propto \log p(\Theta|M) + \mathbb{E}_{(S,V)}\left[\log p(\mathbf{O}, S, V|\Theta, M)\right] \\ &\triangleq \tilde{Q}(\Theta).\end{aligned} \tag{7.59}$$

Here, $\tilde{Q}(\Theta)$ is a VB auxiliary function defined as follows:

$$\begin{aligned}\tilde{Q}(\Theta) &\triangleq \mathbb{E}_{(S,V)}\left[\log p(\mathbf{O}, S, V|\Theta, M)\right] + \log p(\Theta|M) \\ &= \sum_{S,V} \widetilde{q}(S, V|\mathbf{O}, M) \log p(\mathbf{O}, S, V|\Theta, M) + \log p(\Theta|M).\end{aligned} \tag{7.60}$$

On the other hand, the MAP auxiliary function in Eq. (4.16) is defined as follows:

$$Q^{\mathrm{MAP}}(\Theta'|\Theta) \triangleq \sum_{S,V} p(S, V|\mathbf{O}, \Theta, M) \log p(\mathbf{O}, S, V|\Theta', M) + \log p(\Theta'|M). \tag{7.61}$$

Comparing the VB and MAP auxiliary functions, these are almost equivalent except the posterior distributions of latent variables, i.e., $\widetilde{q}(S, V|\mathbf{O}, M)$ vs. $p(S, V|\mathbf{O}, \Theta, M)$. Since $\widetilde{q}(S, V|\mathbf{O}, M)$ is obtained by marginalizing Θ, $\tilde{Q}(\Theta)$ is more appropriate in terms of the Bayesian treatment.

Therefore, this VB-M step is solved by using the result of the MAP-M step solution, except that $\widetilde{q}(S, V|\mathbf{O}, M)$ is obtained by using the VB-E step. The calculated results for the optimal VB posterior distributions for the model parameters are summarized as follows:

$$\begin{aligned}\widetilde{q}(\Theta|M) &\triangleq \prod_{i=1}^{J} \widetilde{q}(\{a_{ij'}\}_{j'=1}^{J}|M) \prod_{j=1}^{J} \widetilde{q}(\{\omega_{jk'}\}_{k'=1}^{K}|M) \prod_{k=1}^{K} \widetilde{q}(\boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk}|M) \\ &\triangleq \prod_{i=1}^{J} \mathrm{Dir}(\{a_{ij'}\}_{j'=1}^{J}|\{\widetilde{\phi}_{ij'}^{a}\}_{j'=1}^{J}) \prod_{j=1}^{J} \mathrm{Dir}(\{\omega_{jk'}\}_{k'=1}^{K}|\{\widetilde{\phi}_{jk'}^{\omega}\}_{k'=1}^{K}) \\ &\quad \times \prod_{k=1}^{K} \prod_{d=1}^{D} \mathcal{N}(\mu_{jkd}|\widetilde{\mu}_{jkd}, (\widetilde{\phi}_{jk}^{\mu} r_{jkd})^{-1}) \mathrm{Gam}_2(r_{jkd}|\widetilde{\phi}_{jk}^{r}, \widetilde{r}_{jkd}).\end{aligned} \tag{7.62}$$

The concrete forms of the distributions are defined as follows:

$$\begin{cases}\mathrm{Dir}(\{a_{ij}\}_{j=1}^{J}|\{\widetilde{\phi}_{ij}^{a}\}_{j=1}^{J}) & \triangleq C_{\mathrm{Dir}}(\{\widetilde{\phi}_{ij}^{a}\}_{j=1}^{J}) \prod_{j=1}^{J} (a_{ij})^{\widetilde{\phi}_{ij}^{a}-1}, \\ \mathrm{Dir}(\{\omega_{jk}\}_{k=1}^{K}|\{\widetilde{\phi}_{jk}^{\omega}\}_{k=1}^{K}) & \triangleq C_{\mathrm{Dir}}(\{\widetilde{\phi}_{jk}^{\omega}\}_{k=1}^{K}) \prod_{k=1}^{K} (\omega_{jk})^{\widetilde{\phi}_{jk}^{\omega}-1}, \\ \mathcal{N}(\mu_{jkd}|\widetilde{\mu}_{jkd}, (\widetilde{\phi}_{jk}^{\mu} r_{jkd})^{-1}) & \triangleq C_{\mathcal{N}}(\widetilde{\phi}_{jk}^{\mu})(r_{jkd})^{\frac{1}{2}} \exp\left(-\frac{\widetilde{\phi}_{jk}^{\mu} r_{jkd}(\mu_{jkd}-\widetilde{\mu}_{jkd})^2}{2}\right), \\ \mathrm{Gam}_2(r_{jkd}|\widetilde{\phi}_{jk}^{r}, \widetilde{r}_{jkd}) & \triangleq C_{\mathrm{Gam}_2}(\widetilde{\phi}_{jk}^{r}, \widetilde{r}_{jkd})(r_{jkd})^{\frac{\widetilde{\phi}_{jk}^{r}}{2}-1} \exp\left(-\frac{\widetilde{r}_{jkd} r_{jkd}}{2}\right),\end{cases} \tag{7.63}$$

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

where the normalization constants are:

$$\begin{cases} C_{\mathrm{Dir}}(\{\widetilde{\phi}^a_{ij}\}_{j=1}^J) & \triangleq \dfrac{\Gamma(\sum_{j=1}^J \widetilde{\phi}^a_{ij})}{\prod_{j=1}^J \Gamma(\widetilde{\phi}^a_{ij})}, \\ C_{\mathrm{Dir}}(\{\widetilde{\phi}^\omega_{jk}\}_{k=1}^K) & \triangleq \dfrac{\Gamma(\sum_{k=1}^K \widetilde{\phi}^\omega_{jk})}{\prod_{k=1}^K \Gamma(\widetilde{\phi}^\omega_{jk})}, \\ C_{\mathcal{N}}(\widetilde{\phi}^\mu_{jk}) & \triangleq \left(\dfrac{\widetilde{\phi}^\mu_{jk}}{2\pi}\right)^{\frac{1}{2}}, \\ C_{\mathrm{Gam}_2}(\widetilde{\phi}^r_{jk}, \widetilde{r}_{jkd}) & \triangleq \dfrac{\left(\frac{\widetilde{r}_{jkd}}{2}\right)^{\frac{\widetilde{\phi}^r_{jk}}{2}}}{\Gamma\left(\frac{\widetilde{\phi}^r_{jk}}{2}\right)}. \end{cases} \tag{7.64}$$

Note that Eqs. (7.55) and (7.62) are members of the same function family, and the only difference is that the set of prior parameters Φ^0 in Eq. (7.55) is replaced with a set of posterior distribution parameters $\widetilde{\Phi}$ in Eq. (7.62), where $\widetilde{\Phi}$ is defined as:

$$\begin{aligned} \widetilde{\Phi} \triangleq & \{\widetilde{\phi}^a_{ij}, \widetilde{\phi}^\omega_{jk}, \widetilde{\phi}^\mu_{jk}, \widetilde{\mu}_{jkd}, \widetilde{\phi}^r_{jk}, \widetilde{r}_{jkd} \\ & |i,j = 1, \dots, J, k = 1, \dots, K, d = 1, \cdots, D\}. \end{aligned} \tag{7.65}$$

The conjugate prior distribution is adopted because the posterior distribution is theoretically a member of the same function family as the prior distribution, and is obtained analytically, which is a characteristic of the exponential distribution family, as discussed in Section 2.1.4. Here, $\widetilde{\Phi}$ values are calculated from:

$$\begin{aligned} \widetilde{\phi}^a_{ij} &= \phi^a_{ij} + \widetilde{\xi}_{ij}, \\ \widetilde{\phi}^\omega_{jk} &= \phi^\omega_{jk} + \widetilde{\gamma}_{jk}, \\ \widetilde{\phi}^\mu_{jk} &= \phi^\mu_{jk} + \widetilde{\gamma}_{jk}, \\ \widetilde{\mu}_{jkd} &= \frac{\phi^\mu_{jk}\mu^0_{jkd} + \gamma^{(1)}_{jkd}}{\phi^\mu_{jk} + \widetilde{\gamma}_{jk}}, \\ \widetilde{\phi}^r_{jk} &= \phi^r_{jk} + \widetilde{\gamma}_{jk}, \\ \widetilde{r}_{jkd} &= \widetilde{\gamma}^{(2)}_{jkd} + \phi^\mu_{jk}(\mu^0_{jkd})^2 - \widetilde{\phi}^\mu_{jk}(\widetilde{\mu}_{jkd})^2 + r^0_{jkd}. \end{aligned} \tag{7.66}$$

$\widetilde{\xi}_{ij}$ denotes the sufficient statistics of the transition matrix, and $\widetilde{\gamma}_{jk}$, $\widetilde{\gamma}^{(1)}_{jkd}$, and $\widetilde{\gamma}^{(2)}_{jkd}$ denote zeroth-, first-, and second-order sufficient statistics of the GMM, respectively, and are defined as follows:

$$\begin{cases} \widetilde{\xi}_{ij} & \triangleq \sum_{t=1}^{T-1} \widetilde{\xi}_t(i,j), \\ \widetilde{\gamma}_{jk} & \triangleq \sum_{t=1}^{T} \widetilde{\gamma}_t(j,k), \\ \widetilde{\gamma}^{(1)}_{jkd} & \triangleq \sum_{t=1}^{T} \widetilde{\gamma}_t(j,k) o_{td}, \\ \widetilde{\gamma}^{(2)}_{jkd} & \triangleq \sum_{t=1}^{T} \widetilde{\gamma}_t(j,k)(o_{td})^2. \end{cases} \tag{7.67}$$

These sufficient statistics $\tilde{\Xi} \triangleq \{\tilde{\xi}_{ij}, \tilde{\gamma}_{jk}, \tilde{\gamma}^{(1)}_{jkd}, \tilde{\gamma}^{(2)}_{jkd} | i,j = 1,\ldots,J, k = 1,\ldots,K, d = 1,\cdots,D\}$ are computed by using $\tilde{\xi}_t(i,j)$ and $\tilde{\gamma}_t(j,k)$, defined as follows:

$$\begin{cases} \tilde{\xi}_t(i,j) & \triangleq & \tilde{q}(s_t = i, s_{t+1} = j | \mathbf{O}, M), \\ \tilde{\gamma}_t(j,k) & \triangleq & \tilde{q}(s_t = j, v_t = k | \mathbf{O}, M). \end{cases} \tag{7.68}$$

Here, $\tilde{\xi}_t(i,j)$ is a VB transition posterior distribution, which denotes the transition probability from a state i to a state j at a frame t, and $\tilde{\gamma}_t(j,k)$ is a VB occupation posterior distribution, which denotes the occupation probability of a mixture component k in a state j at a frame t, in the VB approach. These are similar to those defined in Eq. (3.119) by using the ML–EM algorithm of a CDHMM. Therefore, $\tilde{\Phi}$ can be calculated from Φ^0, $\tilde{\xi}_t(i,j)$, and $\tilde{\gamma}_t(j,k)$, enabling $\tilde{q}(\Theta|\mathbf{O},M)$ to be obtained.

VB E-step

Before we focus on the calculation of $\tilde{\xi}_t(i,j)$ and $\tilde{\gamma}_t(j,k)$, we first focus on the posterior distribution of the joint distribution of the HMM state and mixture component sequences $\tilde{q}(S,V|\Theta,M)$. From Eq. (7.25), $\tilde{q}(S,V|\Theta,M)$ is represented as follows:

$$\tilde{q}(S,V|\mathbf{O},M) = \frac{\tilde{q}(\mathbf{O},S,V|M)}{\tilde{q}(\mathbf{O}|M)} \propto \exp\left(\mathbb{E}_{(\Theta)}\left[\log p(\mathbf{O},S,V|\Theta,M)\right]\right). \tag{7.69}$$

Since $\tilde{q}(\mathbf{O}|M)$ does not depend on S and V, we find that the complete data likelihood marginalized by the model parameter Θ can also be obtained by considering the same expectation with Eq. (7.69):

$$\tilde{q}(\mathbf{O},S,V|M) \propto \exp\left(\mathbb{E}_{(\Theta)}\left[\log p(\mathbf{O},S,V|\Theta,M)\right]\right). \tag{7.70}$$

Once we obtain the complete data likelihood $\tilde{q}(\mathbf{O},S,V|M)$, we can compute the posterior probabilities $\tilde{\xi}_t(i,j)$ and $\tilde{\gamma}_t(j,k)$ by using the forward–backward algorithm, similarly to Section 3.3. Therefore, we focus on how to compute the following expectation of the complete data log likelihood:

$$\mathbb{E}_{(\Theta)}\left[\log p(\mathbf{O},S,V|\Theta,M)\right]. \tag{7.71}$$

By substituting Eq. (7.53) into Eq. (7.71), Eq. (7.71) can be represented as follows:

$$\begin{aligned} &\mathbb{E}_{(\Theta)}\left[\log p(\mathbf{O},S,V|\Theta,M)\right] \\ &\quad = \mathbb{E}_{(\Theta)}\left[\log \prod_{t=1}^{T} a_{s_{t-1}s_t}\omega_{s_tv_t}\mathcal{N}(\mathbf{o}_t|\boldsymbol{\mu}_{s_tv_t},\boldsymbol{\Sigma}_{s_tv_t})\right] \\ &\quad = \sum_{t=1}^{T}\mathbb{E}_{(\Theta)}\left[\log(a_{s_{t-1}s_t}) + \log(\omega_{s_tv_t}) + \log(\mathcal{N}(\mathbf{o}_t|\boldsymbol{\mu}_{s_tv_t},\boldsymbol{\Sigma}_{s_tv_t}))\right] \\ &\quad = \sum_{t=1}^{T}\mathbb{E}_{(\Theta)}\left[\log(a_{s_{t-1}s_t})\right] + \mathbb{E}_{(\Theta)}\left[\log(\omega_{s_tv_t})\right] + \mathbb{E}_{(\Theta)}\left[\log(\mathcal{N}(\mathbf{o}_t|\boldsymbol{\mu}_{s_tv_t},\boldsymbol{\Sigma}_{s_tv_t}))\right]. \end{aligned} \tag{7.72}$$

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

Now we focus on the case when $s_{t-1} = i$, $s_t = j$, and $v_t = k$, where we need to compute the following equations:

$$\begin{aligned} \log \widetilde{a}_{ij} &\triangleq \mathbb{E}_{(a_{ij})}\left[\log(a_{ij})\right], \\ \log \widetilde{\omega}_{jk} &\triangleq \mathbb{E}_{(\omega_{jk})}\left[\log(\omega_{jk})\right], \\ \log \widetilde{b}_{jk}(\mathbf{o}_t) &\triangleq \mathbb{E}_{(\boldsymbol{\mu}_{jk},\boldsymbol{\Sigma}_{jk})}\left[\log(\mathcal{N}(\mathbf{o}_t|\boldsymbol{\mu}_{jk},\boldsymbol{\Sigma}_{jk}))\right]. \end{aligned} \tag{7.73}$$

We can also define the following function based on Eq. (7.73):

$$\tilde{u}(\mathbf{O}, S, V|M) \triangleq \prod_{t=1}^{T} \widetilde{a}_{s_{t-1}s_t}\widetilde{\omega}_{s_t v_t}\widetilde{b}_{s_t v_t}(\mathbf{o}_t). \tag{7.74}$$

This equation behaves similarly to the likelihood function of $p(\mathbf{O}, S, V|\Theta, M)$ in Eq. (7.53). Note that $\tilde{u}(\mathbf{O}, S, V|M)$ is not properly normalized, and cannot be dealt with as a probabilistic distribution. However, from Eq. (7.70), $\tilde{u}(\mathbf{O}, S, V|M)$ is proportional to $\tilde{q}(\mathbf{O}, S, V|M)$, and this function has the following relationship from Eq. (7.69):

$$\begin{aligned} \tilde{q}(\mathbf{O}, S, V|M) &= \tilde{q}(\mathbf{O}|M)\tilde{q}(S, V|\mathbf{O}, M) \\ &= \frac{\tilde{q}(\mathbf{O}|M)}{\sum_{S',V'} \tilde{u}(\mathbf{O}, S', V'|M)}\tilde{u}(\mathbf{O}, S, V|M) \\ &\triangleq \prod_{t=1}^{T} C_a\widetilde{a}_{s_{t-1}s_t} C_{\omega b}\widetilde{\omega}_{s_t v_t}\widetilde{b}_{s_t v_t}(\mathbf{o}_t), \end{aligned} \tag{7.75}$$

where C_a and $C_{\omega b}$ are normalization constants of $\widetilde{a}_{ij}$ and $\widetilde{\omega}_{jk}\widetilde{b}_{jk}(\mathbf{o}_t)$ respectively for each frame, and satisfy the following condition:

$$\frac{\tilde{q}(\mathbf{O}|M)}{\sum_{S',V'} \tilde{u}(\mathbf{O}, S', V'|M)} = (C_a C_{\omega b})^T. \tag{7.76}$$

Note that it is not easy to obtain the normalization factors C_a and $C_{\omega b}$ explicitly, since it requires $\tilde{q}(\mathbf{O}|M)$ and $\sum_{S',V'} \tilde{u}(\mathbf{O}, S', V'|M)$. However, it will be shown later that the calculation of the occupation probabilities does not require computation of the normalization factors explicitly, but only requires $\widetilde{a}_{ij}$, $\widetilde{\omega}_{jk}$, and $\widetilde{b}_{jk}$. Therefore, we can compute various values from Eq. (7.74) (e.g., the forward and backward variables and the occupation probabilities), as discussed in Section 3.3. Thus, the following paragraphs provide the analytical solutions of $\widetilde{a}_{ij}$, $\widetilde{\omega}_{jk}$, and $\widetilde{b}_{jk}$ in detail.

State transition $\widetilde{a}_{ij}$

First, the integral over a_{ij} is solved from Eq. (7.63) by using a partial integral technique and a normalization constant:

$$\begin{aligned} \log \widetilde{a}_{ij} &= \int \tilde{q}(\{a_{ij'}\}_{j'=1}^{J}|M) \log a_{ij} \prod_{j'=1}^{J} da_{ij'} \\ &= C_{\mathrm{Dir}}(\{\widetilde{\phi}^a_{ij'}\}_{j'=1}^{J}) \int \log a_{ij} \prod_{j'=1}^{J} (a_{ij'})^{\widetilde{\phi}^a_{ij'}-1} da_{ij'}. \end{aligned} \tag{7.77}$$

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

Then we use the following derivative formula:

$$\frac{\partial}{\partial \widetilde{\phi}^a_{ij}} (a_{ij})^{\widetilde{\phi}^a_{ij}-1} = (\log a_{ij})(a_{ij})^{\widetilde{\phi}^a_{ij}-1}. \tag{7.78}$$

By substituting Eq. (7.78) into Eq. (7.77), Eq. (7.77) can be rewritten as:

$$\begin{aligned}
\log \widetilde{a}_{ij} &= C_{\text{Dir}}(\{\widetilde{\phi}^a_{ij'}\}_{j'=1}^J) \int \frac{\partial}{\partial \widetilde{\phi}^a_{ij}} (a_{ij})^{\widetilde{\phi}^a_{ij}-1} da_{ij} \int \prod_{j' \neq j}^{J} (a_{ij'})^{\widetilde{\phi}^a_{ij'}-1} da_{ij'} \\
&= C_{\text{Dir}}(\{\widetilde{\phi}^a_{ij'}\}_{j'=1}^J) \frac{\partial}{\partial \widetilde{\phi}^a_{ij}} \int \prod_{j'=1}^{J} (a_{ij'})^{\widetilde{\phi}^a_{ij'}-1} da_{ij'} \\
&= C_{\text{Dir}}(\{\widetilde{\phi}^a_{ij'}\}_{j'=1}^J) \frac{\partial}{\partial \widetilde{\phi}^a_{ij}} \frac{1}{C_{\text{Dir}}(\{\widetilde{\phi}^a_{ij'}\}_{j'=1}^J)},
\end{aligned} \tag{7.79}$$

where we replace the derivative and integral, and the integral can be performed to derive the inverse of the normalization constant of the Dirichlet distribution.

From Eq. (7.64), this derivative can be calculated as follows:

$$\begin{aligned}
&\frac{\partial}{\partial \widetilde{\phi}^a_{ij}} \frac{1}{C_{\text{Dir}}(\{\widetilde{\phi}^a_{ij'}\}_{j'=1}^J)} \\
&\quad = \frac{\partial}{\partial \widetilde{\phi}^a_{ij}} \frac{\prod_{j'=1}^J \Gamma(\widetilde{\phi}^a_{ij'})}{\Gamma(\sum_{j'=1}^J \widetilde{\phi}^a_{ij'})} \\
&\quad = \frac{\left(\frac{\partial}{\partial \widetilde{\phi}^a_{ij}} \Gamma(\widetilde{\phi}^a_{ij})\right) \prod_{j' \neq j}^J \Gamma(\widetilde{\phi}^a_{ij'}) \Gamma(\sum_{j'=1}^J \widetilde{\phi}^a_{ij'}) - \prod_{j'=1}^J \Gamma(\widetilde{\phi}^a_{ij'}) \left(\frac{\partial}{\partial \widetilde{\phi}^a_{ij}} \Gamma(\sum_{j'=1}^J \widetilde{\phi}^a_{ij'})\right)}{\left(\Gamma(\sum_{j'=1}^J \widetilde{\phi}^a_{ij'})\right)^2} \\
&\quad = \frac{\Psi(\widetilde{\phi}^a_{ij}) \prod_{j'=1}^J \Gamma(\widetilde{\phi}^a_{ij'}) \Gamma(\sum_{j'=1}^J \widetilde{\phi}^a_{ij'}) - \prod_{j'=1}^J \Gamma(\widetilde{\phi}^a_{ij'}) \Psi(\sum_{j'=1}^J \widetilde{\phi}^a_{ij'}) \Gamma(\sum_{j'=1}^J \widetilde{\phi}^a_{ij'})}{\left(\Gamma(\sum_{j'=1}^J \widetilde{\phi}^a_{ij'})\right)^2} \\
&\quad = \frac{\prod_{j'=1}^J \Gamma(\widetilde{\phi}^a_{ij'}) \left(\Psi(\widetilde{\phi}^a_{ij}) - \Psi(\sum_{j'=1}^J \widetilde{\phi}^a_{ij'})\right)}{\Gamma(\sum_{j'=1}^J \widetilde{\phi}^a_{ij'})} \\
&\quad = \frac{1}{C_{\text{Dir}}(\{\widetilde{\phi}^a_{ij'}\}_{j'=1}^J)} \left(\Psi(\widetilde{\phi}^a_{ij}) - \Psi(\sum_{j'=1}^J \widetilde{\phi}^a_{ij'})\right),
\end{aligned} \tag{7.80}$$

where $\Psi(y)$ is a di-gamma function, which first appeared in Eq. (5.82), and is defined as

$$\Psi(y) \triangleq \frac{\partial}{\partial y} \log \Gamma(y) = \frac{\frac{\partial}{\partial y}\Gamma(y)}{\Gamma(y)}. \tag{7.81}$$

Thus, $\widetilde{a}_{ij}$ is finally obtained as follows:

$$\log \widetilde{a}_{ij} = \Psi(\widetilde{\phi}^a_{ij}) - \Psi(\sum_{j'=1}^J \widetilde{\phi}^a_{ij'}). \tag{7.82}$$

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

Mixture weight $\widetilde{\omega}_{jk}$

In a way similar to that used for $\widetilde{a}_{ij}$, the integral over ω_{jk} is solved from Eq. (7.63), and $\widetilde{\omega}_{jk}$ is obtained as follows:

$$\log \widetilde{\omega}_{jk} = \Psi(\widetilde{\phi}^{\omega}_{jk}) - \Psi(\sum_{k'=1}^{K} \widetilde{\phi}^{\omega}_{jk'}). \tag{7.83}$$

Gaussian distribution $\widetilde{b}_{jk}(\mathbf{o}_t)$

First, $\log \widetilde{b}_{jk}(\mathbf{o}_t)$ can be factorized for each dimension:

$$\begin{aligned}\log \widetilde{b}_{jk}(\mathbf{o}_t) &= \mathbb{E}_{(\boldsymbol{\mu}_{jk},\boldsymbol{\Sigma}_{jk})}\left[\log(\mathcal{N}(\mathbf{o}_t|\boldsymbol{\mu}_{jk},\boldsymbol{\Sigma}_{jk}))\right] \\ &= \mathbb{E}_{(\boldsymbol{\mu}_{jk},\boldsymbol{\Sigma}_{jk})}\left[\log\left(\prod_{d=1}^{D}\mathcal{N}(o_{td}|\mu_{jkd},\Sigma_{jkd})\right)\right] \\ &= \sum_{d=1}^{D}\mathbb{E}_{(\mu_{jkd},\Sigma_{jkd})}\left[\log\left(\mathcal{N}(o_{td}|\mu_{jkd},\Sigma_{jkd})\right)\right].\end{aligned} \tag{7.84}$$

Therefore, we focus on calculation of the d element. Since the calculation is more complicated than the two previous calculations, the indexes j, k, t, and d are removed to simplify the derivation. By using (7.63), $\log \widetilde{b}(o)$ can be rewritten as follows:

$$\begin{aligned}&\log \widetilde{b}(o) \\ &= \int \mathcal{N}(\mu|\widetilde{\mu},(\widetilde{\phi}^{\mu}r)^{-1})\mathrm{Gam}_2(r|\widetilde{\phi}^{r},\widetilde{r}) \\ &\qquad \times \left(-\frac{1}{2}\left(\log(2\pi) - \log(r) + r(o-\mu)^2\right)\right) d\mu dr.\end{aligned} \tag{7.85}$$

Now we focus on the integral over mean parameter μ. To calculate the integral, we first rewrite the part that is related to μ as follows:

$$\begin{aligned}&\int \mathcal{N}(\mu|\widetilde{\mu},(\widetilde{\phi}^{\mu}r)^{-1})r(o-\mu)^2 d\mu \\ &= r\int \mathcal{N}(\mu|\widetilde{\mu},(\widetilde{\phi}^{\mu}r)^{-1})(o-\mu+\widetilde{\mu}-\widetilde{\mu})^2 d\mu \\ &= r\int \mathcal{N}(\mu|\widetilde{\mu},(\widetilde{\phi}^{\mu}r)^{-1})\left((\mu-\widetilde{\mu})^2+(o-\widetilde{\mu})^2-2(\mu-\widetilde{\mu})(o-\widetilde{\mu})\right)d\mu.\end{aligned} \tag{7.86}$$

The integral of the above terms can be analytically solved. We first consider the following partial derivative:

$$\frac{\partial}{\partial \widetilde{\phi}^{\mu} r}\exp\left(-\frac{1}{2}(\mu-\widetilde{\mu})^2\widetilde{\phi}^{\mu}r\right) = -\frac{1}{2}(\mu-\widetilde{\mu})^2\exp\left(-\frac{1}{2}(\mu-\widetilde{\mu})^2\widetilde{\phi}^{\mu}r\right). \tag{7.87}$$

Therefore, the first integral can be represented as:

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

$$\begin{aligned}
&\int \mathcal{N}(\mu|\widetilde{\mu},(\widetilde{\phi}^{\mu}r)^{-1})(\mu-\widetilde{\mu})^2 d\mu \\
&\quad = (2\pi)^{-\frac{1}{2}}(\widetilde{\phi}^{\mu}r)^{\frac{1}{2}} \int \exp\left(-\frac{1}{2}(\mu-\widetilde{\mu})^2\widetilde{\phi}^{\mu}r\right)(\mu-\widetilde{\mu})^2 d\mu \\
&\quad = (2\pi)^{-\frac{1}{2}}(\widetilde{\phi}^{\mu}r)^{\frac{1}{2}}(-2) \int \frac{\partial}{\partial\widetilde{\phi}^{\mu}r} \exp\left(-\frac{1}{2}(\mu-\widetilde{\mu})^2\widetilde{\phi}^{\mu}r\right) d\mu.
\end{aligned} \tag{7.88}$$

By replacing the integral with the partial derivative, we can solve the integral as:

$$\begin{aligned}
&\int \mathcal{N}(\mu|\widetilde{\mu},(\widetilde{\phi}^{\mu}r)^{-1})(\mu-\widetilde{\mu})^2 d\mu \\
&\qquad = (-2)(2\pi)^{-\frac{1}{2}}(\widetilde{\phi}^{\mu}r)^{\frac{1}{2}} \frac{\partial}{\partial\widetilde{\phi}^{\mu}r} \int \exp\left(-\frac{1}{2}(\mu-\widetilde{\mu})^2\widetilde{\phi}^{\mu}r\right) d\mu \\
&\qquad = (-2)(2\pi)^{-\frac{1}{2}}(\widetilde{\phi}^{\mu}r)^{\frac{1}{2}} \frac{\partial}{\partial\widetilde{\phi}^{\mu}r}(2\pi)^{\frac{1}{2}}(\widetilde{\phi}^{\mu}r)^{-\frac{1}{2}} \\
&\qquad = (-2)(\widetilde{\phi}^{\mu}r)^{\frac{1}{2}}\left(-\frac{1}{2}\right)(\widetilde{\phi}^{\mu}r)^{-\frac{3}{2}} = (\widetilde{\phi}^{\mu}r)^{-1}.
\end{aligned} \tag{7.89}$$

The other two integrals are trivially solved as follows:

$$\begin{aligned}
\int \mathcal{N}(\mu|\widetilde{\mu},(\widetilde{\phi}^{\mu}r)^{-1})(o-\widetilde{\mu})^2 d\mu &= (o-\widetilde{\mu})^2, \\
\int \mathcal{N}(\mu|\widetilde{\mu},(\widetilde{\phi}^{\mu}r)^{-1})(\mu-\widetilde{\mu})(o-\widetilde{\mu}) d\mu &= 0.
\end{aligned} \tag{7.90}$$

Therefore, Eq. (7.86) is solved as:

$$\begin{aligned}
&\int \mathcal{N}(\mu|\widetilde{\mu},(\widetilde{\phi}^{\mu}r)^{-1}) r(o-\mu)^2 d\mu \\
&\qquad = r\left((o-\widetilde{\mu})^2 + (\widetilde{\phi}^{\mu}r)^{-1}\right) = r(o-\widetilde{\mu})^2 + \frac{1}{\widetilde{\phi}^{\mu}}.
\end{aligned} \tag{7.91}$$

Now, we focus on the integral over r, because the integral without $\log(r)$ can be easily computed by the result of the mean value of the gamma distribution in Appendix C.11 as:

$$\begin{aligned}
&\log \widetilde{b}(o) \\
&= \int \mathrm{Gam}_2(r|\widetilde{\phi}^{r},\widetilde{r})\left(-\frac{1}{2}\left(\log(2\pi) - \log(r) + r(o-\widetilde{\mu})^2 + \frac{1}{\widetilde{\phi}^{\mu}}\right)\right) dr \\
&= -\frac{1}{2}\left(\log(2\pi) + \frac{\widetilde{\phi}^{r}}{\widetilde{r}}(o-\widetilde{\mu})^2 + \frac{1}{\widetilde{\phi}^{\mu}}\right) + \frac{1}{2}\int \mathrm{Gam}_2(r|\widetilde{\phi}^{r},\widetilde{r}) \log(r) dr.
\end{aligned} \tag{7.92}$$

Therefore, we focus on the final term. From Eqs. (7.63) and (7.64), the concrete form of the gamma distribution, $\mathrm{Gam}_2(\cdot)$, is defined as follows:

$$\mathrm{Gam}_2(r|\widetilde{\phi}^{r},\widetilde{r}) = C_{\mathrm{Gam}_2}(\widetilde{\phi}^{r},\widetilde{r})(r)^{\frac{\widetilde{\phi}^{r}}{2}-1}\exp\left(-\frac{\widetilde{r}r}{2}\right), \tag{7.93}$$

where

$$C_{\mathrm{Gam}_2}(\widetilde{\phi}^{r},\widetilde{r}) = \frac{\left(\frac{\widetilde{r}}{2}\right)^{\frac{\widetilde{\phi}^{r}}{2}}}{\Gamma\left(\frac{\widetilde{\phi}^{r}}{2}\right)}. \tag{7.94}$$

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

Similarly to the Dirichlet and Gaussian distributions, we consider the following derivative:

$$\frac{\partial}{\partial \widetilde{\phi}^r}(r)^{\frac{\widetilde{\phi}^r}{2}-1} = \frac{1}{2}(r)^{\frac{\widetilde{\phi}^r}{2}-1}\log(r). \tag{7.95}$$

Therefore, the integral is solved by using this relationship as:

$$\begin{aligned}
&\int \mathrm{Gam}_2(r|\widetilde{\phi}^r, \widetilde{r}\,)\log(r)dr \\
&\quad = C_{\mathrm{Gam}_2}(\widetilde{\phi}^r, \widetilde{r}\,)\int (r)^{\frac{\widetilde{\phi}^r}{2}-1}\exp\left(-\frac{\widetilde{r}r}{2}\right)\log(r)dr \\
&\quad = C_{\mathrm{Gam}_2}(\widetilde{\phi}^r, \widetilde{r}\,)\int 2\frac{\partial}{\partial \widetilde{\phi}^r}(r)^{\frac{\widetilde{\phi}^r}{2}-1}\exp\left(-\frac{\widetilde{r}r}{2}\right)dr \\
&\quad = 2C_{\mathrm{Gam}_2}(\widetilde{\phi}^r, \widetilde{r}\,)\frac{\partial}{\partial \widetilde{\phi}^r}\frac{1}{C_{\mathrm{Gam}_2}(\widetilde{\phi}^r, \widetilde{r}\,)}.
\end{aligned} \tag{7.96}$$

The derivative with respect to $\widetilde{\phi}^r$ is calculated as follows:

$$\begin{aligned}
\frac{\partial}{\partial \widetilde{\phi}^r}\frac{\Gamma\left(\frac{\widetilde{\phi}^r}{2}\right)}{\left(\frac{\widetilde{r}}{2}\right)^{\frac{\widetilde{\phi}^r}{2}}} &= \frac{\frac{1}{2}\frac{\partial}{\partial\frac{\widetilde{\phi}^r}{2}}\Gamma\left(\frac{\widetilde{\phi}^r}{2}\right)\left(\frac{\widetilde{r}}{2}\right)^{\frac{\widetilde{\phi}^r}{2}} + \frac{1}{2}\log\left(\frac{\widetilde{r}}{2}\right)\left(\frac{\widetilde{r}}{2}\right)^{\frac{\widetilde{\phi}^r}{2}}\Gamma\left(\frac{\widetilde{\phi}^r}{2}\right)}{\left(\frac{\widetilde{r}}{2}\right)^{\widetilde{\phi}^r}} \\
&= \frac{\frac{1}{2}\Psi\left(\frac{\widetilde{\phi}^r}{2}\right)\Gamma\left(\frac{\widetilde{\phi}^r}{2}\right) + \frac{1}{2}\log\left(\frac{\widetilde{r}}{2}\right)\Gamma\left(\frac{\widetilde{\phi}^r}{2}\right)}{\left(\frac{\widetilde{r}}{2}\right)^{\frac{\widetilde{\phi}^r}{2}}},
\end{aligned} \tag{7.97}$$

where $\Psi(\cdot)$ is a di-gamma function defined in Eq. (7.81). Therefore,

$$2C_{\mathrm{Gam}_2}(\widetilde{\phi}^r, \widetilde{r})\frac{\partial}{\partial \widetilde{\phi}^r}\frac{1}{C_{\mathrm{Gam}_2}(\widetilde{\phi}^r, \widetilde{r}\,)} = \Psi\left(\frac{\widetilde{\phi}^r}{2}\right) + \log\left(\frac{\widetilde{r}}{2}\right). \tag{7.98}$$

Thus, finally $\log \widetilde{b}(o)$ is obtained analytically as follows:

$$\begin{aligned}
&\log \widetilde{b}(o) \\
&\quad = -\frac{1}{2}\int \mathrm{Gam}_2(r|\widetilde{\phi}^r, \widetilde{r})\left(\log(2\pi) + \frac{1}{\widetilde{\phi}^\mu} - \log(r) + r(o - \widetilde{\mu})^2\right)dr \\
&\quad = -\frac{1}{2}\left(\log(2\pi) + \frac{1}{\widetilde{\phi}^\mu} - \Psi\left(\frac{\widetilde{\phi}^r}{2}\right)\right) - \frac{1}{2}\left(\log\left(\frac{\widetilde{r}}{2}\right) + (o - \widetilde{\mu})^2\frac{\widetilde{\phi}^r}{\widetilde{r}}\right).
\end{aligned} \tag{7.99}$$

Reverting to the indexes k, j, t, and d, $\log \widetilde{b}_{jk}(\mathbf{o}_t)$ is represented as

$$\begin{aligned}
\log \widetilde{b}_{jk}(\mathbf{o}_t) = &-\frac{D}{2}\left(\log(2\pi) + \frac{1}{\widetilde{\phi}^\mu_{jk}} - \Psi\left(\frac{\widetilde{\phi}^r_{jk}}{2}\right)\right) \\
&-\frac{1}{2}\sum_{d=1}^{D}\left(\log\left(\frac{\widetilde{r}_{jk}}{2}\right) + \frac{\widetilde{\phi}^r_{jk}(o_{td} - \widetilde{\mu}_{jkd})^2}{\widetilde{r}_{jkd}}\right).
\end{aligned} \tag{7.100}$$

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

Thus, we obtain $\widetilde{a}_{ij}, \widetilde{\omega}_{jk}$ and $\widetilde{b}_{jk}(\mathbf{o}_t)$, which are summarized as follows:

$$\begin{cases} \widetilde{a}_{ij} & \triangleq \exp\big(\Psi(\widetilde{\phi}^a_{ij}) - \Psi(\sum_{j'} \widetilde{\phi}^a_{ij'})\big), \\ \widetilde{\omega}_{jk} & \triangleq \exp\big(\Psi(\widetilde{\phi}^\omega_{jk}) - \Psi(\sum_{k'} \widetilde{\phi}^\omega_{jk'})\big), \\ \widetilde{b}_{jk}(\mathbf{o}_t) & \triangleq \exp\Bigg(-\frac{D}{2}\left(\log(2\pi) + \frac{1}{\widetilde{\phi}^\mu_{jk}} - \Psi\left(\frac{\widetilde{\phi}^r_{jk}}{2}\right)\right) \\ & \qquad -\frac{1}{2}\sum_{d=1}^{D}\left(\log\left(\frac{\widetilde{r}_{jk}}{2}\right) + \frac{\widetilde{\phi}^r_{jk}(o_{td} - \widetilde{\mu}_{jkd})^2}{\widetilde{r}_{jkd}}\right)\Bigg). \end{cases} \tag{7.101}$$

These variables are used to compute the VB transition probability $\widetilde{\xi}_t(i,j)$ and VB occupation probability $\widetilde{\gamma}_t(j,k)$.

VB transition probability $\widetilde{\xi}_t(i,j)$ and occupation probability $\widetilde{\gamma}_t(j,k)$ (VB E-step)

From Eq. (7.25), VB transition probability $\widetilde{\xi}_t(i,j)$ is represented as:

$$\widetilde{\xi}_t(i,j) \triangleq \widetilde{q}(s_t = i, s_{t+1} = j|\mathbf{O}, M). \tag{7.102}$$

Section 3.4.2 shows an efficient computation of the transition probability based on the complete data likelihood $p(\mathbf{O}, S, V|\Theta)$. Here we also consider how to obtain it based on the VB version of the complete data likelihood $\tilde{q}(\mathbf{O}, S, V|M)$, as introduced in Eq. (7.69). However, from Eq. (7.74), $\widetilde{a}_{ij}, \widetilde{\omega}_{jk}$, and $\widetilde{b}_{jk}(\mathbf{o}_t)$ can only compute the unnormalized likelihood function $\tilde{u}(\mathbf{O}, S, V|M)$, that is

$$\tilde{u}(\mathbf{O}, S, V|M) = \prod_{t=1}^{T} \widetilde{a}_{s_{t-1}s_t}\widetilde{\omega}_{s_t v_t}\widetilde{b}_{s_t v_t}(\mathbf{o}_t). \tag{7.103}$$

Therefore, as we discussed in Section 3.4.2, from the dependency of the HMM, we can represent $\widetilde{q}(\mathbf{O}, s_t = i, s_{t+1} = j|M)$ as follows:

$$\begin{aligned} &\widetilde{q}(s_t = i, s_{t+1} = j, \mathbf{O}|M) \\ &\quad = \underbrace{\widetilde{q}(\mathbf{o}_1, \cdots, \mathbf{o}_t, s_t = i|M)}_{=\widetilde{\alpha}_t(i)} \underbrace{\widetilde{q}(\mathbf{o}_{t+1}|s_{t+1} = j, M)}_{=C_{\omega b}\sum_{k=1}^{K}\widetilde{\omega}_{jk}\widetilde{b}_{jk}(\mathbf{o}_{t+1})} \underbrace{\widetilde{q}(\mathbf{o}_{t+2}, \cdots, \mathbf{o}_T|s_{t+1} = j, M)}_{=\widetilde{\beta}_{t+1}(j)} \\ &\qquad \times \underbrace{\widetilde{q}(s_{t+1} = j|s_t = i, M)}_{=C_a\widetilde{a}_{ij}}. \end{aligned} \tag{7.104}$$

Here, $\widetilde{\alpha}_t(i)$ is a forward variable at frame t in state i, as introduced in Eq. (3.50). Similarly, $\widetilde{\beta}_{t+1}(j)$ is a backward variable at frame $t+1$ in state j, as introduced in Eq. (3.55). The forward and backward variables based on the VB formulation are represented as described below.

First, the VB forward variable $\widetilde{\alpha}_t(j)$ is computed by using the following equation:

- Initialization

$$\begin{aligned} \widetilde{\alpha}_1(j) &= \widetilde{q}(\mathbf{o}_1, s_1 = j|M) \\ &= \widetilde{q}(\mathbf{o}_1|s_1 = j, M)\widetilde{q}(s_1 = j|M) \\ &= C_a\widetilde{a}_j C_{\omega b}\sum_{k=1}^{K}\widetilde{\omega}_{jk}\widetilde{b}_{jk}(\mathbf{o}_1), \qquad 1 \le j \le J. \end{aligned} \tag{7.105}$$

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

Then, unnormalized forward variable $\tilde{\tilde{\alpha}}_1(j)$ is defined as:

$$\begin{aligned}\tilde{\tilde{\alpha}}_1(j) &\triangleq \tilde{a}_j \sum_{k=1}^{K} \widetilde{\omega}_{jk} \widetilde{b}_{jk}(\mathbf{o}_1) \\ &= \frac{\widetilde{\alpha}_1(j)}{C_a C_{\omega b}}.\end{aligned} \tag{7.106}$$

- Induction

$$\begin{aligned}\widetilde{\alpha}_t(j) &= \widetilde{q}(\mathbf{o}_1, \cdots, \mathbf{o}_t, s_t = j|M) \\ &= \widetilde{q}(\mathbf{o}_t|s_t = j, M) \sum_{i=1}^{J} \widetilde{q}(s_t = j|s_{t-1} = i, M)\widetilde{q}(\mathbf{o}_1, \cdots, \mathbf{o}_{t-1}, s_{t-1} = i|M) \\ &= \left(C_a \sum_{i=1}^{J} \widetilde{\alpha}_{t-1}(i)\widetilde{a}_{ij} \right) C_{\omega b} \sum_{k=1}^{K} \widetilde{\omega}_{jk} \widetilde{b}_{jk}(\mathbf{o}_t) \\ &= \left(C_a (C_a C_{\omega b})^{t-1} \sum_{i=1}^{J} \tilde{\tilde{\alpha}}_{t-1}(i)\widetilde{a}_{ij} \right) C_{\omega b} \sum_{k=1}^{K} \widetilde{\omega}_{jk} \widetilde{b}_{jk}(\mathbf{o}_t) \\ &= (C_a C_{\omega b})^t \left(\sum_{i=1}^{J} \tilde{\tilde{\alpha}}_{t-1}(i)\widetilde{a}_{ij} \right) \sum_{k=1}^{K} \widetilde{\omega}_{jk} \widetilde{b}_{jk}(\mathbf{o}_t), \qquad \begin{matrix} 2 \leq t \leq T \\ 1 \leq j \leq J, \end{matrix}\end{aligned} \tag{7.107}$$

where the unnormalized forward variable $\tilde{\tilde{\alpha}}_t(j)$ is represented as

$$\tilde{\tilde{\alpha}}_t(j) = \left(\sum_{i=1}^{J} \tilde{\tilde{\alpha}}_{t-1}(i)\widetilde{a}_{ij} \right) \sum_{k=1}^{K} \widetilde{\omega}_{jk} \widetilde{b}_{jk}(\mathbf{o}_t), \qquad \begin{matrix} 2 \leq t \leq T \\ 1 \leq j \leq J. \end{matrix} . \tag{7.108}$$

- Termination

$$\begin{aligned}\widetilde{q}(\mathbf{O}|M) &= \sum_{j=1}^{J} \widetilde{\alpha}_T(j) \\ &= (C_a C_{\omega b})^T \sum_{j=1}^{J} \tilde{\tilde{\alpha}}_T(j).\end{aligned} \tag{7.109}$$

The VB forward variable $\tilde{\alpha}_t(j)$ is obtained with the unnormalized forward variable $\tilde{\tilde{\alpha}}_t(j)$ and normalization constants C_a and $C_{\omega b}$. From this algorithm, we can compute the unnormalized forward variable $\tilde{\tilde{\alpha}}_t(j)$ similarly to the original forward algorithm, but we should be careful that the unnormalized forward variable is not a probability, and probabilistic calculation (sum and product rules etc.) must be performed via the normalized VB forward variable $\tilde{\alpha}_t(j)$.

Similarly, the VB backward variable $\widetilde{\beta}_t(j)$ is computed by using the following equations:

- Initialization

$$\widetilde{\beta}_T(j) = 1, \qquad 1 \leq j \leq J. \tag{7.110}$$

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

- Induction

$$\begin{aligned}\widetilde{\beta}_t(i) &= \widetilde{q}(\mathbf{o}_{t+1}, \cdots, \mathbf{o}_T | s_t = i, M) \\ &= \sum_{j=1}^{J} \widetilde{q}(\mathbf{o}_{t+2}, \cdots, \mathbf{o}_T | s_{t+1} = j, M) \widetilde{q}(\mathbf{o}_{t+1} | s_{t+1} = j, M) \widetilde{q}(s_{t+1} = j | s_t = i, M) \\ &= \sum_{j=1}^{J} C_a \widetilde{a}_{ij} \sum_{k=1}^{K} C_{\omega b} \widetilde{\omega}_{jk} \widetilde{b}_{jk}(\mathbf{o}_{t+1}) \widetilde{\beta}_{t+1}(j) \\ &= (C_a C_{\omega b})^{T-t} \sum_{j=1}^{J} \widetilde{a}_{ij} \sum_{k=1}^{K} \widetilde{\omega}_{jk} \widetilde{b}_{jk}(\mathbf{o}_{t+1}) \tilde{\tilde{\beta}}_{t+1}(j), \\ & t = T-1, T-2, \cdots, 1, \quad 1 \le i \le J, \end{aligned} \tag{7.111}$$

where the unnormalized backward variable $\tilde{\tilde{\beta}}_t(i)$ is represented as

$$\tilde{\tilde{\beta}}_t(i) = \sum_{j=1}^{J} \widetilde{a}_{ij} \sum_{k=1}^{K} \widetilde{\omega}_{jk} \widetilde{b}_{jk}(\mathbf{o}_{t+1}) \tilde{\tilde{\beta}}_{t+1}(j). \tag{7.112}$$

- Termination

$$\begin{aligned}\beta_0 &\triangleq \widetilde{q}(\mathbf{O}|M) \\ &= \sum_{j=1}^{J} \widetilde{a}_j \sum_{k=1}^{K} \widetilde{\omega}_{jk} \widetilde{b}_{jk}(\mathbf{o}_1) \widetilde{\beta}_1(j) \\ &= (C_a C_{\omega b})^{T} \sum_{j=1}^{J} \widetilde{a}_j \sum_{k=1}^{K} \widetilde{\omega}_{jk} \widetilde{b}_{jk}(\mathbf{o}_1) \tilde{\tilde{\beta}}_1(j). \end{aligned} \tag{7.113}$$

Therefore, based on the VB forward and backward variables, we can compute the posterior probabilities as follows:

$$\begin{aligned} & \widetilde{\xi}_t(i,j) \\ &= \frac{\widetilde{\alpha}_t(i) \widetilde{a}_{ij} \left(\sum_{k=1}^{K} \widetilde{\omega}_{jk} \widetilde{b}_{jk}(\mathbf{o}_t) \right) \widetilde{\beta}_{t+1}(j)}{\sum_{i'=1}^{J} \sum_{j'=1}^{J} \widetilde{\alpha}_t(i') \widetilde{a}_{i'j'} \left(\sum_{k=1}^{K} \widetilde{\omega}_{j'k} \widetilde{b}_{j'k}(\mathbf{o}_t) \right) \widetilde{\beta}_{t+1}(j')} \\ &= \frac{(C_a C_{\omega b})^t \tilde{\tilde{\alpha}}_t(i) C_a \widetilde{a}_{ij} C_{\omega b} \left(\sum_{k=1}^{K} \widetilde{\omega}_{jk} \widetilde{b}_{jk}(\mathbf{o}_t) \right) (C_a C_{\omega b})^{T-t-1} \tilde{\tilde{\beta}}_{t+1}(j)}{\sum_{i'=1}^{J} \sum_{j'=1}^{J} (C_a C_{\omega b})^t \tilde{\tilde{\alpha}}_t(i') C_a \widetilde{a}_{i'j'} C_{\omega b} \left(\sum_{k=1}^{K} \widetilde{\omega}_{j'k} \widetilde{b}_{j'k}(\mathbf{o}_t) \right) (C_a C_{\omega b})^{T-t-1} \tilde{\tilde{\beta}}_{t+1}(j')} \\ &= \frac{\tilde{\tilde{\alpha}}_t(i) \widetilde{a}_{ij} \left(\sum_{k=1}^{K} \widetilde{\omega}_{jk} \widetilde{b}_{jk}(\mathbf{o}_t) \right) \tilde{\tilde{\beta}}_{t+1}(j)}{\sum_{i'=1}^{J} \sum_{j'=1}^{J} \tilde{\tilde{\alpha}}_t(i') \widetilde{a}_{i'j'} \left(\sum_{k=1}^{K} \widetilde{\omega}_{j'k} \widetilde{b}_{j'k}(\mathbf{o}_t) \right) \tilde{\tilde{\beta}}_{t+1}(j')}. \end{aligned} \tag{7.114}$$

Thus, we can compute the transition probability with unnormalized VB variables $\widetilde{a}_{ij}$, $\widetilde{\omega}_{jk}$, and $\widetilde{b}_{jk}(\mathbf{o}_t)$, and unnormalized forward and backward variables $\tilde{\tilde{\alpha}}_t(i)$ and $\tilde{\tilde{\beta}}_{t+1}(j)$, where the normalization constants C_a and $C_{\omega b}$ are canceled out. This is based on a well-known

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

scaling property of the HMM forward–backward algorithm (Rabiner & Juang 1993). Similarly the occupation probability is also calculated as:

$$\tilde{\gamma}_t(j,k) = \frac{\tilde{\tilde{\alpha}}_t(j)\tilde{\tilde{\beta}}_t(j)}{\sum_{j'=1}^{J}\tilde{\tilde{\alpha}}_t(j')\tilde{\tilde{\beta}}_t(j')} \cdot \frac{\tilde{\omega}_{jk}\tilde{b}_{jk}(\mathbf{o}_t)}{\sum_{k'=1}^{K}\tilde{\omega}_{jk'}\tilde{b}_{jk'}(\mathbf{o}_t)}. \tag{7.115}$$

These probabilities are obtained similarly to the ML cases in Eqs. (3.126) and (3.122), and the MAP case in Eqs. (4.91) and (4.92) with the VB-based variables obtained by Eq. (7.101). Thus, $\tilde{\xi}_t(i,j)$ and $\tilde{\gamma}_t(j,k)$ are calculated efficiently by using a probabilistic assignment via the familiar *forward–backward algorithm*. This algorithm is called the VB forward–backward algorithm.

Similarly to the VB forward–backward algorithm, the *Viterbi algorithm* is also derived within the VB approach by exchanging the summation over i for the maximization over i in the calculation of the unnormalized forward probability $\tilde{\tilde{\alpha}}_t(j)$. This algorithm is called the VB Viterbi algorithm.

Thus, VB posteriors can be calculated iteratively in the same way as the Baum–Welch algorithm, even for a complicated sequential model that includes latent variables such as HMM and GMM for acoustic models. These calculations are referred to as a VB Baum–Welch algorithm, as proposed in MacKay (1997), Watanabe *et al*. (2002), Beal (2003) and Watanabe *et al*. (2004).

7.3.4 Variational lower bound

This section discusses the VB objective function $\mathcal{F}^M$ for a whole acoustic model topology, i.e., the variational lower bound, and provides general calculation results. The variational lower bound is a criterion for both posterior distribution estimation, and model topology optimization in acoustic model construction. This section begins by focusing on one phoneme category. By substituting the VB posterior distribution obtained in Section 7.3.3, we obtain analytical results for $\mathcal{F}^M$, and therefore, this calculation also requires a VB iterative calculation based on the VB Baum–Welch algorithm used in the VB posterior calculation. We can separate $\mathcal{F}^M$ into two components: one is composed solely of $\tilde{q}(S,V|\mathbf{O},M)$, whereas the other is mainly composed of $\tilde{q}(\Theta|\mathbf{O},M)$. Therefore, we define $\mathcal{F}^M_\Theta$ and $\mathcal{F}^M_{S,V}$, and represent $\mathcal{F}^M$ as follows:

$$\begin{aligned}\mathcal{F}^M &= \mathbb{E}_{(\Theta,S,V)}\left[\log\frac{p(\mathbf{O},S,V|\Theta,M)p(\Theta|M)}{\tilde{q}(\Theta|\mathbf{O},M)}\right] - \mathbb{E}_{(S,V)}\left[\log\tilde{q}(S,V|\mathbf{O},M)\right] \\ &= \mathcal{F}^M_\Theta - \mathcal{F}^M_{S,V},\end{aligned} \tag{7.116}$$

where

$$\begin{aligned}\mathcal{F}^M_\Theta &\triangleq \mathbb{E}_{(\Theta,S,V)}\left[\log\frac{p(\mathbf{O},S,V|\Theta,M)p(\Theta|M)}{\tilde{q}(\Theta|\mathbf{O},M)}\right], \\ \mathcal{F}^M_{S,V} &\triangleq \mathbb{E}_{(S,V)}\left[\log\tilde{q}(S,V|\mathbf{O},M)\right].\end{aligned} \tag{7.117}$$

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

First, we focus on $\mathcal{F}^M_\Theta$. Based on the variational solution of $\widetilde{q}(S, V|\mathbf{O}, M)$ in Eq. (7.42), $\mathcal{F}^M_\Theta$ is rewritten as follows:

$$\begin{aligned}
\mathcal{F}^M_\Theta &\triangleq \mathbb{E}_{(\Theta,S,V)}\left[\log \frac{p(\mathbf{O}, S, V|\Theta, M)p(\Theta|M)}{\frac{1}{Z}p(\Theta|M)\exp\left(\mathbb{E}_{(S,V)}\left[\log p(\mathbf{O}, S, V|\Theta, M)\right]\right)}\right] \\
&= \mathbb{E}_{(\Theta,S,V)}\left[\log p(\mathbf{O}, S, V|\Theta, M)\right] + \mathbb{E}_{(\Theta)}\left[\log p(\Theta|M)\right] \\
&\quad - \mathbb{E}_{(\Theta,S,V)}\left[\log p(\mathbf{O}, S, V|\Theta, M)\right] - \mathbb{E}_{(\Theta)}\left[\log p(\Theta|M)\right] + \log Z \\
&= \log Z, \qquad (7.118)
\end{aligned}$$

where Z is a normalization constant. This equation means that $\mathcal{F}^M_\Theta$ is represented by the logarithmic function of the normalization constant Z. By using the definition of the VB auxiliary function $\tilde{q}(\Theta)$ in Eq. (7.60), Z can be rewritten as

$$\begin{aligned}
Z &\triangleq \int p(\Theta|M)\exp\left(\mathbb{E}_{(S,V)}\left[\log p(\mathbf{O}, S, V|\Theta, M)\right]\right) d\Theta \\
&= \int \exp\left(\tilde{Q}(\Theta)\right) d\Theta. \qquad (7.119)
\end{aligned}$$

Here, from the similarity of the VB and MAP auxiliary functions, as discussed in Section 7.3.3, $\tilde{Q}(\Theta)$ can be obtained by using the analytical results of the MAP auxiliary function. From Eq. (4.38), $\tilde{Q}(\Theta)$ is decomposed into the following auxiliary functions:

$$\tilde{Q}(\Theta) = \tilde{Q}(\mathbf{A}) + \tilde{Q}(\boldsymbol{\omega}) + \tilde{Q}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \qquad (7.120)$$

where $\tilde{Q}(\mathbf{A})$, $\tilde{Q}(\boldsymbol{\omega})$, and $\tilde{Q}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are obtained from the analytical solutions of the corresponding MAP auxiliary functions in Eqs. (4.51), (4.54), and (4.73), as follows:

$$\tilde{Q}(\mathbf{A}) = \sum_{i=1}^{J} \log\left(\mathrm{Dir}(\{a_{ij}\}_{j=1}^J | \{\tilde{\phi}^a_{ij}\}_{j=1}^J)\right) + \sum_{i=1}^{J} \log \frac{C_{\mathrm{Dir}}(\{\phi^a_{ij}\}_{j=1}^J)}{C_{\mathrm{Dir}}(\{\tilde{\phi}^a_{ij}\}_{j=1}^J)}, \qquad (7.121)$$

$$\tilde{Q}(\boldsymbol{\omega}) = \sum_{j=1}^{J} \log\left(\mathrm{Dir}(\{\omega_{jk}\}_{k=1}^K | \{\tilde{\phi}^\omega_{jk}\}_{k=1}^K)\right) + \sum_{j=1}^{J} \log \frac{C_{\mathrm{Dir}}(\{\phi^\omega_{jk}\}_{k=1}^K)}{C_{\mathrm{Dir}}(\{\tilde{\phi}^\omega_{jk}\}_{k=1}^K)}, \qquad (7.122)$$

$$\begin{aligned}
&\tilde{Q}(\boldsymbol{\mu}, \mathbf{R}) \\
&= \sum_{j=1}^{J}\sum_{k=1}^{K} \log\left(\mathcal{N}(\boldsymbol{\mu}_{jk}|\tilde{\boldsymbol{\mu}}_{jk}, (\tilde{\phi}^{\boldsymbol{\mu}}_{jk}\mathbf{R}_{jk})^{-1})\mathcal{W}(\mathbf{R}_{jk}|\tilde{\mathbf{R}}_{jk}, \tilde{\phi}^{\mathbf{R}}_{jk})\right) \\
&+ \sum_{j=1}^{J}\sum_{k=1}^{K}\left(-\sum_{t=1}^{T}\frac{\tilde{\gamma}_t(j,k)D}{2}\log(2\pi) + \frac{D}{2}\log\frac{\phi^{\boldsymbol{\mu}}}{\tilde{\phi}^{\boldsymbol{\mu}}} + \log\frac{C_{\mathcal{W}}(\mathbf{R}^0_{jk}, \phi^{\mathbf{R}}_{jk})}{C_{\mathcal{W}}(\tilde{\mathbf{R}}_{jk}, \tilde{\phi}^{\mathbf{R}}_{jk})}\right). \qquad (7.123)
\end{aligned}$$

If we consider the diagonal covariance matrix, Eq. (7.123) is modified by using the gamma distribution as follows:

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

$$
\begin{aligned}
&\tilde{Q}(\boldsymbol{\mu}, \mathbf{R}) \\
&= \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{d=1}^{D} \log \left(\mathcal{N}(\mu_{jkd} | \tilde{\mu}_{jkd}, (\tilde{\phi}^{\mu}_{jk} r_{jkd})^{-1}) \mathrm{Gam}_2(r_{jkd} | \tilde{r}_{jkd}, \tilde{\phi}^{r}_{jk}) \right) \\
&\quad + \sum_{j=1}^{J} \sum_{k=1}^{K} \left(- \sum_{t=1}^{T} \frac{\tilde{\gamma}_t(j,k) D}{2} \log(2\pi) + \frac{D}{2} \log \frac{\phi^{\mu}_{jk}}{\tilde{\phi}^{\mu}_{jk}} + \log \frac{C_{\mathrm{Gam}_2}(r^0_{jkd}, \phi^r_{jk})}{C_{\mathrm{Gam}_2}(\tilde{r}_{jkd}, \tilde{\phi}^r_{jk})} \right).
\end{aligned}
\tag{7.124}
$$

Here C_{Gam_2} in Eq. (7.124) and C_{Dir} in Eqs. (7.121) and (7.122) are normalization constants of Gamma and Dirichlet distributions, respectively, and are defined as follows:

$$
\begin{cases}
C_{\mathrm{Dir}}(\{\phi_j\}_{j=1}^{J}) & = \dfrac{\Gamma\left(\sum_{j=1}^{J} \phi_j\right)}{\prod_{j=1}^{J} \Gamma(\phi_j)}, \\
C_{\mathrm{Gam}_2}(\phi, r^0) & = \dfrac{\left(\frac{r^0}{2}\right)^{\frac{\phi}{2}}}{\Gamma\left(\frac{\phi}{2}\right)}.
\end{cases}
\tag{7.125}
$$

Therefore, by substituting Eqs. (7.121), (7.122), and (7.124) into Eq. (7.119), and by using the definition of the normalization constant of the Dirichlet and gamma distributions in Eq. (7.125), the integral in Z is performed with the normalization of Θ, and Z is simply obtained as follows:

$$
\begin{aligned}
\mathcal{F}^M_{\Theta} &= \log Z \\
&= \log \left(\int \prod_{i,j} \exp\left(\tilde{Q}(\mathbf{A})\right) da_{ij} \right) + \log \left(\int \prod_{j,k} \exp\left(\tilde{Q}(\boldsymbol{\omega})\right) d\omega_{jk} \right) \\
&\quad + \log \left(\int \prod_{j,k} \exp\left(\tilde{Q}(\boldsymbol{\mu}, \mathbf{R})\right) d\boldsymbol{\mu}_{jk} d\mathbf{R}_{jk} \right) \\
&= \sum_i \log \frac{\Gamma(\sum_j \phi^a_{ij})}{\Gamma(\sum_{j'} \widetilde{\phi}^a_{ij})} \frac{\prod_j \Gamma(\widetilde{\phi}^a_{ij})}{\prod_j \Gamma(\phi^a_{ij})} + \sum_j \log \frac{\Gamma(\sum_k \phi^{\omega}_{jk})}{\Gamma(\sum_k \widetilde{\phi}^{\omega}_{jk})} \frac{\prod_k \Gamma(\widetilde{\phi}^{\omega}_{jk})}{\prod_k \Gamma(\phi^{\omega}_{jk})} \\
&\quad + \sum_{j,k} \log \left((2\pi)^{-\frac{\widetilde{\gamma}_{jk} D}{2}} \left(\frac{\phi^{\mu}_{jk}}{\widetilde{\phi}^{\mu}_{jk}} \right)^{\frac{D}{2}} \frac{\left(\Gamma\left(\frac{\widetilde{\phi}^r_{jk}}{2} \right) \right)^D \prod_d \left(\frac{r^0_{jkd}}{2} \right)^{\frac{\phi^r_{jk}}{2}}}{\left(\Gamma\left(\frac{\phi^r_{jk}}{2} \right) \right)^D \prod_d \left(\frac{\widetilde{r}_{jkd}}{2} \right)^{\frac{\widetilde{\phi}^r_{jk}}{2}}} \right).
\end{aligned}
\tag{7.126}
$$

From Eq. (7.126), $\mathcal{F}^M_{\Theta}$ can be calculated by using the statistics of the posterior distribution parameters $\widetilde{\Phi}$ given in Eq. (7.66). This part is equivalent to the objective function for model selection based on Akaike's Bayesian information criterion (Akaike 1980). The whole $\mathcal{F}^M$ for all categories is obtained by simply summing up the $\mathcal{F}^M$ results obtained in this section for all categories as in Eq. (7.36).

Now we focus on $\mathcal{F}^M_{S,V}$. From the definition in Eq. (7.117), $-\mathcal{F}^M_{S,V}$ can be represented as follows:

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

$$\begin{aligned} -\mathcal{F}^{M}_{S,V} &= -\mathbb{E}_{(S,V)}\left[\log \widetilde{q}(S,V|\mathbf{O},M)\right] \\ &= -\sum_{S,V} \widetilde{q}(S,V|\mathbf{O},M)\log \widetilde{q}(S,V|\mathbf{O},M). \end{aligned} \tag{7.127}$$

Therefore, $\mathcal{F}^{M}_{S,V}$ denotes the entropy of the posterior distribution $\widetilde{q}(S,V|\mathbf{O},M)$. As we discussed in Section 7.69, it is difficult to obtain the analytical form of $\widetilde{q}(S,V|\mathbf{O},M)$ due to the normalization constant, and direct calculation of the above entropy is also difficult. Instead, we focus on the variational complete data likelihood form $\widetilde{q}(\mathbf{O},S,V|M)$, which is obtained by the Bayes theorem as follows:

$$\widetilde{q}(S,V|\mathbf{O},M) = \frac{\widetilde{q}(\mathbf{O},S,V|M)}{\sum_{S,V}\widetilde{q}(\mathbf{O},S,V|M)}. \tag{7.128}$$

Based on the discussion in Eq. (7.73), $\widetilde{q}(\mathbf{O},S,V|M)$ is represented with the unnormalized function $\widetilde{u}(\mathbf{O},S,V|M)$ as follows:

$$\begin{aligned} \widetilde{q}(\mathbf{O},S,V|M) &= C\widetilde{u}(\mathbf{O},S,V|M) \\ &= C\prod_{t=1}^{T} \tilde{a}_{s_{t-1}s_t}\tilde{\omega}_{s_t v_t}\tilde{b}_{s_t v_t}(\mathbf{o}_t), \end{aligned} \tag{7.129}$$

where C is a normalization constant, defined as follows:

$$C = \int \sum_{S,V} \widetilde{u}(\mathbf{O},S,V|M)d\mathbf{O}. \tag{7.130}$$

$\tilde{a}_{ij}$, $\tilde{\omega}_{jk}$, $\tilde{b}_{jk}(\mathbf{o}_t)$ are analytically calculated in the VB-E step (Eq. (7.101)). Therefore, by substituting (7.129) into (7.128), the normalization constant is canceled out, and we can obtain the following equation:

$$\begin{aligned} \widetilde{q}(S,V|\mathbf{O},M) &= \frac{C\widetilde{u}(\mathbf{O},S,V|M)}{\sum_{S',V'} C\widetilde{u}(\mathbf{O},S',V'|M)} \\ &= \frac{\widetilde{u}(\mathbf{O},S,V|M)}{\sum_{S',V'} \widetilde{u}(\mathbf{O},S',V'|M)}. \end{aligned} \tag{7.131}$$

Therefore, $\mathcal{F}^{M}_{S,V}$ can be rewritten as follows:

$$\begin{aligned} \mathcal{F}^{M}_{S,V} &= \sum_{S,V} \widetilde{q}(S,V|\mathbf{O},M)\log \widetilde{q}(S,V|\mathbf{O},M) \\ &= \sum_{S,V} \widetilde{q}(S,V|\mathbf{O},M)\log\left(\widetilde{u}(\mathbf{O},S,V|M)\right) \\ &\quad - \log\left(\sum_{S,V} \widetilde{u}(\mathbf{O},S,V|M)\right). \end{aligned} \tag{7.132}$$

Note that the second term corresponds to the summation of all possible S and V for unnormalized function $\widetilde{u}(\mathbf{O},S,V|M)$, which can be computed in the VB forward algorithm in Eq. (7.109) as follows:

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

$$\sum_{S,V} \widetilde{u}(\mathbf{O}, S, V|M) = \sum_{j=1}^{J} \tilde{\tilde{\alpha}}_T(j). \tag{7.133}$$

Now we focus on the first term in Eq. (7.132). From the discussions in Section 3.4.2, we can convert summation over sequences S, V to a summation over HMM states i and j, and mixture component k in this term. Therefore, this term is represented as follows:

$$\begin{aligned}
&\sum_{S,V} \widetilde{q}(S, V|\mathbf{O}, M) \log\left(\widetilde{u}(\mathbf{O}, S, V|M)\right) \\
&= \sum_{S,V} \widetilde{q}(S, V|\mathbf{O}, M) \sum_{t=1}^{T} \left(\log \tilde{a}_{s_{t-1}s_t} + \log \tilde{\omega}_{s_t v_t} + \log \tilde{b}_{s_t v_t}(\mathbf{o}_t)\right) \\
&= \sum_{i,j,t} \tilde{\xi}_t(i,j) \log \tilde{a}_{ij} + \sum_{j,k,t} \tilde{\gamma}_t(j,k) \left(\log \tilde{\omega}_{jk} + \log \tilde{b}_{jk}(\mathbf{o}_t)\right).
\end{aligned} \tag{7.134}$$

Thus, we obtain the term without computing the summation over S and V.

Finally, we summarize calculation of $\mathcal{F}^M_{S,V}$ by using the definitions of $\tilde{a}_{ij}$, $\tilde{\omega}_{jk}$, $\tilde{b}_{jk}(\mathbf{o}_t)$ in Eq. (7.101), as follows:

$$\begin{aligned}
\mathcal{F}^M_{S,V} = &\sum_{i,j} \widetilde{\xi}_{ij} \left(\Psi\left(\widetilde{\phi}^a_{ij}\right) - \Psi\left(\textstyle\sum_{j'} \widetilde{\phi}^a_{ij'}\right)\right) + \sum_{j,k} \widetilde{\gamma}_{jk} \left(\Psi\left(\widetilde{\phi}^\omega_{jk}\right) - \Psi\left(\textstyle\sum_{k'} \widetilde{\phi}^\omega_{jk'}\right)\right) \\
&- \frac{1}{2} \sum_{j,k} \widetilde{\gamma}_{jk} \left(D\left(\log(2\pi) + \frac{1}{\widetilde{\phi}^\mu_{jk}} - \Psi\left(\frac{\widetilde{\phi}^r_{jk}}{2}\right)\right) + \sum_d \log \frac{\widetilde{r}_{jkd}}{2}\right) \\
&- \frac{1}{2} \sum_{j,k} \left(\widetilde{\phi}^r_{jk} \sum_{t,d} \frac{\widetilde{\gamma}_t(j,k)(o_{td} - \widetilde{\mu}_{jkd})^2}{\widetilde{r}_{jkd}}\right) - \log\left(\sum_j \tilde{\tilde{\alpha}}_T(j)\right).
\end{aligned} \tag{7.135}$$

Thus, we also obtain the analytical result for $\mathcal{F}^M_{S,V}$, which corresponds to the latent variable effect for the variational lower bound.

The analytical result for the variational lower bound $\mathcal{F}^M$ for the CDHMM is determined using $\mathcal{F}^M_\Theta$ in Eq. (7.126) and $\mathcal{F}^M_{S,V}$ in Eq. (7.135). Although the analytical result looks complicated, all variables are already computed in the VB expectation and maximization steps. We also want to emphasize that the computation is quite feasible since it is carried out without a summation over all possible latent variable sequences S and V. The variational lower bound is derived analytically so that it retains the effects of the dependence between model parameters and of the latent variables, defined in the generative model distribution in Eq. (7.53), unlike the conventional Bayesian information criterion and minimum description length (BIC/MDL) approaches, as discussed in Section 6.5. Therefore, the variational lower bound can compare any acoustic models with respect to all topological aspects and their combinations, e.g., contextual and temporal topologies in HMMs, the number of components per GMM in an HMM state, and the dimensional size of feature vectors, based on the following equation:

$$\widetilde{M} = \arg \max_{M \in (\mathbb{T} \times \mathbb{S} \times \mathbb{G} \times \mathbb{D})} \mathcal{F}^M. \tag{7.136}$$

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

Here $\mathbb{T}$, $\mathbb{S}$, $\mathbb{G}$, and $\mathbb{D}$ denote search spaces of HMM-temporal, HMM-contextual, GMM and feature vector topologies, respectively.

Based on the discussion in Section 7.3, the seven steps in Algorithm 11 provide a VB training algorithm for acoustic modeling. Here, τ denotes an iteration count, and ε denotes a threshold that checks whether $\mathcal{F}^M$ converges. Thus, the posterior distribution estimation in the VB framework can be effectively constructed based on the VB Baum–Welch algorithm, which is analogous to the ML Baum–Welch algorithm (Algorithm 4). In addition, VB can realize the model selection using the VB objective function as shown in **Step 9**. Thus, VB can construct an acoustic model consistently based on the Bayesian approach.

Algorithm 11 Variational Bayesian Baum–Welch algorithm for CDHMMs with model selection

```
Require: Set posterior parameter $\tilde{\Phi}[\tau = 0]$ from initialized transition probability
    $\tilde{\xi}[\tau = 0]$, occupation probability $\tilde{\gamma}[\tau = 0]$, and model structure $M$ (prior parameter
    $\Phi^0$ is included) for each category
1: repeat
2:     Compute $\tilde{a}[\tau + 1]$, $\tilde{\omega}[\tau + 1]$, and $\tilde{b}(\mathbf{O})[\tau + 1]$ using $\tilde{\Phi}[\tau]$. (By Eq. (7.101))
3:     Update $\tilde{\xi}[\tau + 1]$ and $\tilde{\gamma}[\tau + 1]$ via the Viterbi algorithm or forward–backward
       algorithm. (By Eqs. (7.114) and (7.115))
4:     Accumulate the sufficient statistics $\tilde{\xi}[\tau + 1]$, $\tilde{\gamma}[\tau + 1]\tilde{\gamma}^{(1)}[\tau + 1]$, $\tilde{\gamma}^{(2)}[\tau + 1]$ (by
       Eq. (7.67)
5:     Compute $\tilde{\Phi}[\tau + 1]$ using $\tilde{\Xi}[\tau + 1]$ and $\Phi^0$. (By Eq. (7.66))
6:     Calculate total $\mathcal{F}^M[\tau + 1]$ for all categories. (By using Eqs. (7.126) and (7.135)
       and summing up all categories' $\mathcal{F}^M$)
7:     Calculate $\Delta = |(\mathcal{F}^M[\tau + 1] - \mathcal{F}^M[\tau])/\mathcal{F}^M[\tau + 1]|$, $\tau \leftarrow \tau + 1$
8: until $\Delta \leq \varepsilon$
   Calculate $\mathcal{F}^M$ for all possible $M$ and find $\tilde{M}(= \arg\max_M \mathcal{F}^M)$
```

Note that if we change $\tilde{\ } \rightarrow \hat{\ }$ (a value with $\hat{\ }$ attached indicates an ML estimate), $\Phi \rightarrow \Theta$ and $\mathcal{F}^M \rightarrow \text{Ł}^M$ (where Ł^M means the log-likelihood for a model M), this algorithm becomes an ML-based framework, except for the model selection. Therefore, in the implementation phase, the VB framework can be realized in the conventional systems of acoustic model construction by adding the prior distribution setting and by changing the estimation procedure and objective function calculation.

7.3.5 VB posterior for Bayesian predictive classification

This subsection deals with the Bayes decision rule based on the VB approach. It is related to the Bayesian predictive classification, as discussed in Section 6.3.1 with the Laplace approximation, but this section deals with the same issue with VB. In this

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

section, we use the following notation to clearly distinguish the training and recognition data, similar to Section 6.3.1:

$$\begin{aligned} \mathbf{O} &: \text{future data}, \\ \mathcal{O} &: \text{training data}. \end{aligned} \tag{7.137}$$

In recognition, $\mathbf{O} = \{\mathbf{O}_t \in \mathbb{R}^D | t = 1, \cdots, T\}$ denotes the feature vector sequence of input speech, and $S = \{s_t \in \{1, \cdots, J\} | t = 1, \cdots, T\}$ denotes the corresponding HMM state sequence. Although our target application is ASR, which outputs the word sequence W, this section simplifies the decoding rule for the explanation. That is, the decoding needs to handle the word sequence W in addition to the state sequence S, but it can be combined with the LVCSR decoder if we can build the Viterbi algorithm. Therefore, this section focuses on formulating the Viterbi algorithm within the VB framework, similarly to that within the ML framework, as discussed in Section 3.3.2.

The Viterbi algorithm can achieve the optimal state sequence $\widetilde{S}$ by using a conditional probability function $p(S|\mathbf{O}, \mathcal{O})$ given input data $\mathbf{O}$ and training data $\mathcal{O}$, as follows:

$$\begin{aligned} \bar{S} = \arg\max_S p(S|\mathbf{O}, \mathcal{O}) &= \arg\max_S \frac{p(\mathbf{O}, S|\mathcal{O})}{p(\mathbf{O}|\mathcal{O})} \\ &= \arg\max_S p(\mathbf{O}, S|\mathcal{O}). \end{aligned} \tag{7.138}$$

$p(\mathbf{O}, S|\mathcal{O})$ is a variant of *predictive distribution* (Berger 1985, Bernardo & Smith 2009), because this distribution predicts the probability of unknown data $\mathbf{O}$ conditioned by training data $\mathcal{O}$. Note that Eq. (7.138) does not depend on parameters Θ and model M, and these can be explicitly involved by considering the following sum rule:

$$p(\mathbf{O}, S|\mathcal{O}) = \sum_M \int p(\mathbf{O}, S|\Theta, \mathcal{O}, M) p(\Theta|\mathcal{O}, M) p(M|\mathcal{O}) d\Theta. \tag{7.139}$$

This predictive distribution based approach involves considering the integrals and true posterior distributions, an approach which is also applied to speech recognition (Huo & Lee 2000, Jiang *et al.* 1999, Lee & Huo 2000, Chien & Liao 2001), as discussed in Section 6.3.1, with the Laplace approximation.

After VB-based acoustic modeling in Section 7.3.4, an appropriate model structure $\widetilde{M}$ is selected based on the VB model selection Eq. (7.136), and the optimal VB posterior distributions are obtained $\widetilde{q}(\Theta|\mathcal{O}, \widetilde{M})$. Therefore, the true posterior distributions can be approximated by the VB posteriors, and Eq. (7.139) is approximated as:

$$\begin{aligned} p(\mathbf{O}, S|\mathcal{O}) &\approx \sum_M \int p(\mathbf{O}, S|\Theta, M) \widetilde{q}(\Theta|\mathcal{O}, M) \delta(M, \widetilde{M}) d\Theta \\ &= \int p(\mathbf{O}, S|\Theta, \widetilde{M}) \widetilde{q}(\Theta|\mathcal{O}, \widetilde{M}) d\Theta \\ &= \mathbb{E}_{(\Theta)} \left[p(\mathbf{O}, S|\Theta, \widetilde{M}) \right]. \end{aligned} \tag{7.140}$$

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

Thus, we can build the Viterbi algorithm for the expectation over the model parameter Θ by using the VB posterior. In the following section we omit the models structure index $\widetilde{M}$ for simplicity.

Similarly to Section 3.3.2, we first define the following *expected* highest probability along a single path, at time t, which accounts for the first t observations and ends in state j:

$$\widetilde{\delta}_t(j) \triangleq \max_{s_1,\cdots,s_{t-1}} \mathbb{E}_{(\Theta)}\left[p(s_1,\cdots,s_t=j,\mathbf{o}_1,\cdots,\mathbf{o}_t|\Theta)\right]. \tag{7.141}$$

By using $\widetilde{\delta}_t(j)$ recursively, we can obtain the most probable state sequence as follows:

- Initialization

$$\begin{aligned} &\widetilde{\delta}_1(i) = \mathbb{E}_{(\boldsymbol{\pi})}\left[\pi_i\right]\sum_{k=1}^{K}\mathbb{E}_{(\boldsymbol{\omega})}\left[\omega_{ik}\right]\mathbb{E}_{(\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[\mathcal{N}(\mathbf{o}_1|\boldsymbol{\mu}_{ik},\boldsymbol{\Sigma}_{ik})\right], \\ &\psi_1(i) = 0, \qquad 1 \le i \le J. \end{aligned} \tag{7.142}$$

- Recursion

$$\begin{aligned} \widetilde{\delta}_t(j) &= \left(\max_{1\le i\le J}\widetilde{\delta}_{t-1}(i)\mathbb{E}_{(\mathbf{A})}\left[a_{ij}\right]\right)\sum_{k=1}^{K}\mathbb{E}_{(\boldsymbol{\omega})}\left[\omega_{ik}\right]\mathbb{E}_{(\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[\mathcal{N}(\mathbf{o}_1|\boldsymbol{\mu}_{jk},\boldsymbol{\Sigma}_{jk})\right], \\ \psi_t(j) &= \left(\arg\max_{1\le i\le J}\widetilde{\delta}_{t-1}(i)\mathbb{E}_{(\mathbf{A})}\left[a_{ij}\right]\right), \qquad \begin{matrix} 2 \le t \le T \\ 1 \le j \le J. \end{matrix} \end{aligned} \tag{7.143}$$

- Termination

$$\begin{aligned} p(\widetilde{S},\mathbf{O}|\mathcal{O}) &= \max_{1\le j\le J}\widetilde{\delta}_T(i), \\ \widetilde{s}_T &= \arg\max_{1\le j\le J}\widetilde{\delta}_T(i). \end{aligned} \tag{7.144}$$

- State sequence backtracking

$$\widetilde{s}_t = \psi_{t+1}(\widetilde{s}_{t+1}), \qquad t = T-1, T-2, \cdots, 1. \tag{7.145}$$

Thus, we can perform the Viterbi algorithm for the predictive distribution based on the VB posteriors. To realize the Viterbi algorithm, we need to consider the following expectation:

$$\begin{aligned} &\mathbb{E}_{(\mathbf{A})}\left[a_{ij}\right], \\ &\mathbb{E}_{(\boldsymbol{\omega})}\left[\omega_{ik}\right], \\ &\mathbb{E}_{(\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[\mathcal{N}(\mathbf{o}_t|\boldsymbol{\mu}_{jk},\boldsymbol{\Sigma}_{jk})\right]. \end{aligned} \tag{7.146}$$

We provide the solution for each of the expected variables. Note that these are different from the expected variables of the CDHMM parameters in the VB E-step, as discussed in Eq. (7.73),[4] i.e.,

[4] Again we omit the initial transition parameters for simplicity.

$$\begin{aligned}\widetilde{a}_{ij} &= \exp\left(\mathbb{E}_{(a_{ij})}\left[\log(a_{ij})\right]\right),\\ \widetilde{\omega}_{jk} &= \exp\left(\mathbb{E}_{(\omega_{jk})}\left[\log(\omega_{jk})\right]\right),\\ \widetilde{b}_{jk}(\mathbf{o}_t) &= \exp\left(\mathbb{E}_{(\boldsymbol{\mu}_{jk},\boldsymbol{\Sigma}_{jk})}\left[\log(\mathcal{N}(\mathbf{o}_t|\boldsymbol{\mu}_{jk},\boldsymbol{\Sigma}_{jk}))\right]\right).\end{aligned} \tag{7.147}$$

This is because the VB E-step is a training step, and we need to optimize these values based on the variational method, which necessitates considering the expectation of the parameters in the logarithmic domain. In the prediction case, since we already have the posterior distributions in the training step, Eq. (7.146) simply performs the expectation determination for the CDHMM parameters directly.

Expected state transition $\mathbb{E}_{(\mathbf{A})}\left[a_{ij}\right]$

We first focus on calculation of the expected state transition $\widetilde{a}_{ij}$. Although this can be obtained as the mean result of the Dirichlet distribution in Appendix C.4, we provide the derivation for its educational value. Based on the definition of the Dirichlet distribution in Eq. (7.62), we can obtain the following equation:

$$\begin{aligned}\mathbb{E}_{(\mathbf{A})}\left[a_{ij}\right] &= \int a_{ij}\mathrm{Dir}(\{a_{ij'}\}_{j'=1}^{J}|\{\widetilde{\phi}_{ij'}^{a}\}_{j'=1}^{J})\prod_{j'=1}^{J} da_{ij'}\\ &= C_{\mathrm{Dir}}(\{\widetilde{\phi}_{ij}^{a}\}_{j=1}^{J})\int a_{ij}\prod_{j'=1}^{J}(a_{ij'})^{\widetilde{\phi}_{ij'}^{a}-1}da_{ij'}.\end{aligned} \tag{7.148}$$

Now we define the following variable:

$$\hat{\phi}_{ij'}^{a} \triangleq \begin{cases}\widetilde{\phi}_{ij'}^{a}+1 & j'=j\\ \widetilde{\phi}_{ij'}^{a} & j'\neq j.\end{cases} \tag{7.149}$$

By using $\hat{\phi}_{ij'}^{a}$, the integral in Eq. (7.148) is solved as:

$$\begin{aligned}\mathbb{E}_{(\mathbf{A})}\left[a_{ij}\right] &= C_{\mathrm{Dir}}(\{\widetilde{\phi}_{ij}^{a}\}_{j=1}^{J})\int \prod_{j'=1}^{J}(a_{ij'})^{\hat{\phi}_{ij'}^{a}-1}da_{ij'}\\ &= \frac{C_{\mathrm{Dir}}(\{\widetilde{\phi}_{ij}^{a}\}_{j=1}^{J})}{C_{\mathrm{Dir}}(\{\hat{\phi}_{ij}^{a}\}_{j=1}^{J})}.\end{aligned} \tag{7.150}$$

The normalization constant of the Dirichlet distribution is defined (Appendix C.4) as

$$C_{\mathrm{Dir}}(\{\phi_j\}_{j=1}^{J}) \triangleq \frac{\Gamma(\sum_{j=1}^{J}\phi_j)}{\prod_{j=1}^{J}\Gamma(\phi_j)}. \tag{7.151}$$

Therefore, by substituting the concrete form of the normalization constant into Eq. (7.150) and by using the definition of $\hat{\phi}_{ij'}^{a}$, Eq. (7.150) can be represented as follows:

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

$$\begin{aligned}
\mathbb{E}_{(\mathbf{A})}\left[a_{ij}\right] &= \frac{\prod_{j'=1}^{J}\Gamma(\hat{\phi}^{a}_{ij'})}{\prod_{j'=1}^{J}\Gamma(\widetilde{\phi}^{a}_{ij'})}\frac{\Gamma(\sum_{j'=1}^{J}\widetilde{\phi}^{a}_{ij'})}{\Gamma(\sum_{j'=1}^{J}\hat{\phi}^{a}_{ij'})} \\
&= \frac{\prod_{j'\neq j}^{J}\Gamma(\widetilde{\phi}^{a}_{ij'})}{\prod_{j'\neq j}^{J}\Gamma(\widetilde{\phi}^{a}_{ij'})}\frac{\Gamma(\widetilde{\phi}^{a}_{ij}+1)}{\Gamma(\widetilde{\phi}^{a}_{ij})}\frac{\Gamma(\sum_{j'=1}^{J}\widetilde{\phi}^{a}_{ij'})}{\Gamma(1+\sum_{j'=1}^{J}\widetilde{\phi}^{a}_{ij'})} \\
&= \frac{\Gamma(\widetilde{\phi}^{a}_{ij}+1)}{\Gamma(\widetilde{\phi}^{a}_{ij})}\frac{\Gamma(\sum_{j'=1}^{J}\widetilde{\phi}^{a}_{ij'})}{\Gamma(1+\sum_{j'=1}^{J}\widetilde{\phi}^{a}_{ij'})}.
\end{aligned} \tag{7.152}$$

Finally, we use the following formula for the gamma function:

$$\Gamma(x+1) = x\Gamma(x). \tag{7.153}$$

Then, Eq. (7.152) is analytically obtained as the following simple equation:

$$\begin{aligned}
\mathbb{E}_{(\mathbf{A})}\left[a_{ij}\right] &= \frac{\widetilde{\phi}^{a}_{ij}\Gamma(\widetilde{\phi}^{a}_{ij})}{\Gamma(\widetilde{\phi}^{a}_{ij})}\frac{\Gamma(\sum_{j'=1}^{J}\widetilde{\phi}^{a}_{ij'})}{\sum_{j'=1}^{J}\widetilde{\phi}^{a}_{ij'}\Gamma(\sum_{j'=1}^{J}\widetilde{\phi}^{a}_{ij'})} \\
&= \frac{\widetilde{\phi}^{a}_{ij}}{\sum_{j'=1}^{J}\widetilde{\phi}^{a}_{ij'}}.
\end{aligned} \tag{7.154}$$

Note that the state transition probability is obtained from the normalized weight, which is proportional to the posterior hyperparameter $\widetilde{\phi}^{a}_{ij}$, and the result is very intuitive.

Expected mixture weight $\mathbb{E}_{(\omega)}\left[\omega_{jk}\right]$

Since the mixture weight ω_{jk} is represented by a multinomial distribution, it is similar to the state transition a_{ij}. Similarly to $\mathbb{E}_{(\mathbf{A})}\left[a_{ij}\right]$, the expected state transition $\widetilde{\omega}_{jk}$ is calculated as follows:

$$\mathbb{E}_{(\omega)}\left[\omega_{jk}\right] = \frac{\widetilde{\phi}^{\omega}_{jk}}{\sum_{k'=1}^{K}\widetilde{\phi}^{\omega}_{jk'}}. \tag{7.155}$$

Again, the mixture weight probability is obtained using the normalized weight of the posterior hyperparameter $\widetilde{\phi}^{\omega}_{jk}$.

Expected Gaussian distribution $\mathbb{E}_{(\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[\mathcal{N}(\mathbf{o}_t|\boldsymbol{\mu}_{jk},\boldsymbol{\Sigma}_{jk})\right]$

Finally, we calculate the expected value of the Gaussian distribution with VB posteriors for Gaussian parameters. First the expectation is factorized for each dimension when we use the diagonal covariance as follows:

$$\mathbb{E}_{(\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[\mathcal{N}(\mathbf{o}_t|\boldsymbol{\mu}_{jk},\boldsymbol{\Sigma}_{jk})\right] = \prod_{d=1}^{D}\mathbb{E}_{(\mu_{jkd},r_{jkd})}\left[\mathcal{N}(o_{td}|\mu_{jkd},(r_{jkd})^{-1})\right]. \tag{7.156}$$

The indexes of state ij, mixture component k, frame t, and dimension d are removed to simplify the derivation.

Based on the definition of the Gaussian and gamma distributions in Eq. (7.62), we can obtain the following equation:

$$
\begin{aligned}
&\mathbb{E}_{(\mu,r)}\left[\mathcal{N}(o|\mu,r^{-1})\right] \\
&\quad = \int \mathcal{N}(\mu|\widetilde{\mu},(\widetilde{\phi}^{\mu}r)^{-1})\mathrm{Gam}_2(r|\widetilde{\phi}^{r},\widetilde{r})\mathcal{N}(o|\mu,r^{-1})d\mu dr \\
&\quad = C_{\mathcal{N}}(\widetilde{\phi}^{\mu})C_{\mathrm{Gam}_2}(\widetilde{\phi}^{r},\widetilde{r})C_{\mathcal{N}} \\
&\quad \times \int r^{\frac{1}{2}}\exp\left(-\frac{\widetilde{\phi}^{\mu}r(\mu-\widetilde{\mu})^2}{2}\right)r^{\frac{\widetilde{\phi}^{r}}{2}-1}\exp\left(-\frac{\widetilde{r}r}{2}\right)r^{\frac{1}{2}}\exp\left(-\frac{r(o-\mu)^2}{2}\right)d\mu dr \\
&\quad \propto \int r^{\frac{\widetilde{\phi}^{r}}{2}}\exp\left(-\frac{\widetilde{r}r}{2}\right)\exp\left(-\frac{\widetilde{\phi}^{\mu}r(\mu-\widetilde{\mu})^2}{2}\right)\exp\left(-\frac{r(o-\mu)^2}{2}\right)d\mu dr.
\end{aligned}
\tag{7.157}
$$

First, we focus on the integration with respect to μ, and completing the square with respect to μ. Then, by integrating with respect to μ, and arranging the equation, the following equation is obtained:

$$
\begin{aligned}
&\int \exp\left(-\frac{r}{2}\left((o-\mu)^2+\widetilde{\phi}^{\mu}(\mu-\widetilde{\mu})^2\right)\right)d\mu \\
&\quad = \int \exp\left(-\frac{r}{2}\left((1+\widetilde{\phi}^{\mu})\left(\mu-\frac{o+\widetilde{\phi}^{\mu}\widetilde{\mu}}{1+\widetilde{\phi}^{\mu}}\right)^2-\frac{(o+\widetilde{\phi}^{\mu}\widetilde{\mu})^2}{1+\widetilde{\phi}^{\mu}}+o^2+\widetilde{\phi}^{\mu}\widetilde{\mu}^2\right)\right)d\mu \\
&\quad \propto r^{-\frac{1}{2}}\exp\left(-\frac{r}{2}\left(-\frac{(o+\widetilde{\phi}^{\mu}\widetilde{\mu})^2}{1+\widetilde{\phi}^{\mu}}+o^2+\widetilde{\phi}^{\mu}\widetilde{\mu}^2\right)\right) \\
&\quad = r^{-\frac{1}{2}}\exp\left(-\frac{r}{2(1+\widetilde{\phi}^{\mu})}\left(-(o+\widetilde{\phi}^{\mu}\widetilde{\mu})^2+(1+\widetilde{\phi}^{\mu})(o^2+\widetilde{\phi}^{\mu}\widetilde{\mu}^2)\right)\right) \\
&\quad = r^{-\frac{1}{2}}\exp\left(-r\frac{\widetilde{\phi}^{\mu}(o-\widetilde{\mu})^2}{2(1+\widetilde{\phi}^{\mu})}\right).
\end{aligned}
\tag{7.158}
$$

Here we discuss the case when the VB posterior for r is the Dirac delta function around the MAP value of r, and the argument of its Dirac delta function is the maximum value of the VB posterior. Then, the result of the integration with respect to r is obtained by changing r to the MAP value $(\widetilde{\phi}^{r}-2)\widetilde{r}^{-1}$ in Eq. (7.158) in Appendix C.11. Therefore, the following equation is obtained:

$$
\begin{aligned}
\mathbb{E}_{(\mu)}\left[\mathcal{N}(o|\mu,r^{-1})\right] &\propto \exp\left(-\frac{\widetilde{\phi}^{r}-2}{\widetilde{r}}\frac{\widetilde{\phi}^{\mu}(o-\widetilde{\mu})^2}{2(1+\widetilde{\phi}^{\mu})}\right) \\
&= \mathcal{N}\left(o\middle|\widetilde{\mu},\frac{1+\widetilde{\phi}^{\mu}}{(\widetilde{\phi}^{r}-2)\widetilde{\phi}^{\mu}}\widetilde{r}\right).
\end{aligned}
\tag{7.159}
$$

Thus, by recovering the omitted indexes, we can obtain

$$
\mathbb{E}_{(\boldsymbol{\mu})}\left[\mathcal{N}(\mathbf{o}_t|\boldsymbol{\mu},\boldsymbol{\Sigma})\right] = \prod_{d=1}^{D}\mathcal{N}\left(o_{td}\middle|\widetilde{\mu}_{jkd},\frac{1+\widetilde{\phi}^{\mu}_{jk}}{(\widetilde{\phi}^{r}_{jk}-2)\widetilde{\phi}^{\mu}_{jk}}\widetilde{r}_{jkd}\right).
\tag{7.160}
$$

This is the analytical result of the expected function of a Gaussian distribution with expectation only over the mean parameter μ.

By substituting Eq. (7.158) into Eq. (7.157), we can obtain the following integral:

$$\int r^{\frac{\widetilde{\phi}^r+1}{2}-1} \exp\left(-r\frac{\widetilde{\phi}^\mu(o-\widetilde{\mu})^2+(1+\widetilde{\phi}^\mu)\widetilde{r}}{2(1+\widetilde{\phi}^\mu)}\right)dr. \tag{7.161}$$

First we use the following notation to simplify the integral:

$$\begin{aligned}\alpha &\triangleq \frac{\widetilde{\phi}^r+1}{2},\\ \beta &\triangleq \frac{\widetilde{\phi}^\mu(o-\widetilde{\mu})^2+(1+\widetilde{\phi}^\mu)\widetilde{r}}{2(1+\widetilde{\phi}^\mu)}.\end{aligned} \tag{7.162}$$

Note that β depends on the observation o. Then, the integral with the explicit range of r is rewritten as follows:

$$\int_0^\infty r^{\alpha-1}e^{-\beta r}dr. \tag{7.163}$$

Now, we convert r with the following variable x:

$$\begin{aligned}r &= \frac{x}{\beta},\\ dr &= \frac{1}{\beta}dx,\\ r \in [0,\infty] &\rightarrow x \in [0,\infty].\end{aligned} \tag{7.164}$$

Now $\widetilde{\phi}^\mu$ is a hyperparameter of the Dirichlet distribution, and $\widetilde{\phi}^\mu > 0$, therefore $\beta > 0$ from Eq. (7.162), and the range of x becomes $[0,\infty]$. Then, the integral can be rewritten as:

$$\begin{aligned}\int_0^\infty r^{\alpha-1}e^{-\beta r}dr &= \int_0^\infty \left(\frac{x}{\beta}\right)^{\alpha-1} e^{-x}\frac{1}{\beta}dx\\ &= \left(\frac{1}{\beta}\right)^\alpha \int_0^\infty x^{\alpha-1}e^{-x}dx.\end{aligned} \tag{7.165}$$

Here, from the formula of the gamma function, we can further rewrite the above integral as:

$$\begin{aligned}\int_0^\infty r^{\alpha-1}e^{-\beta r}dr &= \left(\frac{1}{\beta}\right)^\alpha \Gamma(\alpha) = \left(\frac{1}{\beta}\right)^\alpha (\alpha-1)!\\ &\propto \left(\frac{1}{\beta}\right)^\alpha.\end{aligned} \tag{7.166}$$

Since $(\alpha-1)!$ does not depend on the observation o, we can disregard it as a constant value. Finally, by recovering the variables of α and β from Eq. (7.162), Eq. (7.161) is obtained as the following equation:

$$\begin{aligned}&\int r^{\frac{\widetilde{\phi}^r+1}{2}-1} \exp\left(-r\frac{\widetilde{\phi}^\mu(o-\widetilde{\mu})^2+(1+\widetilde{\phi}^\mu)\widetilde{r}}{2(1+\widetilde{\phi}^\mu)}\right)dr\\ &\propto \left(\frac{\widetilde{\phi}^\mu(o-\widetilde{\mu})^2+(1+\widetilde{\phi}^\mu)\widetilde{r}}{2(1+\widetilde{\phi}^\mu)}\right)^{-\frac{\widetilde{\phi}^r+1}{2}}\\ &\propto \left(1+\frac{\widetilde{\phi}^\mu}{(1+\widetilde{\phi}^\mu)\widetilde{r}}(o-\widetilde{\mu})^2\right)^{-\frac{\widetilde{\phi}^r+1}{2}}.\end{aligned} \tag{7.167}$$

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

Here we refer to the concrete form of the Student's *t*-distribution given in Appendix C.16:

$$\mathrm{St}(x|\mu,\lambda,\kappa) \triangleq C_{\mathrm{St}}\left(1+\frac{1}{\kappa\lambda}(x-\mu)^2\right)^{-\frac{\kappa+1}{2}}. \tag{7.168}$$

The parameters μ, κ, and λ of the Student's *t*-distribution correspond to those of the above equation as follows:

$$\begin{cases} \mu &= \widetilde{\mu}, \\ \lambda &= \frac{(1+\widetilde{\phi}^{\mu})\widetilde{r}}{\widetilde{\phi}^{\mu}\widetilde{\phi}^{r}}, \\ \kappa &= \widetilde{\phi}^{r}. \end{cases} \tag{7.169}$$

Thus, the result of the integral with respect to μ and r (Eq. (7.157)) is represented as the Student's *t*-distribution:

$$\mathrm{St}\left(o\left|\widetilde{\mu}, \frac{(1+\widetilde{\phi}^{\mu})\widetilde{r}}{\widetilde{\phi}^{\mu}\widetilde{\phi}^{r}}, \widetilde{\phi}^{r}\right.\right). \tag{7.170}$$

The third parameter in the *t*-distribution is called the degree of freedom, and if this value is large, the distribution approaches the Gaussian distribution theoretically.

Thus, by recovering the omitted indexes, we can obtain

$$\mathbb{E}_{(\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[\mathcal{N}(\mathbf{o}_t|\boldsymbol{\mu},\boldsymbol{\Sigma})\right] = \prod_{d=1}^{D}\mathrm{St}\left(o_{td}\left|\widetilde{\mu}_{jkd}, \frac{(1+\widetilde{\phi}^{\mu}_{jk})\widetilde{r}_{jkd}}{\widetilde{\phi}^{\mu}_{jk}\widetilde{\phi}^{r}_{jk}}, \widetilde{\phi}^{r}_{jk}\right.\right). \tag{7.171}$$

This is the analytical result of the expected Gaussian distribution with marginalization of both mean and precision parameters μ and r. Compared with Eq. (7.160), the marginalization of both parameters changes the distribution from the Gaussian distribution to the Student's *t*-distribution. The latter is called a long tail distribution since it is a power law function, and it provides a robust classification in general, when the amount of training data is small.

Since the degree of freedom in this solution is the posterior hyperparameter of the precision parameter $\widetilde{\phi}^{r}$, and it is proportional to the amount of data, as shown in Eq. (7.67), this solution approaches the Gaussian distribution. Then, the variance parameters in Eq. (7.171) also approximately approach the following value:

$$\frac{(1+\widetilde{\phi}^{\mu}_{jk})\widetilde{r}_{jkd}}{\widetilde{\phi}^{\mu}_{jk}\widetilde{\phi}^{r}_{jk}} \approx \frac{\widetilde{r}_{jkd}}{\widetilde{\phi}^{r}_{jk}}. \tag{7.172}$$

Thus, Eq. (7.171) is approximated by the following Gaussian distribution when the amount of data $\mathbf{O}$ is large:

$$\mathbb{E}_{(\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[\mathcal{N}(\mathbf{o}_t|\boldsymbol{\mu},\boldsymbol{\Sigma})\right] \approx \prod_{d=1}^{D}\mathcal{N}\left(o_{td}\left|\widetilde{\mu}_{jkd}, \frac{\widetilde{r}_{jkd}}{\widetilde{\phi}^{r}_{jk}}\right.\right). \tag{7.173}$$

This solution corresponds to the MAP estimation result of the Gaussian distribution in CDHMM, as discussed in Section 4.3.5.

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

In Watanabe & Nakamura (2006), experimental results were reported to show the effectiveness of the Bayesian predictive classification without marginalization (it corresponds to the MAP estimation in Section 4.3), with marginalization of only mean parameters (corresponds to Eq. (7.160)), and with marginalization of both mean and covariance parameters (corresponds to Eq. (7.171)). Speaker adaptation experiments for LVCSR (30 000 vocabulary size) show that the Student's t-distribution-based Bayesian predictive classification improved the performance from the MAP estimation and the marginalization results, reducing the WERs by 2.3 % and 1.2 %, respectively, when we only used one utterance (3.3 seconds on average) for the adaptation data. Since all the results use the same prior hyperparameter values, the improvement purely comes from the marginalization effect. In addition, if the amount of adaptation data increased, the performance of these three methods converged to the same value, which is also expected, based on the discussion of analytical results of the Student's t-distribution in Eq. (7.173).

The use of VB-based Bayesian predictive classification makes acoustic modeling in speech recognition a totally Bayesian framework that follows a consistent concept, whereby all acoustic procedures (model parameter estimation, model selection, and speech classification) are carried out based on posterior distributions. For example, compare the variational Bayesian speech recognition framework with a conventional ML-BIC approach: the model parameter estimation, model selection and speech classification are based on ML (Chapter 3) and BIC (Chapter 6). BIC is an asymptotic criterion that is theoretically effective only when the amount of training data is sufficiently large. Therefore, for a small amount of training data, model selection does not perform well because of the uncertainty of the ML estimates. The next section aims at solving the problem caused by a small amount of training data by using VB.

7.3.6 Decision tree clustering

This section revisits decision tree clustering of the context-dependent HMM states, as we discussed in Section 6.5, based on the VB framework (Watanabe *et al.* 2004, Hashimoto, Zen, Nankaku *et al.* 2008, Shiota, Hashimoto, Nankaku *et al.* 2009). Similarly to Eq. (6.74), we approximate the Bayes factor in Section 6.2 by selecting an appropriate question at each split, chosen to increase the variational lower bound/VB objective function $\mathcal{F}^M$ in the VB framework, as discussed in Section 7.3.4. When node n is split into a Yes node (n^Q_Y) and No node (n^Q_N) by question Q (we use $M_{Q(n)}$ with this hypothesized model, obtained from question Q, and M_n with the original model), the appropriate question $\widetilde{Q}(n)$ is chosen from a set of questions as follows:

$$\begin{aligned}\widetilde{Q}(n) &= \arg\max_Q \log\left(\frac{p(M_{Q(n)}|\mathbf{O})}{p(M_n|\mathbf{O})}\right)\\ &\approx \arg\max_Q \Delta\mathcal{F}_{Q(n)},\end{aligned} \tag{7.174}$$

where $\Delta\mathcal{F}^{Q(n)}$ is the gain in the VB objective function when node n is split by Q, which is defined as:

$$\Delta\mathcal{F}_{Q(n)} = \arg\max_Q \mathcal{F}_{\Omega(n^Q_Y)} + \mathcal{F}_{\Omega(n^Q_N)} - \mathcal{F}_{\Omega(n)}. \tag{7.175}$$

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

$\Omega(n)$ denotes a set of (non-shared) context-dependent HMM states at node n in a decision tree. The question is chosen to maximize the gain in $\mathcal{F}^M$ by splitting. The VB objective function for decision tree construction is also simply calculated under the following same constraints as the ML approach:

- Data alignments $\widetilde{\gamma}_t(j,k)$ and $\widetilde{\xi}_t(i,j)$ for each state are fixed while splitting.
- Emission probability distribution in a state is represented by a single Gaussian distribution (i.e., $K = 1$).
- Covariance matrices have only diagonal elements.
- A contribution of state transitions a_{ij} and initial weights π_j for likelihood is disregarded.

By using these conditions, the objective function is obtained without iterative calculations, which reduces the calculation time. Under conditions of fixed data assignment and single Gaussian assumptions, the latent variable part of $\mathcal{F}^M$ can be disregarded, i.e., Eq. (7.116) is approximated as

$$\mathcal{F}^M \approx \mathcal{F}^M_\Theta. \tag{7.176}$$

In the VB objective function of model parameter $\mathcal{F}^M_\Theta$ (Eq. (7.126)), the factors of posterior parameters of state transition $\widetilde{\phi}^a$ and mixture component $\widetilde{\phi}^\omega$ can also be disregarded under the above conditions. Therefore, the objective function $\mathcal{F}_\Omega$ in node n (we omit the index n for simplicity) for assigned data set $\mathbf{O}_\Omega = \{\mathbf{O}(i)|i \in \Omega\}$, where $\mathbf{O}(i) = \{\mathbf{o}_t(i) \in \mathbb{R}^D|t = 1, \cdots, T(i)\}$ can be obtained from the modification of $\mathcal{F}^M_\Theta$ in Eq. (7.126) as follows:

$$\mathcal{F}_\Omega = \log\left((2\pi)^{-\frac{\widetilde{\gamma}_\Omega D}{2}} \left(\frac{\phi^\mu_\Omega}{\widetilde{\phi}^\mu_\Omega}\right)^{\frac{D}{2}} \frac{2^{\frac{\widetilde{\phi}^r_\Omega D}{2}} \left(\Gamma\left(\frac{\widetilde{\phi}^r_\Omega}{2}\right)\right)^D \prod_{d=1}^D \left(r^0_{\Omega d}\right)^{\frac{\phi^r_\Omega}{2}}}{2^{\frac{\phi^r_\Omega D}{2}} \left(\Gamma\left(\frac{\phi^r_\Omega}{2}\right)\right)^D \prod_{d=1}^D \left(\widetilde{r}_{\Omega d}\right)^{\frac{\widetilde{\phi}^r_\Omega}{2}}}\right), \tag{7.177}$$

where $\{\widetilde{\phi}^\mu_\Omega, \widetilde{\boldsymbol{\mu}}_\Omega, \widetilde{\phi}^r_\Omega, \{\widetilde{r}_{\Omega d}\}_{d=1}^D\}(\triangleq \widetilde{\Psi}_\Omega)$ is a subset of the posterior parameters in Eq. (7.66), and is represented by:

$$\begin{cases} \widetilde{\phi}^\mu_\Omega & = \phi^\mu_\Omega + \widetilde{\gamma}_\Omega, \\ \widetilde{\boldsymbol{\mu}}_\Omega & = \frac{\phi^\mu_\Omega \boldsymbol{\mu}^0_\Omega + \widetilde{\boldsymbol{\gamma}}^{(1)}_\Omega}{\phi^\mu_\Omega + \widetilde{\gamma}_\Omega}, \\ \widetilde{\phi}^r_\Omega & = \phi^r_\Omega + \widetilde{\gamma}_\Omega, \\ \widetilde{r}_{\Omega d} & = \widetilde{\gamma}^{(2)}_{\Omega d} + \phi^\mu_\Omega(\mu^0_{\Omega d})^2 - \widetilde{\phi}^\mu_\Omega(\widetilde{\mu}_{\Omega d})^2 + r^0_{\Omega d}. \end{cases} \tag{7.178}$$

$\widetilde{\gamma}_\Omega, \widetilde{\boldsymbol{\gamma}}^{(1)}_\Omega$ and $\widetilde{\gamma}^{(2)}_{\Omega d}$ are the sufficient statistics of a set of states in node n, as defined as follows.

$$\begin{cases} \widetilde{\gamma}_\Omega & \triangleq & \sum_{i\in\Omega} \sum_{t=1}^{T(i)}, \\ \widetilde{\boldsymbol{\gamma}}^{(1)}_\Omega & \triangleq & \sum_{i\in\Omega} \sum_{t=1}^{T(i)} \mathbf{o}_t(i), \\ \widetilde{\gamma}^{(2)}_{\Omega d} & \triangleq & \sum_{i\in\Omega} \sum_{t=1}^{T(i)} (o_{td}(i))^2. \end{cases} \tag{7.179}$$

Note that since we use the hard aligned data $\mathbf{O}(i)$ (based on the Viterbi algorithm), the assignment information $\widetilde{\gamma}_t(i)$ is included in this representation. Here,

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

$\{\phi^{\mu}_{\Omega}, \boldsymbol{\mu}^{0}_{\Omega}, \phi^{r}_{\Omega}, \{r^{0}_{\Omega d}\}^{D}_{d=1}\}(\triangleq \Psi_{\Omega})$ is a set of prior parameters. One choice of setting the prior hyperparameters $\boldsymbol{\mu}^{0}_{\Omega}$ and $r^{0}_{\Omega d}$ would be to set them by using monophone (root node) HMM state statistics ($\tilde{\gamma}_{\text{root}}$, $\tilde{\gamma}^{(1)}_{\text{root}}$ and $\tilde{\gamma}^{(2)}_{\text{root}}$) as follows:

$$\begin{aligned}\boldsymbol{\mu}^{0}_{\Omega} &= \frac{\tilde{\boldsymbol{\gamma}}^{(1)}_{\text{root}}}{\tilde{\gamma}_{\text{root}}},\\ r^{0}_{\Omega d} &= \phi^{r}_{\Omega}\left(\frac{\tilde{\gamma}^{(2)}_{\text{root, d}}}{\tilde{\gamma}_{\text{root}}} - \left(\mu^{0}_{\text{root},d}\right)^{2}\right).\end{aligned} \tag{7.180}$$

The other parameters ϕ^{μ}_{Ω} and ϕ^{r}_{Ω} are set manually. By substituting Eq. (7.178) into Eq. (7.177), the gain $\Delta\mathcal{F}_{Q(n)}$ can be obtained when n is split into n^{Q}_{Y}, n^{Q}_{N} by question Q:

$$\Delta\mathcal{F}_{Q(n)} = f(\tilde{\Psi}_{\Omega(n^{Q}_{Y})}) + f(\tilde{\Psi}_{\Omega(n^{Q}_{N})}) - f(\tilde{\Psi}_{\Omega(n)}) - f(\Psi_{\Omega(n)}). \tag{7.181}$$

Here, $f(\Psi)$ is defined by:

$$f(\Psi) \triangleq -\frac{D}{2}\log\phi^{\mu} - \frac{\phi^{r}}{2}\sum_{d=1}^{D}\log r_{d} + D\log\Gamma\left(\frac{\phi^{r}}{2}\right). \tag{7.182}$$

The terms that do not contribute to $\Delta\mathcal{F}^{Q(n)}$ are disregarded. The final term in Eq. (7.181) is only computed from the prior hyperparameter Ψ. Similarly to the BIC criterion in Eq. (6.92), node splitting stops when the condition

$$\Delta\mathcal{F}_{Q(n)} \leq 0 \tag{7.183}$$

is satisfied. A model structure based on the VB framework can be obtained by executing this construction for all trees, resulting in the maximization of total $\mathcal{F}^{M}$. This implementation based on the decision tree method does not require iterative calculations, and can construct clustered-state HMMs efficiently. There is another major method for the construction of clustered-state HMMs that uses a successive state splitting algorithm, and which does not remove latent variables in HMMs (Takami & Sagayama 1992, Ostendorf & Singer 1997). Therefore, this requires the VB Baum–Welch algorithm and calculation of the latent variable part of the lower bound/VB objective function for each splitting. This is realized as the VB SSS algorithm by Jitsuhiro & Nakamura (2004).

The relationship between VB model selection and the conventional BIC model selection, based on Eqs. (7.181) and (6.89), respectively, is discussed below. Based on the condition of a sufficiently large amount of data, the posterior hyperparameters in Eq. (7.178) are approximated as follows:

$$\begin{aligned}\tilde{\phi}^{\mu}_{\Omega}, \tilde{\phi}^{r}_{\Omega} &\to \tilde{\gamma}_{\Omega},\\ \tilde{\boldsymbol{\mu}}_{\Omega} &\to \frac{\tilde{\boldsymbol{\gamma}}^{(1)}_{\Omega}}{\tilde{\gamma}_{\Omega}},\\ \tilde{r}_{\Omega d} &\to \tilde{\gamma}^{(2)}_{\Omega d} - \frac{\left(\tilde{\gamma}^{(1)}_{\Omega d}\right)^{2}}{\tilde{\gamma}_{\Omega}}.\end{aligned} \tag{7.184}$$

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

In addition, from Stirling's approximation, the logarithmic gamma function has the following relationship:

$$\log \Gamma\left(\frac{x}{2}\right) \to \frac{x}{2}\log\left(\frac{x}{2}\right) - \frac{x}{2} - \frac{1}{2}\log\left(\frac{x}{2\pi}\right), \tag{7.185}$$

when $|x| \to \infty$. By substituting Eq. (7.184) into Eq. (7.182) and using Eq. (7.185), $f(\widetilde{\Psi})$ is approximated as

$$\begin{aligned} f(\widetilde{\Psi}) \to & -\frac{D}{2}\log\widetilde{\gamma}_\Omega - \frac{\widetilde{\gamma}_\Omega}{2}\sum_{d=1}^{D}\log\left(\widetilde{\gamma}^{(2)}_{\Omega d} - \frac{\left(\widetilde{\gamma}^{(1)}_{\Omega d}\right)^2}{\widetilde{\gamma}_\Omega}\right) \\ & + D\left(\frac{\widetilde{\gamma}_\Omega}{2}\log\left(\frac{\widetilde{\gamma}_\Omega}{2}\right) - \frac{\widetilde{\gamma}_\Omega}{2} - \frac{1}{2}\log\left(\frac{\widetilde{\gamma}_\Omega}{2\pi}\right)\right) \\ = & -\frac{\widetilde{\gamma}_\Omega}{2}\left(D\left(1+\log(2\pi)\right) + \underbrace{\sum_{d=1}^{D}\log\left(\frac{\widetilde{\gamma}^{(2)}_{\Omega d}}{\widetilde{\gamma}_\Omega} - \frac{\left(\widetilde{\gamma}^{(1)}_{\Omega d}\right)^2}{(\widetilde{\gamma}_\Omega)^2}\right)}_{\approx \log|\boldsymbol{\Sigma}^{\mathrm{ML}}_\Omega|}\right) - D\log\widetilde{\gamma}_\Omega. \end{aligned} \tag{7.186}$$

Then, an asymptotic form of Eq. (7.182) is composed of a log-likelihood gain term and a penalty term depending on the number of free parameters ($2D$ in this diagonal covariance Gaussian case) and the amount of training data, i.e., the asymptotic form becomes the BIC-type objective function form, as shown in Eq. (6.88). Therefore, VB theoretically involves the BIC objective function, and so BIC model selection is asymptotically equivalent to VB model selection, which demonstrates the advantages of VB, especially for small amounts of training data.

7.3.7 Determination of HMM topology

Once a clustered-state model structure is obtained, acoustic model selection is completed by determining the number of mixture components per state. GMMs include latent variables, and their determination requires the VB Baum–Welch algorithm and computation of the latent variable part of the variational lower bound, unlike the clustering triphone HMM states in Section 7.4.5. Therefore, this section deals with determination of the number of GMM components per state by considering the latent variable effects. Then, the effectiveness of VB model selection in latent variable models is confirmed (Jitsuhiro & Nakamura 2004) for the successive state splitting algorithm, and the effectiveness of VB model selection for GMMs is re-confirmed (Valente & Wellekens 2003). In general, there are two methods for determining the number of mixture components. With the first method, the number of mixture components per state is the same for all states. The objective function $\mathcal{F}^M$ is calculated for each number of mixture components, and the number of mixture components that maximizes the total $\mathcal{F}^M$ is determined as being the appropriate one (fixed-number GMM method). With the second method, the number of mixture components per state can vary by state; here, Gaussians are split and merged to increase $\mathcal{F}^M$ and determine the number of mixture components

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

in each state (varying-number GMM method). A model obtained by the varying-number GMM method is expected to be more accurate than one obtained by the fixed-number GMM method, although the varying-number GMM method requires more computation time.

We require the variational lower bound for each state to determine the number of mixture components. In this case, the state alignments vary and states are expressed as GMMs. Therefore, the model includes latent variables and the component $\mathcal{F}^M_{S,V}$ cannot be disregarded, unlike the case of triphone HMM state clustering. However, since the number of mixture components is determined for each state and the state alignments do not change greatly, the contribution of the state transitions to the objective function is expected to be small, and can be ignored. Therefore, the objective function $\mathcal{F}^M$ for a particular state j is represented from Eqs. (7.126) and (7.135) as follows:

$$(\mathcal{F}^M)_j = (\mathcal{F}^M_\Theta)_j - (\mathcal{F}^M_V)_j, \tag{7.187}$$

where $(\mathcal{F}^M_\Theta)_j$ is represented by removing the HMM terms in Eq. (7.126) as follows:

$$\begin{aligned}(\mathcal{F}^M_\Theta)_j = \log & \frac{\Gamma(\sum_k \phi^\omega_{jk})}{\Gamma(\sum_k \widetilde{\phi}^\omega_{jk})} \frac{\prod_k \Gamma(\widetilde{\phi}^\omega_{jk})}{\prod_k \Gamma(\phi^\omega_{jk})} \\ & + \sum_k \log \left((2\pi)^{-\frac{\widetilde{\gamma}_{jk} D}{2}} \left(\frac{\phi^\mu_{jk}}{\widetilde{\phi}^\mu_{jk}} \right)^{\frac{D}{2}} \frac{\left(\Gamma\left(\frac{\widetilde{\phi}^r_{jk}}{2} \right) \right)^D \prod_d \left(\frac{r^0_{jkd}}{2} \right)^{\frac{\phi^r_{jk}}{2}}}{\left(\Gamma\left(\frac{\phi^r_{jk}}{2} \right) \right)^D \prod_d \left(\frac{\widetilde{r}_{jkd}}{2} \right)^{\frac{\widetilde{\phi}^r_{jk}}{2}}} \right).\end{aligned} \tag{7.188}$$

Similarly, $(\mathcal{F}^M_V)_j$ is also represented as follows:

$$\begin{aligned}(\mathcal{F}^M_V)_j = & \sum_k \widetilde{\gamma}_{jk} \left(\Psi\left(\widetilde{\phi}^\omega_{jk} \right) - \Psi\left(\sum_{k'} \widetilde{\phi}^\omega_{jk'} \right) \right) \\ & - \frac{1}{2} \sum_k \widetilde{\gamma}_{jk} \left(D \left(\log(2\pi) + \frac{1}{\widetilde{\phi}^\mu_{jk}} - \Psi\left(\frac{\widetilde{\phi}^r_{jk}}{2} \right) \right) + \sum_d \log \frac{\widetilde{r}_{jkd}}{2} \right) \\ & - \frac{1}{2} \sum_k \left(\widetilde{\phi}^r_{jk} \sum_{t,d} \frac{\widetilde{\gamma}_t(j,k)(o_{td} - \widetilde{\mu}_{jkd})^2}{\widetilde{r}_{jkd}} \right) - \log \left(\sum_V \tilde{u}(\mathbf{O}, V|M) \right).\end{aligned} \tag{7.189}$$

Therefore, with the fixed-number GMM method, the total $\mathcal{F}^M$ is obtained by summing up all states' $(\mathcal{F}^M)_j$, which determines the number of mixture components per state. With the varying-number GMM method, the change of $(\mathcal{F}^M)_j$ per state is compared after merging or splitting the Gaussians, which also determines the number of mixture components. The number of mixture components is also automatically determined by using the BIC/MDL objective function (Chen & Gopinath 1999, Shinoda & Iso 2001). However, the BIC/MDL objective function is based on the asymptotic condition and cannot be applied to latent models in principle. On the other hand, the variational lower bound derived by VB does not need the asymptotic condition and can determine an appropriate model structure with latent variables.

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

Table 7.2 Automatic determination of acoustic model topology.

	Read speech (JNAS)	Read speech (WSJ)	Isolated word (JEIDA)	Lecture (CSJ)
VB	91.7 %	91.3 %	97.9 %	74.5 %
# states	912	2504	254	1986
# components	40	32	35	32
ML + Dev. Set	91.4 %	91.3 %	98.1 %	74.2 %
# states	1000	7500	1000	3000
# components	30	32	15	32

Table 7.2 shows experimental results for automatic determination of the acoustic model topology by using VB and the conventional heuristic approach that determines the model topology by evaluating ASR performance on development sets. Note that VB was only used for the model topology determination, and the other procedures (e.g., training and decoding) were performed by using the conventional (ML) approaches. Therefore, Table 7.2 simply shows the effectiveness of the model selection. We used two tasks based on read speech recognition of news articles, *JNAS* (Shikano, Kawahara, Kobayashi *et al*. 1999) and *WSJ* (Paul & Baker 1992), an isolated word speech recognition task (JEIDA 100 city name recognition), and a lecture speech recognition task, CSJ (Furui, Maekawa & Isahara 2000). Table 7.2 provides the ASR performance of the determined model topology with the number of total HMM states and a mixture component in an HMM state, where we used the same number of mixture component for all states. In the various ASR tasks, VB achieved comparable performance to the conventional method by selecting appropriate model topologies without using a development set. Thus, these experiments proved that the VB model selection method can *automatically* determine an appropriate acoustic model topology with a comparable performance to that obtained by using a development set.

7.4 Structural Bayesian linear regression for hidden Markov model

As discussed in Section 3.5, a Bayesian treatment of the affine transformation parameters of CDHMM is an important issue to improve the generalization capability of the model adaptation. While the regression tree used in the conventional maximum likelihood linear regression (MLLR) can be considered one form of prior knowledge, i.e., how various Gaussian distributions are related, another approach is to explicitly construct and use *prior knowledge of regression parameters* in an approximated Bayesian paradigm.

For example, maximum a-posteriori linear regression (MAPLR) (Chesta, Siohan & Lee 1999) replaces the ML criterion with the MAP criterion introduced in Chapter 4 in the estimation of regression parameters. Quasi-Bayes linear regression (Chien 2002) also replaces the ML/MAP criterion with a quasi-Bayes criterion. With the explicit prior knowledge acting as a regularization term, MAPLR appears to be less susceptible to the

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

overfitting problem. The MAPLR is extended to the structural MAP (SMAP) (Shinoda & Lee 2001) and the structural MAPLR (SMAPLR) (Siohan, Myrvoll & Lee 2002), both of which fully utilize the Gaussian tree structure used in the model selection approach to efficiently set the hyperparameters in prior distributions. In SMAP and SMAPLR, the hyperparameters in the prior distribution in a target node are obtained from the statistics in its parent node. Since the total number of speech frames assigned to a set of Gaussians in the parent node is always larger than that in the target node, the statistics obtained in the parent node are more reliable than those in the target node, and these can be good prior knowledge for transformation parameter estimation in the target node.

Another extension of MAPLR is to replace MAP approximation by a fully Bayesian treatment of latent models, using VB. This section employs VB for the linear regression problem (Watanabe & Nakamura 2004, Yu & Gales 2006, Watanabe, Nakamura & Juang 2013), but we focus on model selection and efficient prior utilization at the same time, in addition to estimation of the linear transformation parameters of HMMs proposed in previous work (Watanabe & Nakamura 2004, Yu & Gales 2006). In particular, we consistently use the variational lower bound as the optimization criterion for the model structure and hyperparameters, in addition to the posterior distributions of the transformation parameters and the latent variables. As we discussed in Section 7.2, since this optimization leads the approximated variational posterior distributions to the true posterior distributions theoretically in the sense of minimizing Kullback–Leibler divergence between them, the above consistent approach leads to improved generalization capability (Neal & Hinton 1998, Attias 1999, Ueda & Ghahramani 2002).

This section provides an analytical solution to the variational lower bound by marginalizing all possible transformation parameters and latent variables introduced in the linear regression problem. The solution is based on a variance-normalized representation of Gaussian mean vectors to simplify the solution as normalized domain MLLR. As a result of variational calculation, we can marginalize the transformation parameters in all nodes used in the structural prior setting. This is a part of the solution of the variational message passing algorithm (Winn & Bishop 2006), which is a general framework of variational inference in a graphical model. Furthermore, the optimization of the model topology and hyperparameters in the proposed approach yields an additional benefit in the improvement of the generalization capability. For example, this approach infers the linear regression without controlling the Gaussian cluster topology and hyperparameters as the tuning parameters. Thus linear regression for HMM parameters is accomplished without excessive parameterization in a Bayesian sense.

7.4.1 Variational Bayesian linear regression

This section provides an analytical solution for Bayesian linear regression by using a variational lower bound. The previous section only considers a regression matrix in leaf node $j \in \mathcal{J}_M$, but we also consider a regression matrix in leaf or non-leaf node $i \in \mathcal{I}_M$ in the Gaussian tree given model structure M. Then we focus on a set of regression matrices in all nodes $\mathbf{\Lambda}_{\mathcal{I}_M} = \{\mathbf{W}_i | i = 1, \cdots, |\mathcal{I}_M|\}$, instead of $\mathbf{\Lambda}_{\mathcal{J}_M}$, and marginalize $\mathbf{\Lambda}_{\mathcal{I}_M}$ in a Bayesian manner. This extension involves the structural prior setting as proposed in

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

SMAP and SMAPLR (Shinoda & Lee 2001, Siohan *et al.* 2002, Yamagishi, Kobayashi, Nakano *et al.* 2009).

In this section, we mainly deal with:

- prior distribution of model parameters $p(\mathbf{\Lambda}_{\mathcal{I}_M}; M, \Psi)$;
- true posterior distribution of model parameters and latent variables $p(\mathbf{\Lambda}_{\mathcal{I}_M}, Z|\mathbf{O}; M, \Psi)$;
- variational posterior distribution of model parameters and latent variables $q(\mathbf{\Lambda}_{\mathcal{I}_M}, Z|\mathbf{O}; M, \Psi)$;
- generative model distribution $p(\mathbf{O}, Z|\mathbf{\Lambda}_{\mathcal{I}_M}; \Theta)$.

Note that the prior and generative model distributions are given, as shown in the generative process of Algorithm 12, and we obtain the variational posterior distribution, which is an approximation of the true posterior distribution.

7.4.2 Generative model

As discussed in Section 3.5, the generative model distribution with the expectation with respect to the posterior distributions of latent variables is represented as follows:

$$\mathbb{E}_{(Z)}\left[\log p(\mathbf{O}, Z|\mathbf{\Lambda}_{\mathcal{I}_M}; \Theta)\right] = \sum_{k=1}^{K}\sum_{t=1}^{T} \gamma_t(k) \log \mathcal{N}(\mathbf{o}_t|\boldsymbol{\mu}_k^{ad}, \mathbf{\Sigma}_k), \tag{7.190}$$

where $p(\mathbf{O}, Z|\mathbf{\Lambda}_{\mathcal{I}_M}; \Theta)$ is the generative model distribution of the transformed HMM parameters with transformed mean vectors $\boldsymbol{\mu}_k^{ad}$. We use $\boldsymbol{\mu}_k^{ad}$ based on the following variance normalized representation:

$$\boldsymbol{\mu}_k^{ad} = \mathbf{C}_k \mathbf{W}_j \boldsymbol{\xi}_k. \tag{7.191}$$

$\mathbf{C}_k$ is the Cholesky decomposition matrix of $\mathbf{\Sigma}_k$, and $\boldsymbol{\xi}_k = [1, ((\mathbf{C}_k)^{-1}\boldsymbol{\mu}_k^{ini})^\intercal]^\intercal$ is obtained based on the initial mean vector $\boldsymbol{\mu}_k^{ini}$. This representation makes the calculation simple.[5]

7.4.3 Variational lower bound

With regard to variational Bayesian approaches, we first focus on the following marginal log-likelihood $p(\mathbf{O}; \Theta, M, \Psi)$ with a set of HMM parameters Θ, a set of hyperparameters Ψ, and a model structure:[6,7]

$$\begin{aligned}&\log p(\mathbf{O}; \Theta, M, \Psi)\\&= \log\left(\int \sum_{Z} p(\mathbf{O}, Z|\mathbf{\Lambda}_{\mathcal{I}_M}; \Theta) p(\mathbf{\Lambda}_{\mathcal{I}_M}; M, \Psi) d\mathbf{\Lambda}_{\mathcal{I}_M}\right).\end{aligned} \tag{7.192}$$

[5] Hahm, Ogawa, Fujimoto *et al.* (2012) discuss the use of conventional MLLR estimation without the variance normalization in the VB framework and its application to feature-space MLLR (fVBLR).

[6] Ψ and M can also be marginalized by setting their distributions. This section point-estimates Ψ and M by a MAP approach, similar to the evidence approximation in Chapter 5.

[7] We can also marginalize the HMM parameters Θ. This corresponds to jointly optimizing HMM and linear regression parameters.

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

$p(\mathbf{\Lambda}_{\mathcal{I}_M}; M, \Psi)$ is a prior distribution of transformation matrices $\mathbf{\Lambda}_{\mathcal{I}_M}$. In the following explanation, we omit Θ, M, and Ψ in the prior distribution and generative model distribution for simplicity, i.e., $p(\mathbf{\Lambda}_{\mathcal{I}_M}; M, \Psi) \to p(\mathbf{\Lambda}_{\mathcal{I}_M})$, and $p(\mathbf{O}, Z|\mathbf{\Lambda}_{\mathcal{I}_M}; \Theta) \to p(\mathbf{O}, Z|\mathbf{\Lambda}_{\mathcal{I}_M})$.

Similarly to Eq. (7.15), since the variational Bayesian approach focuses on the variational lower bound of the marginal log likelihood $\mathcal{F}(M, \Psi)$ with a set of hyperparameters Ψ and a model structure M, Eq. (7.192) is represented as follows:

$$\begin{aligned}
&\log p(\mathbf{O}; \Theta, M, \Psi) \\
&\quad = \log \left(\int \sum_{Z} \frac{p(\mathbf{O}, Z|\mathbf{\Lambda}_{\mathcal{I}_M}) p(\mathbf{\Lambda}_{\mathcal{I}_M})}{q(\mathbf{\Lambda}_{\mathcal{I}_M}, Z)} q(\mathbf{\Lambda}_{\mathcal{I}_M}, Z) d\mathbf{\Lambda}_{\mathcal{I}_M} \right) \\
&\quad \geq \underbrace{\mathbb{E}_{(\mathbf{\Lambda}_{\mathcal{I}_M}, Z)} \left[\log \frac{p(\mathbf{O}, Z|\mathbf{\Lambda}_{\mathcal{I}_M}) p(\mathbf{\Lambda}_{\mathcal{I}_M})}{q(\mathbf{\Lambda}_{\mathcal{I}_M}, Z)} \right]}_{\triangleq \mathcal{F}(M, \Psi)}.
\end{aligned} \tag{7.193}$$

The inequality in Eq. (7.193) is supported by the Jensen's inequality in Eq. (7.10). $q(\mathbf{\Lambda}_{\mathcal{I}_M}, Z)$ is an arbitrary distribution, and is optimized by using a variational method to be discussed later. For simplicity, we omit M, Ψ, and $\mathbf{O}$ from the distributions. As discussed in Section 7.1, the variational lower bound is a better approximation of the marginal log likelihood than the auxiliary functions of maximum likelihood EM and maximum a-posteriori EM algorithms that point-estimate model parameters, especially for small amount of training data. Therefore, the variational Bayes can mitigate the sparse data problem that the conventional approaches must resolve.

The variational Bayes regards the variational lower bound $\mathcal{F}(M, \Psi)$ as an objective function for the model structure and hyperparameter, and an objective functional for the joint posterior distribution of the transformation parameters and latent variables (Attias 1999, Ueda & Ghahramani 2002). In particular, if we consider the true posterior distribution $p(\mathbf{\Lambda}_{\mathcal{I}_M}, Z|\mathbf{O})$ (we omit conditional variables M and Ψ for simplicity), we obtain the following relationship:

$$\mathrm{KL}\left(q(\mathbf{\Lambda}_{\mathcal{I}_M}, Z) \| p(\mathbf{\Lambda}_{\mathcal{I}_M}, Z|\mathbf{O}) \right) = \log p(\mathbf{O}; \Theta, M, \Psi) - \mathcal{F}(M, \Psi). \tag{7.194}$$

This equation means that maximizing the variational lower bound $\mathcal{F}(M, \Psi)$ with respect to $q(\mathbf{\Lambda}_{\mathcal{I}_M}, Z)$ corresponds to minimizing the KL divergence between $q(\mathbf{\Lambda}_{\mathcal{I}_M}, Z)$ and $p(\mathbf{\Lambda}_{\mathcal{I}_M}, Z|\mathbf{O})$ indirectly. Therefore, this optimization leads to finding $q(\mathbf{\Lambda}_{\mathcal{I}_M}, Z)$, which approaches the true posterior distribution.[8]

[8] The following sections assume factorization forms of $q(\mathbf{\Lambda}_{\mathcal{I}_M}, Z)$ to make solutions mathematically tractable. However, this factorization assumption weakens the relationship between the KL divergence and the variational lower bound. For example, if we assume $q(\mathbf{\Lambda}_{\mathcal{I}_M}, Z) = q(\mathbf{\Lambda}_{\mathcal{I}_M})q(Z)$, and focus on the KL divergence between $q(\mathbf{\Lambda}_{\mathcal{I}_M})$ and $p(\mathbf{\Lambda}_{\mathcal{I}_M}|\mathbf{O})$, we obtain the following inequality:

$$\mathrm{KL}\left(q(\mathbf{\Lambda}_{\mathcal{I}_M}) \| p(\mathbf{\Lambda}_{\mathcal{I}_M}|\mathbf{O}) \right) \leq \log p(\mathbf{O}; \Theta, M, \Psi) - \mathcal{F}(M, \Psi). \tag{7.195}$$

Compared with Eq. (7.194), the relationship between the KL divergence and the variational lower bound is less direct due to the inequality relationship. In general, the factorization assumption distances optimal variational posteriors from the true posterior within the VB framework.

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

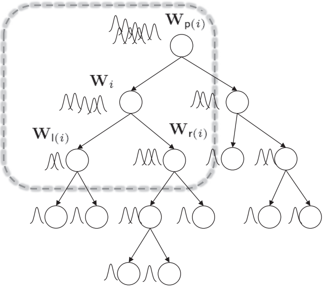
$\mathbf{W}_{\mathsf{p}(i)}$
$\mathbf{W}_i$
$\mathbf{W}_{\mathsf{l}(i)}$
$\mathbf{W}_{\mathsf{r}(i)}$

Figure 7.1 Binary tree structure with transformation matrices. If we focus on node i, the transformation matrices in the parent node, left child node, and right child node are represented as $\mathbf{W}_{\mathsf{p}(i)}$, $\mathbf{W}_{\mathsf{l}(i)}$, and $\mathbf{W}_{\mathsf{r}(i)}$, respectively.

Thus, in principle, we can straightforwardly obtain the (sub) optimal model structure, hyperparameters, and posterior distribution, as follows:

$$\begin{aligned}\tilde{m} &= \arg\max_{M} \mathcal{F}(M, \Psi), \\ \tilde{\Psi} &= \arg\max_{\Psi} \mathcal{F}(M, \Psi), \\ \tilde{q}(\mathbf{\Lambda}_{\mathcal{I}_M}, Z) &= \arg \max_{q(\mathbf{\Lambda}_{\mathcal{I}_M}, Z)} \mathcal{F}(M, \Psi). \end{aligned} \tag{7.196}$$

These optimization steps are performed alternately, and finally lead to local optimum solutions, similar to the EM algorithm. However, it is difficult to deal with the joint distribution $q(\mathbf{\Lambda}_{\mathcal{I}_M}, Z)$ directly, and we propose factorizing them by utilizing a Gaussian tree structure. In addition, we also set a conjugate form of the prior distribution $p(\mathbf{\Lambda}_{\mathcal{I}_M})$. This procedure is a typical recipe of VB to make a solution mathematically tractable similarly to that of the classical Bayesian adaptation approach.

Structural prior distribution setting in a binary tree

We utilize a Gaussian tree structure to factorize the prior distribution $p(\mathbf{\Lambda}_{\mathcal{I}_M})$. We consider a binary tree structure, but the formulation is applicable to a general non-binary tree. We define the parent node of i as $\mathsf{p}(i)$, the left child node of i as $\mathsf{l}(i)$, and the right child node of i as $\mathsf{r}(i)$, as shown in Figure 7.1, where a transformation matrix is prepared for each corresponding node i. If we define $\mathbf{W}_1$ as the transformation matrix in the root node, we assume the following factorization for the hierarchical prior distribution $p(\mathbf{\Lambda}_{\mathcal{I}_M})$:

$$\begin{aligned} p(\mathbf{\Lambda}_{\mathcal{I}_M}) &= p(\mathbf{W}_1, \cdots, \mathbf{W}_{|\mathcal{I}_M|}) \\ &= p(\mathbf{W}_1) p(\mathbf{W}_{\mathsf{l}(1)}|\mathbf{W}_1) p(\mathbf{W}_{\mathsf{r}(1)}|\mathbf{W}_1) \\ &\quad p(\mathbf{W}_{\mathsf{l}(\mathsf{l}(1))}|\mathbf{W}_{\mathsf{l}(1)}) p(\mathbf{W}_{\mathsf{r}(\mathsf{l}(1))}|\mathbf{W}_{\mathsf{l}(1)}) \\ &\quad p(\mathbf{W}_{\mathsf{l}(\mathsf{r}(1))}|\mathbf{W}_{\mathsf{r}(1)}) p(\mathbf{W}_{\mathsf{r}(\mathsf{r}(1))}|\mathbf{W}_{\mathsf{r}(1)}) \cdots \\ &= \prod_{i \in \mathcal{I}_M} p(\mathbf{W}_i|\mathbf{W}_{\mathsf{p}(i)}). \end{aligned} \tag{7.197}$$

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

To make the prior distribution a product form in the last line of Eq. (7.197), we define $p(\mathbf{W}_1) \triangleq p(\mathbf{W}_1|\mathbf{W}_{\mathsf{p}(1)})$. As seen, the effect of the transformation matrix in a target node propagates to its child nodes.

This hierarchical prior setting is based on an intuitive assumption that the statistics in a target node are highly correlated with the statistics in its parent node. In addition, since the total number of speech frames assigned to a set of Gaussians in the parent node is always larger than that in the target node, the statistics obtained in the parent node are more reliable than in the target node, and these can be good prior knowledge for the transformation parameter estimation in the target node.

With a Bayesian approach, we need to set a practical form of the above prior distributions. A conjugate distribution in Section 2.1.4 is preferable as far as obtaining an analytical solution is concerned, and we set a matrix variate Gaussian distribution similar to maximum a-posteriori linear regression (MAPLR (Chesta *et al.* 1999)). A matrix variate Gaussian distribution is defined in Appendix C.9 as follows:

$$\begin{aligned} p(\mathbf{W}_i) &= \mathcal{N}(\mathbf{W}_i|\mathbf{M}_i, \mathbf{\Phi}_i, \mathbf{\Omega}_i) \\ &\triangleq C_{\mathcal{N}}(\mathbf{\Phi}_i, \mathbf{\Omega}_i) \exp\left(-\frac{1}{2}\mathrm{tr}\left[(\mathbf{W}_i - \mathbf{M}_i)^{\mathsf{T}} \mathbf{\Phi}_i^{-1} (\mathbf{W}_i - \mathbf{M}_i)\mathbf{\Omega}_i^{-1}\right]\right), \end{aligned} \tag{7.198}$$

where $C_{\mathcal{N}}(\mathbf{\Phi}_i, \mathbf{\Omega}_i)$ is a normalization constant defined as:

$$C_{\mathcal{N}}(\mathbf{\Phi}_i, \mathbf{\Omega}_i) \triangleq (2\pi)^{-\frac{D(D+1)}{2}} |\mathbf{\Omega}_i|^{-\frac{D}{2}} |\mathbf{\Phi}_i|^{-\frac{D+1}{2}}. \tag{7.199}$$

$\mathbf{M}_i$ is a $D \times (D+1)$ location matrix, $\mathbf{\Omega}_i$ is a $(D+1) \times (D+1)$ symmetric scale matrix, and $\mathbf{\Phi}_i$ is a $D \times D$ symmetric scale matrix. The term $\mathbf{\Omega}_i$ represents correlation of column vectors, and $\mathbf{\Phi}_i$ represents correlation of raw vectors. These are hyperparameters of the matrix variate Gaussian distribution. There are many hyperparameters to be set, and this makes the implementation complicated. In this section, we try to find another conjugate distribution with fewer hyperparameters than Eq. (7.198). To obtain a simple solution for the final analytical results, we use a spherical Gaussian distribution that has the following constraints on $\mathbf{\Omega}_i$ and $\mathbf{\Phi}_i$:

$$\begin{aligned} \mathbf{\Phi}_i &\approx \mathbf{I}_D, \\ \mathbf{\Omega}_i &\approx \rho_i^{-1}\mathbf{I}_{D+1}, \end{aligned} \tag{7.200}$$

where $\mathbf{I}_D$ is the $D \times D$ identity matrix and ρ_i indicates a precision parameter. Then Eq. (7.198) can be rewritten as follows:

$$\begin{aligned} &\mathcal{N}(\mathbf{W}_i|\mathbf{M}_i, \mathbf{I}_D, \rho_i^{-1}\mathbf{I}_{D+1}) \\ &\quad = C_{\mathcal{N}}(\mathbf{I}_D, \rho_i^{-1}\mathbf{I}_{D+1}) \exp\left(-\frac{1}{2}\mathrm{tr}\left[\rho_i(\mathbf{W}_i - \mathbf{M}_i)^{\mathsf{T}}(\mathbf{W}_i - \mathbf{M}_i)\right]\right), \end{aligned} \tag{7.201}$$

where $C_{\mathcal{N}}(\mathbf{I}_D, \rho_i^{-1}\mathbf{I}_{D+1})$ is a normalization factor, and is defined as

$$C_{\mathcal{N}}(\mathbf{I}_D, \rho_i^{-1}\mathbf{I}_{D+1}) \triangleq \left(\frac{\rho_i}{2\pi}\right)^{\frac{D(D+1)}{2}}. \tag{7.202}$$

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

This approximation means that matrix elements do not have any correlation with each other. This can produce simple solutions for Bayesian linear regression.[9]

Based on the spherical matrix variate Gaussian distribution, the conditional prior distribution $p(\mathbf{W}_i|\mathbf{W}_{\mathsf{p}(i)})$ in Eq. (7.197) is obtaining by setting the location matrix as the transformation matrix $\mathbf{W}_{\mathsf{p}(i)}$ in the parent node with the precision parameter ρ_i as follows:

$$p(\mathbf{W}_i|\mathbf{W}_{\mathsf{p}(i)}) = \mathcal{N}(\mathbf{W}_i|\mathbf{W}_{\mathsf{p}(i)}, \mathbf{I}_D, \rho_i^{-1}\mathbf{I}_{D+1}). \tag{7.204}$$

Note that in the following sections $\mathbf{W}_i$ and $\mathbf{W}_{\mathsf{p}(i)}$ are marginalized. In addition, we set the location matrix in the root node as the deterministic value of $\mathbf{W}_{\mathsf{p}(1)} = [\mathbf{0}, \mathbf{I_D}]$. Since $\boldsymbol{\mu}_k^{ad} = \mathbf{C}_k\mathbf{W}_{\mathsf{p}(1)}\boldsymbol{\xi}_k = \boldsymbol{\mu}_k^{ini}$ from Eq. (7.191), this hyperparameter setting means that the initial mean vectors are not changed if we only use the prior knowledge. This makes sense in the case of a small amount of data by fixing the HMM parameters as their initial values; this in a sense also inherits the philosophical background of Bayesian adaptation, although the objective function has been changed from a-posteriori probability to a lower bound of the marginal likelihood. Therefore, we just have $\{\rho_i|i = 1, \cdots, |\mathcal{I}_M|\}$ as a set of hyperparameters Ψ, which will also be optimized in our framework.

Algorithm 12 Generative process of structural Bayesian transformation of CDHMM

```
Require: $\Psi$ and $\Theta$
1: Draw $\boldsymbol{\Lambda}_{\mathcal{I}_M}$ from $p(\boldsymbol{\Lambda}_{\mathcal{I}_M})$
2: Update $\Theta^{ad}$ from transformation matrices in leaf nodes $\boldsymbol{\Lambda}_{\mathcal{J}_M}$
3: Draw $\mathbf{O}$ from CDHMM with $\Theta^{ad}$
```

Variational calculus

In VB, we also assume the following factorization form for the posterior distribution $q(Z, \boldsymbol{\Lambda}_{\mathcal{I}_M})$:

$$q(Z, \boldsymbol{\Lambda}_{\mathcal{I}_M}) = q(Z)q(\boldsymbol{\Lambda}_{\mathcal{I}_M}) = q(Z)\prod_{i\in\mathcal{I}_M} q(\mathbf{W}_i). \tag{7.205}$$

Then, from the general variational calculation for $\mathcal{F}(M, \Psi)$ with respect to $q(\mathbf{W}_i)$ based on Eq. (7.25), we obtain the following (sub) optimal solution for $q(\mathbf{W}_i)$:

[9] A matrix variate Gaussian distribution in Eq. (7.198) is also represented by the following multivariate Gaussian distribution (Dawid 1981):

$$\begin{aligned}&\mathcal{N}(\mathbf{W}_i|\mathbf{M}_i, \boldsymbol{\Phi}_i, \boldsymbol{\Omega}_i)\\ &\propto \exp\left(-\frac{1}{2}\operatorname{vec}(\mathbf{W}_i - \mathbf{M}_i)^{\mathsf{T}}(\boldsymbol{\Omega}_i \otimes \boldsymbol{\Phi}_i)^{-1}\operatorname{vec}(\mathbf{W}_i - \mathbf{M}_i)^{-1}\right),\end{aligned} \tag{7.203}$$

where $\operatorname{vec}(\mathbf{W}_i - \mathbf{M}_i)$ is a vector formed by the concatenation of the columns of $(\mathbf{W}_i - \mathbf{M}_i)$, and $\otimes$ denotes the Kronecker product. Based on this form, a VB solution in this section could be extended without considering the variance normalized representation used (Chien 2002).

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

$$\begin{aligned}
&\log \tilde{q}(\mathbf{W}_i) \\
&\quad \propto \mathbb{E}_{(Z,\mathbf{W}_{\setminus i})}\left[\log p(\mathbf{O}, Z, \mathbf{\Lambda}_{\mathcal{I}_M})\right] \\
&\quad \propto \mathbb{E}_{(Z,\mathbf{W}_{\setminus i})}\left[\log p(\mathbf{O}, Z|\mathbf{\Lambda}_{\mathcal{I}_M})\right] + \mathbb{E}_{(\mathbf{W}_{\setminus i})}\left[\log p(\mathbf{\Lambda}_{\mathcal{I}_M})\right].
\end{aligned} \tag{7.206}$$

$\mathbf{W}_{\setminus i}$ means a set of transformation matrices at a set of nodes for $\mathcal{I}_M$ that does not include $\mathbf{W}_i$, i.e.,

$$\mathbf{W}_{\setminus i} = \{\mathbf{W}_{i'} | i' \in \mathcal{I}_M \setminus i\}. \tag{7.207}$$

Then, by using Eqs. (7.197) for $p(\mathbf{\Lambda}_{\mathcal{I}_M})$ and (7.205) for $q(\mathbf{W}_{\setminus i})$, we can rewrite the equation, as follows:

$$\begin{aligned}
&\log \tilde{q}(\mathbf{W}_i) \\
&\propto \mathbb{E}_{(\mathbf{W}_{\setminus i})}\left[\log \prod_{i' \in \mathcal{I}_M} p(\mathbf{W}_{i'}|\mathbf{W}_{\mathsf{p}(i')})\right] + \mathbb{E}_{(Z,\mathbf{W}_{\setminus i})}\left[\log p(\mathbf{O}, Z|\mathbf{\Lambda}_{\mathcal{I}_M})\right] \\
&\propto \sum_{i' \in \mathcal{I}_M} \mathbb{E}_{(\mathbf{W}_{\setminus i})}\left[\log p(\mathbf{W}_{i'}|\mathbf{W}_{\mathsf{p}(i')})\right] + \mathbb{E}_{(Z,\mathbf{W}_{\setminus i})}\left[\log p(\mathbf{O}, Z|\mathbf{\Lambda}_{\mathcal{I}_M})\right].
\end{aligned} \tag{7.208}$$

Note that the second term depends on two nodes i' and $\mathsf{p}(i')$, and the expectation over i' is not trivial. In this expectation, we can consider the following two cases of variational posterior distributions:

1) Leaf node

We first focus on the initial term of Eq. (7.208). If i is a leaf node, we can disregard the expectation with respect to $\prod_{i' \neq i \in \mathcal{I}_M} q(\mathbf{W}_{i'})$ in the nodes other than the parent node $\mathsf{p}(i)$ of the target leaf node. Thus, we obtain the following simple solution:

$$\log \tilde{q}(\mathbf{W}_i) \propto \mathbb{E}_{(\mathbf{W}_{\mathsf{p}(i)})}\left[\log p(\mathbf{W}_i|\mathbf{W}_{\mathsf{p}(i)})\right] + \mathbb{E}_{(Z,\mathbf{W}_{\setminus i})}\left[\log p(\mathbf{O}, Z|\mathbf{\Lambda}_{\mathcal{I}_M})\right]. \tag{7.209}$$

2) Non-leaf node (with child nodes)

Similarly, if i is a non-leaf node, in addition to the parent node $\mathsf{p}(i)$ of the target node, we also have to consider the child nodes $\mathsf{l}(i)$ and $\mathsf{r}(i)$ of the target node for the expectation, as follows:

$$\log \tilde{q}(\mathbf{W}_i) \propto \quad \mathbb{E}_{(\mathbf{W}_{\mathsf{p}(i)})}\left[\log p(\mathbf{W}_i|\mathbf{W}_{\mathsf{p}(i)})\right] \tag{7.210}$$

$$+ \mathbb{E}_{(\mathbf{W}_{\mathsf{l}(i)})}\left[\log p(\mathbf{W}_{\mathsf{l}(i)}|\mathbf{W}_i)\right] \tag{7.211}$$

$$+ \mathbb{E}_{(\mathbf{W}_{\mathsf{r}(i)})}\left[\log p(\mathbf{W}_{\mathsf{r}(i)}|\mathbf{W}_i)\right] \tag{7.212}$$

$$+ \mathbb{E}_{(Z,\mathbf{W}_{\setminus i})}\left[\log p(\mathbf{O}, Z|\mathbf{\Lambda}_{\mathcal{I}_M})\right]. \tag{7.213}$$

In both cases, the posterior distribution of the transformation matrix in the target node depends on those in the parent and child nodes. Therefore, the posterior distributions are iteratively calculated. This inference is known as a variational message passing algorithm (Winn & Bishop 2006), and Eqs. (7.209)–(7.213) are specific solutions of the variational message passing algorithm to a binary tree structure. The next section provides a concrete form of the posterior distribution of the transformation matrix.

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

Posterior distribution of transformation matrix

We first focus on Eq. (7.213), which is a general form of Eq. (7.209) that has additional terms based on child nodes to Eq. (7.209). Equation (7.213) is based on the expectation with respect to $\prod_{i' \neq i \in \mathcal{I}_M} q(\mathbf{W}_{i'})$ and $q(Z)$. The term with $q(Z)$ is represented as the following expression similar to Eqs. (3.168) and (3.161):

$$\begin{aligned}
&\mathbb{E}_{(Z)}\left[\log p(\mathbf{O}, Z|\boldsymbol{\Lambda}_{\mathcal{I}_M})\right] \\
&= \sum_{k=1}^{K}\sum_{t=1}^{T} \gamma_t(k) \log \mathcal{N}(\mathbf{o}_t|\boldsymbol{\mu}_k^{ad}, \boldsymbol{\Sigma}_k) \\
&= \sum_{i\in\mathcal{I}_M}\left(\sum_{k\in\mathcal{K}_i} \gamma_k \log C_{\mathcal{N}}(\boldsymbol{\Sigma}_k) - \frac{1}{2}\mathrm{tr}\left[\mathbf{W}_i^{\mathsf{T}}\mathbf{W}_i\boldsymbol{\Xi}_i - 2\mathbf{W}_i^{\mathsf{T}}\mathbf{Z}_i + \sum_{k\in\mathcal{K}_i}\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\Gamma}_k\right]\right).
\end{aligned} \tag{7.214}$$

This equation is calculated from the sufficient statistics (γ_k, $\mathbf{S}_k$, $\boldsymbol{\Xi}_i$, and $\mathbf{Z}_i$ in Eqs. (3.169) and (3.167)), that is

$$\begin{cases}
\gamma_k = \sum_{t=1}^{T} \gamma_t(k), \\
\boldsymbol{\gamma}_k = \sum_{t=1}^{T} \gamma_t(k)\mathbf{o}_t, \\
\boldsymbol{\Gamma}_k = \sum_{t=1}^{T} \gamma_t(k)\mathbf{o}_t\mathbf{o}_t^{\mathsf{T}},
\end{cases} \tag{7.215}$$

$$\begin{cases}
\boldsymbol{\Xi}_i \triangleq \sum_{k\in\mathcal{K}_i} \boldsymbol{\xi}_k\boldsymbol{\xi}_k^{\mathsf{T}}\gamma_k, \\
\mathbf{Z}_i \triangleq \sum_{k\in\mathcal{K}_i} (\mathbf{C}_k)^{-1}\boldsymbol{\gamma}_k\boldsymbol{\xi}_k^{\mathsf{T}}.
\end{cases} \tag{7.216}$$

These are computed by the VB-E step (e.g., $\gamma_t(k) = q(v_t = k)$), which is described in the next section. This equation form means that the term can be factorized by node i. This factorization property is important for the following analytic solutions and algorithm. Actually, by considering the expectation with respect to $\prod_{i' \neq i \in \mathcal{I}_M} q(\mathbf{W}_{i'})$, we can integrate out the terms that do not depend on $\mathbf{W}_i$, as follows:

$$\mathbb{E}_{(\mathbf{W}_{\backslash i})}\left[\mathbb{E}_{(Z)}\left[\log p(\mathbf{O}, Z|\boldsymbol{\Lambda}_{\mathcal{I}_M})\right]\right] \propto -\frac{1}{2}\mathrm{tr}\left[\mathbf{W}_i^{\mathsf{T}}\mathbf{W}_i\boldsymbol{\Xi}_i - 2\mathbf{W}_i^{\mathsf{T}}\mathbf{Z}_i\right]. \tag{7.217}$$

Thus, we can obtain the simple quadratic form for this expectation.

Next, we consider Eq. (7.210). Since we use a conjugate prior distribution, $q(\mathbf{W}_{\mathsf{p}(i)})$ is also represented by the following matrix variate Gaussian distribution as the same distribution family with the prior distribution:

$$q(\mathbf{W}_{\mathsf{p}(i)}) = \mathcal{N}(\mathbf{W}_{\mathsf{p}(i)}|\mathbf{M}_{\mathsf{p}(i)}, \mathbf{I}_D, \boldsymbol{\Omega}_{\mathsf{p}(i)}). \tag{7.218}$$

Note that the posterior distribution has a unique form in that the first covariance matrix is an identity matrix while the second is a symmetric matrix. We discuss this form with the analytical solution, later.

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

By substituting Eqs. (7.197) and (7.218) into Eq. (7.210), Eq. (7.210) is represented as follows:

$$\begin{aligned}&\mathbb{E}_{(\mathbf{W}_{\mathsf{p}(i)})}\left[\log p(\mathbf{W}_i|\mathbf{W}_{\mathsf{p}(i)})\right]\\&\quad=\int \mathcal{N}(\mathbf{W}_{\mathsf{p}(i)}|\mathbf{M}_{\mathsf{p}(i)},\mathbf{I}_D,\mathbf{\Omega}_{\mathsf{p}(i)})\log\mathcal{N}(\mathbf{W}_i|\mathbf{W}_{\mathsf{p}(i)},\mathbf{I}_D,\rho_i^{-1}\mathbf{I}_{D+1})d\mathbf{W}_{\mathsf{p}(i)}.\end{aligned}\tag{7.219}$$

To solve the integral, we use the following matrix distribution formula:

$$\begin{aligned}\int \mathcal{N}(\mathbf{W}_{\mathsf{p}(i)}|\mathbf{M}_{\mathsf{p}(i)},\mathbf{I}_D,\mathbf{\Omega}_{\mathsf{p}(i)})d\mathbf{W}_{\mathsf{p}(i)}&=1,\\\int \mathbf{W}_{\mathsf{p}(i)}\mathcal{N}(\mathbf{W}_{\mathsf{p}(i)}|\mathbf{M}_{\mathsf{p}(i)},\mathbf{I}_D,\mathbf{\Omega}_{\mathsf{p}(i)})d\mathbf{W}_{\mathsf{p}(i)}&=\mathbf{M}_{\mathsf{p}(i)}.\end{aligned}\tag{7.220}$$

Then, by using the concrete form of the prior distribution in Eq. (7.201) and by disregarding the terms that do not depend on $\mathbf{W}_i$, Eq. (7.219) can be solved as the logarithmic function of the matrix variate Gaussian distribution that has the posterior distribution parameter $\mathbf{M}_{\mathsf{p}(i)}$ as a hyperparameter:

$$\begin{aligned}&\mathbb{E}_{(\mathbf{W}_{\mathsf{p}(i)})}\left[\log p(\mathbf{W}_i|\mathbf{W}_{\mathsf{p}(i)})\right]\\&\quad\propto \rho_i\int \mathrm{tr}\left[\mathbf{W}_i^{\mathsf{T}}\mathbf{W}_{\mathsf{p}(i)}\right]\mathcal{N}(\mathbf{W}_{\mathsf{p}(i)}|\mathbf{M}_{\mathsf{p}(i)},\mathbf{I}_D,\mathbf{\Omega}_{\mathsf{p}(i)})d\mathbf{W}_{\mathsf{p}(i)}\\&\qquad-\frac{\rho_i}{2}\int \mathrm{tr}\left[\mathbf{W}_i^{\mathsf{T}}\mathbf{W}_i\right]\mathcal{N}(\mathbf{W}_{\mathsf{p}(i)}|\mathbf{M}_{\mathsf{p}(i)},\mathbf{I}_D,\mathbf{\Omega}_{\mathsf{p}(i)})d\mathbf{W}_{\mathsf{p}(i)}\\&\quad\propto \rho_i\mathrm{tr}\left[\mathbf{W}_i^{\mathsf{T}}\mathbf{M}_{\mathsf{p}(i)}\right]-\frac{\rho_i}{2}\mathrm{tr}\left[\mathbf{W}_i^{\mathsf{T}}\mathbf{W}_i\right]\\&\quad\propto\log\mathcal{N}(\mathbf{W}_i|\mathbf{M}_{\mathsf{p}(i)},\mathbf{I}_D,\rho_i^{-1}\mathbf{I}_{D+1}).\end{aligned}\tag{7.221}$$

Similarly, Eqs. (7.211) and (7.212) are solved as follows:

$$\begin{aligned}\mathbb{E}_{(\mathbf{W}_{\mathsf{l}(i)})}\left[\log p(\mathbf{W}_{\mathsf{l}(i)}|\mathbf{W}_i)\right]&\propto\log\mathcal{N}(\mathbf{W}_i|\mathbf{M}_{\mathsf{l}(i)},\mathbf{I}_D,\rho_{\mathsf{l}(i)}^{-1}\mathbf{I}_{D+1}),\\\mathbb{E}_{(\mathbf{W}_{\mathsf{r}(i)})}\left[\log p(\mathbf{W}_{\mathsf{r}(i)}|\mathbf{W}_i)\right]&\propto\log\mathcal{N}(\mathbf{W}_i|\mathbf{M}_{\mathsf{r}(i)},\mathbf{I}_D,\rho_{\mathsf{r}(i)}^{-1}\mathbf{I}_{D+1}).\end{aligned}\tag{7.222}$$

Thus, the expected value terms of the three prior distributions in Eq. (7.210) are represented as the following matrix variate Gaussian distribution:

$$\begin{aligned}&\mathbb{E}_{(\mathbf{W}_{\mathsf{p}(i)})}\left[\log p(\mathbf{W}_i|\mathbf{W}_{\mathsf{p}(i)})\right]+\mathbb{E}_{(\mathbf{W}_{\mathsf{l}(i)})}\left[\log p(\mathbf{W}_{\mathsf{l}(i)}|\mathbf{W}_i)\right]\\&\qquad+\mathbb{E}_{(\mathbf{W}_{\mathsf{r}(i)})}\left[\log p(\mathbf{W}_{\mathsf{r}(i)}|\mathbf{W}_i)\right]\\&\quad\propto\log\mathcal{N}(\mathbf{W}_i|\mathbf{M}_{\mathsf{p}(i)},\mathbf{I}_D,\rho_i^{-1}\mathbf{I}_{D+1})+\log\mathcal{N}(\mathbf{W}_i|\mathbf{M}_{\mathsf{l}(i)},\mathbf{I}_D,\rho_{\mathsf{l}(i)}^{-1}\mathbf{I}_{D+1})\\&\qquad+\log\mathcal{N}(\mathbf{W}_i|\mathbf{M}_{\mathsf{r}(i)},\mathbf{I}_D,\rho_{\mathsf{r}(i)}^{-1}\mathbf{I}_{D+1})\\&\quad\propto\rho_i\mathrm{tr}\left[\mathbf{W}_i^{\mathsf{T}}\mathbf{M}_{\mathsf{p}(i)}\right]-\frac{\rho_i}{2}\mathrm{tr}\left[\mathbf{W}_i^{\mathsf{T}}\mathbf{W}_i\right]+\rho_{\mathsf{l}(i)}\mathrm{tr}\left[\mathbf{W}_i^{\mathsf{T}}\mathbf{M}_{\mathsf{l}(i)}\right]-\frac{\rho_{\mathsf{l}(i)}}{2}\mathrm{tr}\left[\mathbf{W}_i^{\mathsf{T}}\mathbf{W}_i\right]\\&\qquad+\rho_{\mathsf{r}(i)}\mathrm{tr}\left[\mathbf{W}_i^{\mathsf{T}}\mathbf{M}_{\mathsf{r}(i)}\right]-\frac{\rho_{\mathsf{r}(i)}}{2}\mathrm{tr}\left[\mathbf{W}_i^{\mathsf{T}}\mathbf{W}_i\right]\end{aligned}$$

$$\propto -\frac{\rho_i + \rho_{\mathsf{l}(i)} + \rho_{\mathsf{r}(i)}}{2}\mathrm{tr}\left[\mathbf{W}_i^{\mathsf{T}}\mathbf{W}_i\right] + \mathrm{tr}\left[\mathbf{W}_i^{\mathsf{T}}\left(\rho_i\mathbf{M}_{\mathsf{p}(i)} + \rho_{\mathsf{l}(i)}\mathbf{M}_{\mathsf{l}(i)} + \rho_{\mathsf{r}(i)}\mathbf{M}_{\mathsf{r}(i)}\right)\right]$$
$$\propto \log\mathcal{N}\left(\mathbf{W}_i\left|\frac{\rho_i\mathbf{M}_{\mathsf{p}(i)} + \rho_{\mathsf{l}(i)}\mathbf{M}_{\mathsf{l}(i)} + \rho_{\mathsf{r}(i)}\mathbf{M}_{\mathsf{r}(i)}}{\rho_i + \rho_{\mathsf{l}(i)} + \rho_{\mathsf{r}(i)}}, \mathbf{I}_D, (\rho_i + \rho_{\mathsf{l}(i)} + \rho_{\mathsf{r}(i)})^{-1}\mathbf{I}_{D+1}\right.\right). \tag{7.223}$$

It is an intuitive solution, since the location parameter $\mathbf{W}_i$ is represented as a linear interpolation of the location values of the posterior distributions in the parent and child nodes. The precision parameters control the linear interpolation ratio.

Similarly, we can also obtain the expected value term of the prior term in Eq. (7.209), and we summarize the prior terms of the non-leaf and leaf node cases as follows:

$$\hat{q}(\mathbf{W}_i) = \mathcal{N}(\mathbf{W}_i|\hat{\mathbf{M}}_i, \mathbf{I}_D, \hat{\rho}_i^{-1}\mathbf{I}_{D+1}), \tag{7.224}$$

where

$$\hat{\mathbf{M}}_i = \begin{cases} \frac{\rho_i\mathbf{M}_{\mathsf{p}(i)}+\rho_{\mathsf{l}(i)}\mathbf{M}_{\mathsf{l}(i)}+\rho_{\mathsf{r}(i)}\mathbf{M}_{\mathsf{r}(i)}}{\rho_i+\rho_{\mathsf{l}(i)}+\rho_{\mathsf{r}(i)}} & \text{Non-leaf node,} \\ \mathbf{M}_{\mathsf{p}(i)} & \text{Leaf node,} \end{cases}$$
$$\hat{\rho}_i = \begin{cases} \rho_i + \rho_{\mathsf{l}(i)} + \rho_{\mathsf{r}(i)} & \text{Non-leaf node,} \\ \rho_i & \text{Leaf node.} \end{cases} \tag{7.225}$$

Thus, the effect of prior distributions becomes different depending on whether the target node is a non-leaf node or leaf node. The solution is different from our previous solution (Watanabe, Nakamura & Juang 2011), since the previous solution does not marginalize the transformation parameters in non-leaf nodes. In the Bayesian sense, this solution is stricter than the previous solution.

Based on Eqs. (7.214) and (7.224), we can finally derive the quadratic form of $\mathbf{W}_i$ as follows:

$$\log\tilde{q}(\mathbf{W}_i) \propto -\frac{1}{2}\mathrm{tr}\left[\hat{\rho}_i\mathbf{W}_i^{\mathsf{T}}\mathbf{W}_i + \mathbf{W}_i^{\mathsf{T}}\mathbf{W}_i\boldsymbol{\Xi}_i - 2\hat{\rho}_i\mathbf{W}_i^{\mathsf{T}}\hat{\mathbf{M}}_i - 2\mathbf{W}_i^{\mathsf{T}}\mathbf{Z}_i\right]$$
$$\propto -\frac{1}{2}\mathrm{tr}\left[\mathbf{W}_i^{\mathsf{T}}\mathbf{W}_i(\hat{\rho}_i\mathbf{I}_{D+1} + \boldsymbol{\Xi}_i) - 2\mathbf{W}_i^{\mathsf{T}}(\hat{\rho}_i\hat{\mathbf{M}}_i + \mathbf{Z}_i)\right], \tag{7.226}$$

where we disregard the terms that do not depend on $\mathbf{W}_i$. Thus, by defining the following matrix variables:

$$\tilde{\boldsymbol{\Omega}}_i = \left(\hat{\rho}_i\mathbf{I}_{D+1} + \boldsymbol{\Xi}_i\right)^{-1}$$
$$= \begin{cases} \left((\rho_i + \rho_{\mathsf{l}(i)} + \rho_{\mathsf{r}(i)})\mathbf{I}_{D+1} + \boldsymbol{\Xi}_i\right)^{-1} & \text{Non-leaf node,} \\ (\rho_i\mathbf{I}_{D+1} + \boldsymbol{\Xi}_i)^{-1} & \text{Leaf node,} \end{cases}$$
$$\tilde{\mathbf{M}}_i = \left(\hat{\rho}_i\hat{\mathbf{M}}_i + \mathbf{Z}_i\right)\tilde{\boldsymbol{\Omega}}$$
$$= \begin{cases} \left(\rho_i\mathbf{M}_{\mathsf{p}(i)} + \rho_{\mathsf{l}(i)}\mathbf{M}_{\mathsf{l}(i)} + \rho_{\mathsf{r}(i)}\mathbf{M}_{\mathsf{r}(i)} + \mathbf{Z}_i\right)\tilde{\boldsymbol{\Omega}} & \text{Non-leaf node,} \\ \left(\rho_i\mathbf{M}_{\mathsf{p}(i)} + \mathbf{Z}_i\right)\tilde{\boldsymbol{\Omega}} & \text{Leaf node,} \end{cases} \tag{7.227}$$

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

we can derive the posterior distribution of $\mathbf{W}_i$ analytically. The analytical solution is expressed as

$$\begin{aligned}\tilde{q}(\mathbf{W}_i) &= \mathcal{N}(\mathbf{W}_i|\tilde{\mathbf{M}}_i, \mathbf{I}_D, \tilde{\boldsymbol{\Omega}}_i)\\ &= C_{\mathcal{N}}(\mathbf{I}_D, \tilde{\boldsymbol{\Omega}}_i)\exp\left(-\frac{1}{2}\mathrm{tr}\left[(\mathbf{W}_i - \tilde{\mathbf{M}}_i)^{\mathsf{T}}(\mathbf{W}_i - \tilde{\mathbf{M}}_i)\tilde{\boldsymbol{\Omega}}_i^{-1}\right]\right),\end{aligned} \tag{7.228}$$

where

$$C_{\mathcal{N}}(\mathbf{I}_D, \tilde{\boldsymbol{\Omega}}_i) = (2\pi)^{-\frac{D(D+1)}{2}}|\tilde{\boldsymbol{\Omega}}_i|^{-\frac{D}{2}}. \tag{7.229}$$

The posterior distribution also becomes a matrix variate Gaussian distribution, since we use a conjugate prior distribution for $\mathbf{W}_i$. From Eq. (7.227), $\tilde{\mathbf{M}}_i$ are linearly interpolated by hyperparameter $\hat{\mathbf{M}}_i$ and the first-order statistics of the linear regression matrix $\mathbf{Z}_i$. $\hat{\rho}_i$ controls the balance between the effects of the prior distribution and adaptation data. This solution is the M-step of the VB–EM algorithm, and corresponds to that of the ML–EM algorithm in Section 3.5.

Compared with Eq. (7.201), Eq. (7.228) keeps the first covariance matrix as a diagonal matrix, while the second covariance matrix $\tilde{\boldsymbol{\Omega}}$ has off-diagonal elements. This means that the posterior distribution only considers the correlation between column vectors in $\mathbf{W}$. This unique property comes from the variance normalized representation introduced in Section 3.5, which makes multivariate Gaussian distributions in HMMs uncorrelated, and this relationship is taken over to the VB solutions.

Although the solution for a non-leaf node would make the prior distribution robust by taking account of the child node hyperparameters, this structure makes the dependency of the target node on the other linked nodes complex. Therefore, in the implementation step, we approximate the hyperparameters of the posterior distribution for a non-leaf node to those for a leaf node by $\hat{\mathbf{M}}_i \approx \mathbf{M}_{\mathsf{p}(i)}$ and $\hat{\rho}_i \approx \rho_i$ in the Eq. (7.225), and this makes an algorithm simple.

The next section explains the E-step of the VB–EM algorithm, which computes sufficient statistics γ_k, $\boldsymbol{\Gamma}_k$, $\boldsymbol{\Xi}_i$, and $\mathbf{Z}_i$ in Eqs. (3.169) and (3.167). These are obtained by using $\tilde{q}(\mathbf{W}_i)$, of which mode $\tilde{\mathbf{M}}_i$ is used for the Gaussian mean vector transformation.

Posterior distribution of latent variables

From the variational calculation of $\mathcal{F}(M, \Psi)$ with respect to $q(Z)$ based on Eq. (7.25), we also obtain the following posterior distribution:

$$\log \tilde{q}(Z) \propto \mathbb{E}_{(\boldsymbol{\Lambda}_{\mathcal{I}_M})}\left[\log p(\mathbf{O}, Z|\boldsymbol{\Lambda}_{\mathcal{I}_M})\right]. \tag{7.230}$$

By using the factorization form of the variational posterior (Eq. (7.205)), we can disregard the expectation with respect to the variational posteriors other than that of the target node i. Therefore, to obtain the above VB posteriors of latent variables, we have to consider the following integral:

$$\int \tilde{q}(\mathbf{W}_i)\log\mathcal{N}(\mathbf{o}_t|\mathbf{C}_k\mathbf{W}_i\boldsymbol{\xi}_k, \boldsymbol{\Sigma}_k)d\mathbf{W}_i. \tag{7.231}$$

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

Since the Gaussian mean vectors are only updated in the leaf nodes, node i in this section is regarded as a leaf node. By substituting Eqs. (7.228) and (3.162) into Eq. (7.231), the equation can be represented as:

$$\begin{aligned}\int & \tilde{q}(\mathbf{W}_i) \log \mathcal{N}(\mathbf{o}_t | \mathbf{C}_k \mathbf{W}_i \boldsymbol{\xi}_k, \boldsymbol{\Sigma}_k) d\mathbf{W}_i \\ &= \log \mathcal{N}(\mathbf{o}_t | \tilde{\boldsymbol{\mu}}_k, \boldsymbol{\Sigma}_k) - \frac{1}{2} \mathrm{tr}\left[\boldsymbol{\xi}_k \boldsymbol{\xi}_k^\intercal \tilde{\boldsymbol{\Omega}}_i\right]. \end{aligned} \tag{7.232}$$

where

$$\tilde{\boldsymbol{\mu}}_k = \mathbf{C}_k \tilde{\mathbf{M}}_i \boldsymbol{\xi}_k. \tag{7.233}$$

The analytical result is almost equivalent to the E-step of conventional MLLR, which means that the computation time is almost the same as that of the conventional MLLR E-step.

We derive the posterior distribution of latent variables $\tilde{q}(Z)$, introduced in Section 7.4.3, based on the VB framework. In this derivation, we omit indexes i, k, and t for simplicity. By substituting the concrete form (Eq. (3.162)) of the multivariate Gaussian distribution into Eq. (7.231), the equation can be represented as:

$$\begin{aligned}\int & \tilde{q}(\mathbf{W}) \log \mathcal{N}(\mathbf{o} | \mathbf{C}\mathbf{W}\boldsymbol{\xi}, \boldsymbol{\Sigma}) d\mathbf{W} \\ &= -\frac{D}{2} \log(2\pi|\boldsymbol{\Sigma}|) - \frac{1}{2} \int \tilde{q}(\mathbf{W}) \underbrace{\left((\mathbf{o} - \mathbf{C}\mathbf{W}\boldsymbol{\xi})^\intercal \boldsymbol{\Sigma}^{-1} (\mathbf{o} - \mathbf{C}\mathbf{W}\boldsymbol{\xi})\right)}_{(*1)} d\mathbf{W}, \end{aligned} \tag{7.234}$$

where we use the following equation for the normalization term:

$$\int \tilde{q}(\mathbf{W}) d\mathbf{W} = 1. \tag{7.235}$$

Let us now focus on the quadratic form $(*1)$ of Eq. (7.234). By considering $\boldsymbol{\Sigma} = \mathbf{C}(\mathbf{C})^\intercal$ in Eq. (3.164), $(*1)$ can be rewritten as follows:

$$\begin{aligned}(*1) &= (\mathbf{C}^{-1}\mathbf{o} - \mathbf{W}\boldsymbol{\xi})^\intercal (\mathbf{C}^{-1}\mathbf{o} - \mathbf{W}\boldsymbol{\xi}) \\ &= \mathrm{tr}\left[(\mathbf{C}^{-1}\mathbf{o} - \mathbf{W}\boldsymbol{\xi})(\mathbf{C}^{-1}\mathbf{o} - \mathbf{W}\boldsymbol{\xi})^\intercal\right] \\ &= \mathrm{tr}\left[\mathbf{R}\mathbf{W}^\intercal\mathbf{W} - 2\mathbf{W}\mathbf{Y}^\intercal + \mathbf{U}\right], \end{aligned} \tag{7.236}$$

where we use the fact that the trace of the scalar value is equal to the original scalar value and the cyclic property of the trace in Appendix B:

$$a = \mathrm{tr}[a], \tag{7.237}$$

$$\mathrm{tr}[\mathbf{ABC}] = \mathrm{tr}[\mathbf{BCA}]. \tag{7.238}$$

We also define $(D+1) \times (D+1)$ matrix $\mathbf{R}$, $D \times (D+1)$ matrix $\mathbf{Y}$, and $D \times D$ matrix $\mathbf{U}$ in Eq. (7.236) as follows:

$$\begin{aligned}\mathbf{R} &\triangleq \boldsymbol{\xi}\boldsymbol{\xi}^\intercal, \\ \mathbf{Y} &\triangleq \mathbf{C}^{-1}\mathbf{o}\boldsymbol{\xi}^\intercal, \\ \mathbf{U} &\triangleq \boldsymbol{\Sigma}^{-1}\mathbf{o}\mathbf{o}^\intercal. \end{aligned} \tag{7.239}$$

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

The integral of Eq. (7.236) over $\mathbf{W}$ can be decomposed into the following three terms:

$$\begin{aligned}
&\int \tilde{q}(\mathbf{W})\mathrm{tr}\left[\mathbf{R}\mathbf{W}^{\mathsf{T}}\mathbf{W} - 2\mathbf{W}\mathbf{Y}^{\mathsf{T}} + \mathbf{U}\right] d\mathbf{W} \\
&= \underbrace{\int \tilde{q}(\mathbf{W})\mathrm{tr}\left[\mathbf{R}\mathbf{W}^{\mathsf{T}}\mathbf{W}\right] d\mathbf{W}}_{(*2)} - 2\underbrace{\int \tilde{q}(\mathbf{W})\mathrm{tr}\left[\mathbf{W}\mathbf{Y}^{\mathsf{T}}\right] d\mathbf{W}}_{(*3)} + \mathrm{tr}\left[\mathbf{U}\right],
\end{aligned} \tag{7.240}$$

where we use the distributive property of the trace in Appendix B:

$$\mathrm{tr}[\mathbf{A}(\mathbf{B}+\mathbf{C})] = \mathrm{tr}[\mathbf{A}\mathbf{B}+\mathbf{A}\mathbf{C}], \tag{7.241}$$

and use Eq. (7.235) in the third term of the second line in Eq. (7.240).

We focus on the integrals $(*2)$ and $(*3)$. Since $\tilde{q}(\mathbf{W})$ is a scalar value, $(*3)$ can be rewritten as follows:

$$\begin{aligned}
(*3) &= \int \mathrm{tr}\left[\tilde{q}(\mathbf{W})\mathbf{W}\mathbf{Y}^{\mathsf{T}}\right] d\mathbf{W} \\
&= \mathrm{tr}\left[\int \tilde{q}(\mathbf{W})\mathbf{W}\mathbf{Y}^{\mathsf{T}} d\mathbf{W}\right].
\end{aligned} \tag{7.242}$$

Here, we use the following matrix properties:

$$\mathrm{tr}[a\mathbf{A}] = a\,\mathrm{tr}[\mathbf{A}], \tag{7.243}$$

$$\int \mathrm{tr}[f(\mathbf{A})]d\mathbf{A} = \mathrm{tr}\left[\int f(\mathbf{A})d\mathbf{A}\right]. \tag{7.244}$$

Thus, the integral is finally solved as

$$\begin{aligned}
(*3) &= \mathrm{tr}\left[\left(\int \tilde{q}(\mathbf{W})\mathbf{W}d\mathbf{W}\right)\mathbf{Y}^{\mathsf{T}}\right] \\
&= \mathrm{tr}\left[\tilde{\mathbf{M}}\mathbf{Y}^{\mathsf{T}}\right],
\end{aligned} \tag{7.245}$$

where we use

$$\int \tilde{q}(\mathbf{W})\mathbf{W}d\mathbf{W} = \tilde{\mathbf{M}}. \tag{7.246}$$

Similarly, we also rewrite $(*2)$ in Eq. (7.240) based on Eqs. (7.243) and (7.244) as follows:

$$\begin{aligned}
(*2) &= \int \mathrm{tr}\left[\tilde{q}(\mathbf{W})\mathbf{R}\mathbf{W}^{\mathsf{T}}\mathbf{W}\right] d\mathbf{W} \\
&= \mathrm{tr}\left[\int \tilde{q}(\mathbf{W})\mathbf{R}\mathbf{W}^{\mathsf{T}}\mathbf{W}d\mathbf{W}\right] \\
&= \mathrm{tr}\left[\mathbf{R}\int \tilde{q}(\mathbf{W})\mathbf{W}^{\mathsf{T}}\mathbf{W}d\mathbf{W}\right].
\end{aligned} \tag{7.247}$$

Thus, the integral is finally solved as

$$(*2) = \mathrm{tr}\left[\mathbf{R}\left(\tilde{\mathbf{\Omega}} + \tilde{\mathbf{M}}^{\mathsf{T}}\tilde{\mathbf{M}}\right)\right], \tag{7.248}$$

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

where we use

$$\int \tilde{q}(\mathbf{W})\mathbf{W}^{\mathsf{T}}\mathbf{W}d\mathbf{W} = \tilde{\boldsymbol{\Omega}} + \tilde{\mathbf{M}}^{\mathsf{T}}\tilde{\mathbf{M}}. \tag{7.249}$$

Thus, we have solved all the integrals in Eq. (7.240).

Finally, we substitute the integral results of ($*2$) and ($*3$) (i.e., Eqs. (7.248) and (7.245)) into Eq. (7.240), and rewrite Eq. (7.240) based on the concrete forms of $\mathbf{R}$, $\mathbf{Y}$, and $\mathbf{U}$ defined in Eq. (7.239) as follows:

$$\begin{aligned}
&\text{Eq. (7.240)} \\
&\quad = \mathrm{tr}\left[\mathbf{R}\left(\tilde{\boldsymbol{\Omega}} + \tilde{\mathbf{M}}^{\mathsf{T}}\tilde{\mathbf{M}}\right) - 2\tilde{\mathbf{M}}\mathbf{Y}^{\mathsf{T}} + \mathbf{U}\right] \\
&\quad = \mathrm{tr}\left[\boldsymbol{\xi}\boldsymbol{\xi}^{\mathsf{T}}(\tilde{\boldsymbol{\Omega}} + \tilde{\mathbf{M}}^{\mathsf{T}}\tilde{\mathbf{M}}) - 2\tilde{\mathbf{M}}\boldsymbol{\xi}\mathbf{o}^{\mathsf{T}}(\mathbf{C}^{-1})^{\mathsf{T}} + \boldsymbol{\Sigma}^{-1}\mathbf{o}\mathbf{o}^{\mathsf{T}}\right].
\end{aligned} \tag{7.250}$$

Then, by using the cyclic property in Eq. (7.238) and $\boldsymbol{\Sigma} = \mathbf{C}(\mathbf{C})^{\mathsf{T}}$ in Eq. (3.164), we can further rewrite Eq. (7.240) as follows:

$$\begin{aligned}
&\text{Eq. (7.240)} \\
&\quad = \mathrm{tr}\left[\boldsymbol{\xi}\boldsymbol{\xi}^{\mathsf{T}}\tilde{\boldsymbol{\Omega}} + \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\Sigma}\tilde{\mathbf{M}}\boldsymbol{\xi}\boldsymbol{\xi}^{\mathsf{T}}\tilde{\mathbf{M}}^{\mathsf{T}} - 2\mathbf{C}\tilde{\mathbf{M}}\boldsymbol{\xi}\mathbf{o}^{\mathsf{T}} + \mathbf{o}\mathbf{o}^{\mathsf{T}}\right)\right] \\
&\quad = \mathrm{tr}\left[\boldsymbol{\xi}\boldsymbol{\xi}^{\mathsf{T}}\tilde{\boldsymbol{\Omega}} + \boldsymbol{\Sigma}^{-1}\left(\mathbf{o} - \mathbf{C}\tilde{\mathbf{M}}\boldsymbol{\xi}\right)\left(\mathbf{o} - \mathbf{C}\tilde{\mathbf{M}}\boldsymbol{\xi}\right)^{\mathsf{T}}\right].
\end{aligned} \tag{7.251}$$

Thus, we obtain the quadratic form with respect to $\mathbf{o}$, which becomes a multivariate Gaussian distribution form. By recovering the omitted indexes i, k, and t, and substituting the integral result in Eq. (7.251) into Eq. (7.234), we finally solve Eq. (7.231) as:

$$\begin{aligned}
&\int \tilde{q}(\mathbf{W}_i)\log\mathcal{N}(\mathbf{o}_t|\mathbf{C}_k\mathbf{W}_i\boldsymbol{\xi}_k, \boldsymbol{\Sigma}_k)d\mathbf{W}_i \\
&= -\frac{D}{2}\log(2\pi|\boldsymbol{\Sigma}_k|) - \frac{1}{2}\mathrm{tr}\left[\boldsymbol{\xi}_k\boldsymbol{\xi}_k^{\mathsf{T}}\tilde{\boldsymbol{\Omega}}_i + (\boldsymbol{\Sigma}_k)^{-1}\left(\mathbf{o}_t - \mathbf{C}_k\tilde{\mathbf{M}}_i\boldsymbol{\xi}_k\right)\left(\mathbf{o}_t - \mathbf{C}_k\tilde{\mathbf{M}}_i\boldsymbol{\xi}_k\right)^{\mathsf{T}}\right] \\
&= \log\mathcal{N}(\mathbf{o}_t|\mathbf{C}_k\tilde{\mathbf{M}}_i\boldsymbol{\xi}_k, \boldsymbol{\Sigma}_k) - \frac{1}{2}\mathrm{tr}\left[\boldsymbol{\xi}_k\boldsymbol{\xi}_k^{\mathsf{T}}\tilde{\boldsymbol{\Omega}}_i\right].
\end{aligned} \tag{7.252}$$

Here, we use the concrete form of the multivariate Gaussian distribution in Eq. (3.162).

Note that the Gaussian mean vectors are updated in the leaf nodes in this result, while the posterior distributions of the transformation parameters are updated for all nodes.

Variational lower bound

By using the factorization form (Eq. (7.205)) of the variational posterior distribution, the variational lower bound defined in Eq. (7.193) is decomposed as follows:

$$\begin{aligned}
\mathcal{F}(M, \Psi) &= \mathbb{E}_{(Z,\boldsymbol{\Lambda}_{\mathcal{I}_M})}\left[\log\frac{p(\mathbf{O}, Z|\boldsymbol{\Lambda}_{\mathcal{I}_M})p(\boldsymbol{\Lambda}_{\mathcal{I}_M})}{q(Z)\prod_{i\in\mathcal{I}_M}q(\mathbf{W}_i)}\right] \\
&= \underbrace{\mathbb{E}_{(Z,\boldsymbol{\Lambda}_{\mathcal{I}_M})}\left[\log\frac{p(\mathbf{O}, Z|\boldsymbol{\Lambda}_{\mathcal{I}_M})p(\boldsymbol{\Lambda}_{\mathcal{I}_M})}{\prod_{i\in\mathcal{I}_M}q(\mathbf{W}_i)}\right]}_{\triangleq\mathcal{L}(M,\Psi)} - \mathbb{E}_{(Z)}\left[\log q(Z)\right].
\end{aligned} \tag{7.253}$$

The second term, which contains $q(Z)$, is an entropy value and is calculated at the E-step in the VB–EM algorithm. The first term ($\mathcal{L}(M, \Psi)$) is a logarithmic evidence term for M and $\Psi = \{\rho_i | i = 1, \cdots |\mathcal{I}_M|\}$, and we can obtain an analytical solution of $\mathcal{L}(M, \Psi)$. Because of the factorization forms in Eqs. (7.205), (7.197), and (7.214), $\mathcal{L}(M, \Psi)$ can be represented as the summation over i, as follows:

$$\begin{aligned}\mathcal{L}(M, \Psi) &= \mathbb{E}_{(Z, \mathbf{\Lambda}_{\mathcal{I}_M})}\left[\log \frac{\prod_{i \in \mathcal{I}_M} p(\mathbf{O}, Z | \mathbf{W}_i) p(\mathbf{W}_i | \mathbf{W}_{\mathsf{p}(i)})}{\prod_{i \in \mathcal{I}_M} q(\mathbf{W}_i)}\right] \\ &= \sum_{i \in \mathcal{I}_M} \mathcal{L}_i(\rho_i, \rho_{\mathsf{l}(i)}, \rho_{\mathsf{r}(i)}), \end{aligned} \tag{7.254}$$

where

$$\mathcal{L}_i(\rho_i, \rho_{\mathsf{l}(i)}, \rho_{\mathsf{r}(i)}) \triangleq \sum_{i \in \mathcal{I}_M} \mathbb{E}_{(Z, \mathbf{\Lambda}_{\mathcal{I}_M})}\left[\log \frac{p(\mathbf{O}, Z | \mathbf{W}_i) p(\mathbf{W}_i | \mathbf{W}_{\mathsf{p}(i)})}{q(\mathbf{W}_i)}\right]. \tag{7.255}$$

Note that this factorization form has some dependencies from parent and child node parameters through Eqs. (7.225) and (7.227). To derive an analytical solution, we first consider the expectation with respect to $q(Z)$ only for cluster i. By substituting Eqs. (3.168), (7.201), and (7.228) into $\mathcal{L}_i(\rho_i, \rho_{\mathsf{l}(i)}, \rho_{\mathsf{r}(i)})$, the expectation can be rewritten, as follows:

$$\begin{aligned}&\mathbb{E}_{(Z)}\left[\log \frac{p(\mathbf{O}, Z | \mathbf{W}_i) p(\mathbf{W}_i | \mathbf{W}_{\mathsf{p}(i)})}{q(\mathbf{W}_i)}\right] \\ &\quad = \sum_{k \in \mathcal{K}_i} \gamma_k \log C_{\mathcal{N}}(\mathbf{\Sigma}_k) - \frac{1}{2}\mathrm{tr}\left[\mathbf{W}_i^{\mathsf{T}} \mathbf{W}_i \mathbf{\Xi}_i - 2\mathbf{W}_i^{\mathsf{T}} \mathbf{Z}_i + \sum_{k \in \mathcal{K}_i} \mathbf{\Sigma}_k^{-1} \mathbf{\Gamma}_k\right] \\ &\qquad + \log C_{\mathcal{N}}(\mathbf{I}_D, \rho_i^{-1} \mathbf{I}_{D+1}) - \frac{1}{2}\mathrm{tr}\left[\rho_i (\mathbf{W}_i - \mathbf{W}_{\mathsf{p}(i)})^{\mathsf{T}} (\mathbf{W}_i - \mathbf{W}_{\mathsf{p}(i)})\right] \\ &\qquad - \log C_{\mathcal{N}}(\mathbf{I}_D, \tilde{\mathbf{\Omega}}_i) + \frac{1}{2}\mathrm{tr}\left[(\mathbf{W}_i - \tilde{\mathbf{M}}_i)^{\mathsf{T}} (\mathbf{W}_i - \tilde{\mathbf{M}}_i) \tilde{\mathbf{\Omega}}_i^{-1}\right] \\ &\quad = \sum_{k \in \mathcal{K}_i} \gamma_k \log C_{\mathcal{N}}(\mathbf{\Sigma}_k) + \log \frac{C_{\mathcal{N}}(\mathbf{I}_D, \hat{\rho}_i^{-1} \mathbf{I}_{D+1})}{C_{\mathcal{N}}(\mathbf{I}_D, \tilde{\mathbf{\Omega}}_i)} + (*). \end{aligned} \tag{7.256}$$

If we consider only the leaf node case, by using Eq. (7.227), the expectation of $(*)$ part can be rewritten as:

$$\begin{aligned}&(*) \text{ in Eq. (7.256)} \\ &\quad = -\frac{1}{2}\mathrm{tr}\left[\hat{\rho}_i \hat{\mathbf{M}}_i^{\mathsf{T}} \hat{\mathbf{M}}_i - \tilde{\mathbf{M}}_i^{\mathsf{T}} \tilde{\mathbf{M}}_i \tilde{\mathbf{\Omega}}_i^{-1} + \sum_{k \in \mathcal{K}_i} \mathbf{\Sigma}_k^{-1} \mathbf{\Gamma}_k\right]. \end{aligned} \tag{7.257}$$

The result obtained does not depend on $\mathbf{W}_i$. Therefore, the expectation with respect to $q(\mathbf{W}_i)$ can be disregarded in $\mathcal{L}_i(\rho_i, \rho_{\mathsf{l}(i)}, \rho_{\mathsf{r}(i)})$. Consequently, we can obtain the following analytical result for the lower bound:

$$
\begin{aligned}
&\mathcal{L}_i(\rho_i, \rho_{\mathsf{l}(i)}, \rho_{\mathsf{r}(i)}) \\
&= -\frac{D}{2}\log(2\pi)\sum_{k\in\mathcal{K}_i}\gamma_k - \frac{1}{2}\sum_{k\in\mathcal{K}_i}\gamma_k \log|\mathbf{\Sigma}_k| \\
&\quad + \frac{D(D+1)}{2}\log\hat{\rho}_i + \frac{D}{2}\log|\tilde{\mathbf{\Omega}}_i| \\
&\quad - \frac{1}{2}\mathrm{tr}\left[\hat{\rho}_i\hat{\mathbf{M}}_i'\hat{\mathbf{M}}_i - \tilde{\mathbf{M}}_i'\tilde{\mathbf{M}}_i\tilde{\mathbf{\Omega}}_i^{-1} + \sum_{k\in\mathcal{K}_i}\mathbf{\Sigma}_k^{-1}\mathbf{\Gamma}_k\right].
\end{aligned}
\tag{7.258}
$$

The first line of the result obtained corresponds to the likelihood value given the amount of data and the covariance matrices of the Gaussians. The other terms consider the effect of the prior and posterior distributions of the model parameters. This result is used as an optimization criterion with respect to the model structure M and the hyperparameters Ψ.

Note that the objective function can be represented as a summation over i because of the factorization form of the prior and posterior distributions. This representation property is used for our model structure optimization in Section 7.4.5 for a binary tree structure representing a set of Gaussians used in the conventional MLLR.

7.4.4 Optimization of hyperparameters and model structure

In this section, we describe how to optimize hyperparameters Ψ and model structure M by using the variational lower bound as an objective function. Once we obtain the variational lower bound, we can obtain an appropriate model structure and hyperparameters that maximize the lower bound at the same time as follows:

$$
\{\widetilde{\Psi}, \widetilde{M}\} = \arg\max_{M,\Psi}\mathcal{F}(M,\Psi). \tag{7.259}
$$

We use two approximations for the variational lower bound to make the inference algorithm practical. First, we fix latent variables Z during the above optimization. Then, $\mathbb{E}_{(Z)}\left[\log q(Z)\right]$ in Eq. (7.253) is also fixed for M and Ψ, and can be disregarded in the objective function. Thus, we can only focus on $\mathcal{L}(M,\Psi)$ in the optimization step, which reduces computational cost greatly, as follows:

$$
\{\widetilde{\Psi}, \widetilde{M}\} \approx \arg\max_{M,\Psi}\mathcal{L}(M,\Psi). \tag{7.260}
$$

This approximation is widely used in acoustic model selection (likelihood criterion (Odell 1995) and Bayesian criterion (Watanabe *et al.* 2004)). Second, as we discussed in Section 7.4.3, the solution for a non-leaf node (Eq. (7.224)) makes the dependency of the target node on the other linked nodes complex. Therefore, we approximate $\mathcal{L}_i(\rho_i, \rho_{\mathsf{l}(i)}, \rho_{\mathsf{r}(i)}) \approx \mathcal{L}_i(\rho_i)$ by $\hat{\rho}_i \approx \rho_i$ and so on, where $\mathcal{L}_i(\rho_i)$ is defined in the next section. Therefore, in the implementation step, we approximate the posterior distribution for a non-leaf node to that for a leaf node to make the algorithm simple.

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

7.4.5 Hyperparameter optimization

Even though we marginalize all of transformation matrix ($\mathbf{W}_i$), we still have to set the precision hyperparameters ρ_i for all nodes. Since we can derive the variational lower bound, we can optimize the precision hyperparameter, and can remove the manual tuning of the hyperparameters with the proposed approach. This is an advantage of the proposed approach with regard to SMAPLR (Siohan *et al.* 2002), since SMAPLR has to hand-tune its hyperparameters corresponding to $\{\rho_i\}_i$.

Based on the leaf node approximation for variational posterior distributions, in addition to the fixed latent variable approximation ($\mathcal{F}(M, \Psi) \approx \mathcal{L}(M, \Psi)$), in this section the method we implement approximately optimizes the precision hyperparameter as follows:

$$
\begin{aligned}
\tilde{\rho}_i &= \arg\max_{\rho_i} \mathcal{L}(M, \Psi) \\
&= \begin{cases} \arg\max_{\rho_i} \left(\mathcal{L}_i(\rho_i, \rho_{\mathsf{l}(i)}, \rho_{\mathsf{r}(i)}) + \mathcal{L}_{\mathsf{p}(i)}(\rho_{\mathsf{p}(i)}, \rho_i, \rho_{\mathsf{r}(\mathsf{p}(i))})\right) \\ \qquad i \text{ is a left child node of } \mathsf{p}(i) \\ \arg\max_{\rho_i} \left(\mathcal{L}_i(\rho_i, \rho_{\mathsf{l}(i)}, \rho_{\mathsf{r}(i)}) + \mathcal{L}_{\mathsf{p}(i)}(\rho_{\mathsf{p}(i)}, \rho_{\mathsf{l}(\mathsf{p}(i))}, \rho_i)\right) \\ \qquad i \text{ is a right child node of } \mathsf{p}(i) \end{cases} \\
&\approx \arg\max_{\rho_i} \mathcal{L}_i(\rho_i),
\end{aligned} \tag{7.261}
$$

where

$$
\begin{aligned}
\mathcal{L}_i(\rho_i) \triangleq{} & \frac{D(D+1)}{2} \log \rho_i + \frac{D}{2} \log |\tilde{\boldsymbol{\Omega}}_i| \\
& - \frac{1}{2} \mathrm{tr} \left[\rho_i \mathbf{M}_{\mathsf{p}(i)}^{\mathsf{T}} \mathbf{M}_{\mathsf{p}(i)} - \tilde{\mathbf{M}}_i^{\mathsf{T}} \tilde{\mathbf{M}}_i \tilde{\boldsymbol{\Omega}}_i^{-1} \right].
\end{aligned} \tag{7.262}
$$

This approximation makes the algorithm simple because we can optimize the precision hyperparameter within the target and parent nodes, and do not have to consider the child nodes. Since we only have one scalar parameter for this optimization step, we simply used a line search algorithm to obtain the optimal precision hyperparameter. If we consider a more complex precision structure (e.g., a precision matrix instead of a scalar precision parameter in the prior distribution setting Eq. (7.200)), the line search algorithm may not be adequate. In that case, we need to update hyperparameters by using some other optimization technique (e.g., gradient ascent).

Model selection

The remaining tuning parameter in the proposed approach is how many clusters we prepare. This is a model selection problem, and we can also automatically obtain the number of clusters by optimizing the variational lower bound. In the binary tree structure, we focus on a subtree composed of a target non-leaf node i and its child nodes $\mathsf{l}(i)$

Algorithm 13 Structural Bayesian linear regression.

```
Prepare an initial Gaussian tree with a set of nodes $\mathcal{I}$
Initialize $\tilde{\Psi} = \{\tilde{\rho}_i, \tilde{\mathbf{M}}_i | i = 1, \cdots |\mathcal{I}|\}$
repeat
    VB E-step
    $\mathcal{L}(M, \mathbf{\Phi})$ = Prune_tree(root node) // prune a tree by model selection
    # of leaf nodes = Transform_HMM(root node) // Transform HMMs in the pruned tree
until Total lower bound is converged or a specified number of iterations has been reached.
```

and $\mathsf{r}(i)$. We compute the following difference based on Eq. (7.262) of the parent and that of the child nodes:[10]

$$\Delta\mathcal{L}_i(\rho_i) \triangleq \mathcal{L}_{\mathsf{l}(i)}(\rho_{\mathsf{l}(i)}) + \mathcal{L}_{\mathsf{r}(i)}(\rho_{\mathsf{r}(i)}) - \mathcal{L}_i(\rho_i). \tag{7.263}$$

This difference function is used for a stopping criterion in a top-down clustering strategy. This difference function similarly appeared in the model selection of the context-dependent CDHMM topologies in Sections 6.5 and 7.3.6. Then if the sign of $\Delta\mathcal{L}$ is negative, the target non-leaf node is regarded as a new leaf node determined by the model selection in terms of optimizing the lower bound. Next we prune the child nodes $\mathsf{l}(i)$ and $\mathsf{r}(i)$. By checking the signs of $\Delta\mathcal{L}_i$ for all possible nodes, and pruning the child nodes when $\Delta\mathcal{L}_i$ have negative signs, we can obtain the pruned tree structure, which corresponds to maximizing the variational lower bound locally. This optimization is efficiently accomplished by using a depth-first search.

Thus, by optimizing the hyperparameters and model structure, we can avoid setting any tuning parameters. We summarize this optimization in Algorithms 13, 14, and 15. Algorithm 13 prepares a large Gaussian tree with a set of nodes $\mathcal{I}$, prunes a tree based on the model selection (Algorithm 14), and transforms HMMs (Algorithm 15). Algorithm 14 first optimizes the precision hyperparameters Ψ, and then the model structure M. Algorithm 15 transforms Gaussian mean vectors in HMMs at the new root nodes in the pruned tree $\mathcal{I}_M$ obtained by Algorithm 14.

Watanabe *et al.* (2013) compare the VB linear regression method (VBLR) with MLLR and SMAPLR, as regards the *WSJ*, for various amounts of adaptation data by using LVCSR experiments for the Corpus of Spontaneous Japanese (CSJ). With a small amount of adaptation data, VBLR outperforms the conventional approaches by about 1.0% absolute accuracy improvement, while with a large amount of adaptation data, the accuracies of all approaches are comparable. This property is theoretically reasonable

[10] Since we approximate the posterior distribution for a non-leaf node to that for a leaf node, the contribution of the variational lower bounds from the non-leaf nodes to the total lower bounds can be disregarded, and Eq. (7.263) is used as a pruning criterion. If we do not use this approximation, we just compare the difference between the values $\mathcal{L}_i(\rho_i, \rho_{\mathsf{l}(i)}, \rho_{\mathsf{r}(i)})$ of the leaf and non-leaf node cases in Eq. (7.258).

Algorithm 14 Prune_tree(node *i*)

```
if First iteration then
    $\tilde{\rho}_i = \arg\max_{\rho_i} \mathcal{L}_i(\rho_i)$ // These are used as
    Update $\tilde{q}(\mathbf{W}_i)$   // hyperparameters of parent nodes
end if
if Node $i$ has child nodes then
    $\tilde{\rho}_i = \arg\max_{\rho_i} \mathcal{L}_i(\rho_i)$
    Update $\tilde{q}(\mathbf{W}_i)$
    $\Delta\mathcal{L} =$ Prune_tree(node $left(i)$) + Prune_tree(node $right(i)$) $-\mathcal{L}_i(\tilde{\rho}_i)$
    if $\Delta\mathcal{L} < 0$ then
        Prune child nodes // this node becomes a leaf node
    end if
    return $\mathcal{L}_i(\tilde{\rho}_i)$
else
    $\tilde{\rho}_i = \arg\max_{\rho_i} \mathcal{L}_i(\rho_i)$
    Update $q(\mathbf{W}_i)$
    return $\mathcal{L}_i(\tilde{\rho}_i)$
end if
```

Algorithm 15 Transform_HMM(node *i*)

```
if Node $i$ has child nodes then
    return Transform_HMM(node $left(i)$) + Transform_HMM(node $right(i)$)
else
    Update $\tilde{\boldsymbol{\mu}}_k = C_k \tilde{\mathbf{M}}_i \boldsymbol{\xi}_k$
    return 1
end if
```

since the variational lower bound would be tighter than the EM-based objective function for a small amount of data, while the lower bound would approach it for a large amount of data asymptotically. Therefore, it can be concluded that this improvement comes from the optimization of the hyperparameters and the model structure in VBLR, in addition to mitigation of the sparse data problem arising in the Bayesian approach.

7.5 Variational Bayesian speaker verification

This section describes an application of VB to speaker verification (Zhao, Dong, Zhao *et al*. 2009, Kenny 2010, Villalba & Brümmer 2011). The main goal of this approach is to obtain the feature representation that only holds speaker specific characteristics. As discussed in Section 4.6.2, state-of-the-art speaker verification systems use the super vector obtained by the GMM–UBM or MLLR techniques, as a feature. However, the super vector still includes various factors other than the speaker characteristics with very high dimensional representation. Use of factor analysis is critical in speaker verification to remove these irrelevant factors of the super vector and find the lower dimensional representation (Kenny 2010, Kinnunen & Li 2010, Dehak *et al.* 2011). In addition, a Bayesian treatment of the factor analysis yields robust modeling of the speaker verification. This section discusses a VB treatment of the factor analysis model by providing a generative model of the super vector (Section 7.5.1), prior distributions (Section 7.5.2), variational posteriors (Section 7.5.3), and variational lower bound (Section 7.5.4).

7.5.1 Generative model

Let $\mathbf{O} = \{\mathbf{o}_n \in \mathbb{R}^D | n = 1, \cdots, N\}$ be a D dimensional feature vector of n recordings. Note that $\mathbf{o}_n$ is a super vector, and it can be the Gaussian super vector, vectorized form of the MLLR matrix, or the factor vector obtained after the initial factor analysis process.[11] If we use the Gaussian super vector, the number of dimensions D would be the product of multiplying the number of mixture components in GMM (usually $K = 1024$) and the number of speech feature dimensions (usually $D_{\text{MFCC}} = 39$ when we use MFCC and delta features) when we use GMM–UBM, that is

$$\mathbf{o}_n = [\boldsymbol{\mu}_1^\mathsf{T}, \cdots, \boldsymbol{\mu}_k^\mathsf{T}, \cdots, \boldsymbol{\mu}_K^\mathsf{T}]^\mathsf{T}, \quad \boldsymbol{\mu}_k \in \mathbb{R}^{D_{\text{MFCC}}}. \tag{7.264}$$

Therefore, D would be $1024 \times 39 \approx 40$ thousands, and it is much larger than the number of speech feature dimensions that we are dealing with at CDHMMs. Note also that the feature $\mathbf{o}_n$ is obtained for each recording (utterance), and the frame level process is performed when super vectors are extracted by GMM–UBM.

The generative model is represented as follows (Kenny, Boulianne, Ouellet *et al.* 2007, Kenny 2010):

$$\mathbf{o}_n = \mathbf{m} + \mathbf{U}_1\mathbf{x}_1 + \mathbf{U}_2\mathbf{x}_{2n} + \boldsymbol{\varepsilon}_n, \tag{7.265}$$

where $\mathbf{m} \in \mathbb{R}^D$ is a global mean vector for the feature vectors and can be regarded as a bias vector of the feature vectors. Vector $\mathbf{x}_1 \in \mathbb{R}^{D_1}$ is a vector having a D_1 dimensional standard Gaussian distribution, which does not depend on the recording n, and it can represent stationary speaker characteristics across recordings. On the other hand, $\mathbf{x}_{2n} \in \mathbb{R}^{D_2}$ is a vector having a D_2 dimensional standard Gaussian distribution depending on the recording n, and it denotes channel characteristics changing over a recording. We also define $\mathbf{X}_2 \triangleq \{\mathbf{x}_{2n} | n = 1, \cdots, N\}$. $\boldsymbol{\varepsilon}_n \in \mathbb{R}^D$ as a D dimensional vector having a Gaussian distribution with $\mathbf{0}$ mean vector, and $\mathbf{R} \in \mathbb{R}^{D \times D}$ precision matrix.

In this book, $\mathbf{m}$, $\mathbf{U}_1 \in \mathbb{R}^{D \times D_1}$, $\mathbf{U}_2 \in \mathbb{R}^{D \times D_2}$, and $\mathbf{R}$ ($\Psi = \{\mathbf{m}, \mathbf{U}_1, \mathbf{U}_2, \mathbf{R}\}$) are regarded as non-probabilistic model parameters and assumed to be estimated without

[11] A Bayesian treatment of the factor analysis in a state-of-the-art speaker verification system can be performed after the first step of the factor analysis, called *i* vector analysis (Dehak *et al.* 2011). That is, Bayesian factor analysis is often performed for the first-step factor vector (*i* vector) for each utterance, instead of the Gaussian super vector. This book discusses Bayesian factor analysis for the Gaussian super vector for simplicity.

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

the Bayesian framework based on ML/MAP. However, Villalba & Brümmer (2011) deal with these parameters as probabilistic variables, and provide a fully Bayesian solution of the factor analysis based speaker modeling by VB. The other probabilistic variables $\mathbf{x}_1$, $\mathbf{x}_{2n}$, and $\boldsymbol{\varepsilon}_n$ are generated from the following Gaussian distributions:

$$\begin{aligned}\mathbf{x}_1 &\sim \mathcal{N}(\mathbf{0}, u_1^{-1}\mathbf{I}_{D_1}),\\ \mathbf{x}_{2n} &\sim \mathcal{N}(\mathbf{0}, u_{2n}^{-1}\mathbf{I}_{D_2}),\\ \boldsymbol{\varepsilon}_n &\sim \mathcal{N}(\mathbf{0}, v_n^{-1}\mathbf{R}^{-1}),\end{aligned} \tag{7.266}$$

where we assume a zero mean spherical Gaussian distribution for $\mathbf{x}_1$ and $\mathbf{x}_{2n}$. The model parameters $u_1 \in \mathbb{R}_{>0}$, $u_{2n} \in \mathbb{R}_{>0}$, and $v_n \in \mathbb{R}_{>0}$ are positive, and the probabilistic treatment of these parameters is discussed later.

In summary, the conditional distribution of $\mathbf{O}$ is represented as follows:

$$p(\mathbf{O}|\mathbf{x}_1, \mathbf{X}_2, \{v_n\}_{n=1}^N, \Psi) = \prod_{n=1}^N \mathcal{N}(\mathbf{o}_n|\mathbf{m} + \mathbf{U}_1\mathbf{x}_1 + \mathbf{U}_2\mathbf{x}_{2n}, v_n^{-1}\mathbf{R}^{-1}). \tag{7.267}$$

The conditional joint distribution of $\mathbf{O}$, $\mathbf{x}_1$, and $\mathbf{X}_2$ is represented as follows:

$$\begin{aligned}&p(\mathbf{O}, \mathbf{x}_1, \mathbf{X}_2|\{v_n\}_{n=1}^N, \Psi, u_1, \{u_{2n}\}_{n=1}^N)\\ &\quad = p(\mathbf{O}|\mathbf{x}_1, \mathbf{X}_2, \{v_n\}_{n=1}^N, \Psi)p(\mathbf{x}_1|u_1)\prod_{n=1}^N p(\mathbf{x}_{2n}|u_{2n}),\end{aligned} \tag{7.268}$$

where

$$\begin{aligned}p(\mathbf{x}_1|u_1) &= \mathcal{N}(\mathbf{x}_1|\mathbf{0}, u_1^{-1}\mathbf{I}_{D_1}),\\ p(\mathbf{x}_{2n}|u_{2n}) &= \mathcal{N}(\mathbf{x}_{2n}|\mathbf{0}, u_{2n}^{-1}\mathbf{I}_{D_2}).\end{aligned} \tag{7.269}$$

The following section regards u_1, u_{2n}, and v_n as probabilistic variables.

7.5.2 Prior distributions

We provide the conjugate prior distributions for u_1, u_{2n}, and v_n that are represented by a gamma distribution, as we discussed in Section 2.1.3. Kenny (2010) provides a simple hyperparameter setting for each gamma distribution in Appendix C.11 by using only one hyperparameter for each distribution, i.e., the model parameters u_1, u_{2n}, and v_n are generated from the following prior distributions:

$$\begin{aligned}u_1 &\sim \mathrm{Gam}\left(\frac{\phi_1}{2}, \frac{\phi_1}{2}\right),\\ u_{2n} &\sim \mathrm{Gam}\left(\frac{\phi_2}{2}, \frac{\phi_2}{2}\right),\\ v_n &\sim \mathrm{Gam}\left(\frac{\phi_v}{2}, \frac{\phi_v}{2}\right),\end{aligned} \tag{7.270}$$

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

where ϕ_1, ϕ_2, and ϕ_v ($\triangleq$ Φ) are hyperparameters in this model. Since the mean and variance of the gamma distribution $\mathrm{Gam}(y|\alpha,\beta)$ are $\frac{\alpha}{\beta}$ and $\frac{\alpha}{\beta^2}$, respectively, this parameterization means that u_1, u_{2n}, and v_n have the same mean value with 1, but the variance values are changed with $\frac{2}{\phi_1}$, $\frac{2}{\phi_2}$, and $\frac{2}{\phi_v}$, respectively. Thus, we can provide the following concrete forms of the prior distributions for u_1, u_{2n}, and v_n:

$$\begin{aligned} p(u_1|\phi_1) &= \mathrm{Gam}\left(u_1 \middle| \frac{\phi_1}{2}, \frac{\phi_1}{2}\right), \\ p(u_{2n}|\phi_2) &= \mathrm{Gam}\left(u_{2n} \middle| \frac{\phi_2}{2}, \frac{\phi_2}{2}\right), \\ p(v_n|\phi_v) &= \mathrm{Gam}\left(v_n \middle| \frac{\phi_v}{2}, \frac{\phi_v}{2}\right). \end{aligned} \tag{7.271}$$

In this model, $Z \triangleq \{\mathbf{x}_1, u_1, \{\mathbf{x}_{2n}, u_{2n}, v_n\}_{n=1}^N\}$ is a set of hidden variables, and the posterior distribution of each variable can be obtained by using variational Bayes. Thus, the joint prior distribution given hyperparameters Φ is represented as follows:

$$\begin{aligned} p(Z|\Phi) &= p(\mathbf{x}_1, u_1, \{\mathbf{x}_{2n}, u_{2n}, v_n\}_{n=1}^N|\Phi) \\ &= p(\mathbf{x}_1|u_1)p(u_1|\phi_1)\prod_{n=1}^N p(\mathbf{x}_{2n}|u_{2n})p(u_{2n}|\phi_2)p(v_n|\phi_v). \end{aligned} \tag{7.272}$$

Note that $\mathbf{x}_1$ depends on u_1, and we cannot fully factorize them. A similar discussion applies to $\mathbf{x}_{2n}$ and u_{2n}.

Now, we can provide the complete data likelihood function given hyperparameters Ψ and Φ based on Eqs. (7.267) and (7.272), and this can be used to obtain the variational posteriors:

$$\begin{aligned} &p(\mathbf{O}, Z|\Psi, \Phi) \\ &\quad = p(\mathbf{O}|\Psi, Z)p(Z|\Phi) \\ &\quad = \mathcal{N}(\mathbf{x}_1|\mathbf{0}, u_1^{-1}\mathbf{I}_{D_1})\mathrm{Gam}\left(u_1 \middle| \frac{\phi_1}{2}, \frac{\phi_1}{2}\right) \\ &\qquad \times \prod_{n=1}^N \mathcal{N}(\mathbf{o}_n|\mathbf{m} + \mathbf{U}_1\mathbf{x}_1 + \mathbf{U}_2\mathbf{x}_{2n}, v_n^{-1}\mathbf{R}^{-1})\mathcal{N}(\mathbf{x}_{2n}|\mathbf{0}, u_{2n}^{-1}\mathbf{I}_{D_2}) \\ &\qquad \times \mathrm{Gam}\left(u_{2n} \middle| \frac{\phi_2}{2}, \frac{\phi_2}{2}\right)\mathrm{Gam}\left(v_n \middle| \frac{\phi_v}{2}, \frac{\phi_v}{2}\right). \end{aligned} \tag{7.273}$$

These probabilistic distributions are represented by Gaussian and gamma distributions. In the following sections we simplify the complete data likelihood function $p(\mathbf{O}, Z|\Psi, \boldsymbol{\Phi})$ to $p(\mathbf{O}, Z)$ to avoid complicated equations. Algorithm 16 provides a generative process for the joint factor analysis speaker model with Eq. (7.273).

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

Algorithm 16 Generative process for joint factor analysis speaker model

```
Require: $\Psi = \{\mathbf{m}, \mathbf{U}_1, \mathbf{U}_2, \mathbf{R}\}$ and $\Phi = \{\phi_1, \phi_2, \phi_v\}$
Draw $u_1$ from $\mathrm{Gam}\left(u_1 \middle| \frac{\phi_1}{2}, \frac{\phi_1}{2}\right)$
Draw $\mathbf{x}_1$ from $\mathcal{N}(\mathbf{x}_1|\mathbf{0}, u_1^{-1}\mathbf{I}_{D_1})$
for $n = 1, \cdots, N$ do
    Draw $u_{2n}$ from $\mathrm{Gam}\left(u_{2n} \middle| \frac{\phi_2}{2}, \frac{\phi_2}{2}\right)$
    Draw $v_n$ from $\mathrm{Gam}\left(v_n \middle| \frac{\phi_v}{2}, \frac{\phi_v}{2}\right)$
    Draw $\mathbf{x}_{2n}$ from $\mathcal{N}(\mathbf{x}_{2n}|\mathbf{0}, u_{2n}^{-1}\mathbf{I}_{D_2})$
    Draw $\mathbf{o}_n$ from $\mathcal{N}(\mathbf{o}_n|\mathbf{m} + \mathbf{U}_1\mathbf{x}_1 + \mathbf{U}_2\mathbf{x}_{2n}, v_n^{-1}\mathbf{R}^{-1})$
end for
```

7.5.3 Variational posteriors

To deal with the variational posteriors, we assume the following factorization based on the VB recipe:

$$q(Z|\mathbf{O}) = q(\mathbf{x}_1|\mathbf{O})q(u_1|\mathbf{O}) \prod_{n=1}^{N} q(\mathbf{x}_{2n}|\mathbf{O})q(u_{2n}|\mathbf{O})q(v_n|\mathbf{O}). \tag{7.274}$$

Section 7.1.3 discusses how we can obtain the following general solution for approximated variational posteriors:

$$\widetilde{q}(Z_i|\mathbf{O}) \propto \exp\left(\mathbb{E}_{(Z_{\backslash i}|\mathbf{O})}\left[\log p(\mathbf{O}, Z)\right]\right). \tag{7.275}$$

We focus on the actual solutions for $q(\mathbf{x}_1|\mathbf{O})$, $q(u_1|\mathbf{O})$, $q(\mathbf{x}_{2n})$, $q(u_{2n}|\mathbf{O})$, and $q(v_n|\mathbf{O})$.

- $q(\mathbf{x}_1|\mathbf{O})$: this is calculated by substituting the factors depending on $\mathbf{x}_1$ in Eq. (7.273) into Eq. (7.275) as follows:

$$\begin{aligned}
&\log q(\mathbf{x}_1|\mathbf{O}) \\
&\quad \propto \mathbb{E}_{(Z_{\backslash \mathbf{x}_1})}[\log p(\mathbf{O}, Z)] \\
&\quad \propto \mathbb{E}_{(Z_{\backslash \mathbf{x}_1})}\left[\log\left(\mathcal{N}(\mathbf{x}_1|\mathbf{0}, u_1^{-1}\mathbf{I}_{D_1}) \prod_{n=1}^{N} \mathcal{N}(\mathbf{o}_n|\mathbf{m} + \mathbf{U}_1\mathbf{x}_1\right.\right. \\
&\qquad\qquad\qquad\qquad\qquad\qquad \left.\left. + \, \mathbf{U}_2\mathbf{x}_{2n}, v_n^{-1}\mathbf{R}^{-1})\right)\right] \\
&\quad \propto \mathbb{E}_{(u_1)}[\log \mathcal{N}(\mathbf{x}_1|\mathbf{0}, u_1^{-1}\mathbf{I}_{D_1})] \\
&\qquad + \sum_{n=1}^{N} \mathbb{E}_{(\mathbf{x}_{2n}, v_n)}[\log \mathcal{N}(\mathbf{o}_n|\mathbf{m} + \mathbf{U}_1\mathbf{x}_1 + \mathbf{U}_2\mathbf{x}_{2n}, v_n^{-1}\mathbf{R}^{-1})].
\end{aligned} \tag{7.276}$$

Now let us consider the two expectations in the above equation. From the definition of the multivariate Gaussian distribution in Appendix C.6, we can obtain the following equation by disregarding the terms that do not depend on $\mathbf{x}_1$:

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

$$\begin{aligned}\mathbb{E}_{(u_1)}[\log \mathcal{N}(\mathbf{x}_1|\mathbf{0}, u_1^{-1}\mathbf{I}_{D_1})] &\propto \mathbb{E}_{(u_1)}\left[-\frac{u_1}{2}\mathbf{x}_1^{\mathsf{T}}\mathbf{x}_1\right] \\ &\propto -\frac{\mathbb{E}\,[u_1]}{2}\mathbf{x}_1^{\mathsf{T}}\mathbf{x}_1,\end{aligned} \tag{7.277}$$

where we omit the subscript (u_1) in the expectation, as it is trivial. Similarly, the rest of the expectations in Eq. (7.276) are also represented as follows:

$$\begin{aligned}&\mathbb{E}_{(\mathbf{x}_{2n},v_n)}\left[\log \mathcal{N}(\mathbf{o}_n|\mathbf{m}+\mathbf{U}_1\mathbf{x}_1+\mathbf{U}_2\mathbf{x}_{2n}, v_n^{-1}\mathbf{R}^{-1})\right] \\ &\propto \mathbb{E}_{(\mathbf{x}_{2n},v_n)}\left[-\frac{v_n}{2}(\mathbf{o}_n-(\mathbf{m}+\mathbf{U}_1\mathbf{x}_1+\mathbf{U}_2\mathbf{x}_{2n}))^{\mathsf{T}}\right. \\ &\qquad\qquad \left.\times\mathbf{R}(\mathbf{o}_n-(\mathbf{m}+\mathbf{U}_1\mathbf{x}_1+\mathbf{U}_2\mathbf{x}_{2n}))\right] \\ &\propto \mathbb{E}_{(\mathbf{x}_{2n},v_n)}\left[-\frac{v_n}{2}\left(\mathbf{x}_1^{\mathsf{T}}\mathbf{U}_1^{\mathsf{T}}\mathbf{R}\mathbf{U}_1\mathbf{x}_1\right)+v_n(\mathbf{o}_n-\mathbf{m}-\mathbf{U}_2\mathbf{x}_{2n})^{\mathsf{T}}\mathbf{R}\mathbf{U}_1\mathbf{x}_1\right] \\ &\propto -\frac{\mathbb{E}[v_n]}{2}\mathbf{x}_1^{\mathsf{T}}\mathbf{U}_1^{\mathsf{T}}\mathbf{R}\mathbf{U}_1\mathbf{x}_1+\mathbb{E}[v_n](\mathbf{o}_n-\mathbf{m}-\mathbf{U}_2\mathbb{E}[\mathbf{x}_{2n}])^{\mathsf{T}}\mathbf{R}\mathbf{U}_1\mathbf{x}_1.\end{aligned} \tag{7.278}$$

Thus, by substituting Eqs. (7.277) and (7.278) into Eq. (7.276), we find that

$$\begin{aligned}&\log q(\mathbf{x}_1|\mathbf{O}) \\ &\propto -\frac{\mathbb{E}\,[u_1]}{2}\mathbf{x}_1^{\mathsf{T}}\mathbf{x}_1+\sum_{n=1}^{N}-\frac{\mathbb{E}[v_n]}{2}\mathbf{x}_1^{\mathsf{T}}\mathbf{U}_1^{\mathsf{T}}\mathbf{R}\mathbf{U}_1\mathbf{x}_1 \\ &\quad+\sum_{n=1}^{N}\mathbb{E}[v_n](\mathbf{o}_n-\mathbf{m}-\mathbf{U}_2\mathbb{E}[\mathbf{x}_{2n}])^{\mathsf{T}}\mathbf{R}\mathbf{U}_1\mathbf{x}_1 \\ &\propto -\frac{1}{2}\mathrm{tr}\left[\left(\mathbb{E}\,[u_1]\,\mathbf{I}_{D_1}+\sum_{n=1}^{N}\mathbb{E}[v_n]\mathbf{U}_1^{\mathsf{T}}\mathbf{R}\mathbf{U}_1\right)\mathbf{x}_1\mathbf{x}_1^{\mathsf{T}}\right] \\ &\quad+\mathrm{tr}\left[\mathbf{x}_1^{\mathsf{T}}\sum_{n=1}^{N}\mathbb{E}[v_n]\mathbf{U}_1^{\mathsf{T}}\mathbf{R}(\mathbf{o}_n-\mathbf{U}_2\mathbb{E}[\mathbf{x}_{2n}]-\mathbf{m})\right] \\ &\propto \log \mathcal{N}(\mathbf{x}_1|\tilde{\boldsymbol{\mu}}_{\mathbf{x}_1}, \tilde{\boldsymbol{\Sigma}}_{\mathbf{x}_1}),\end{aligned} \tag{7.279}$$

where we use the trace form definition of the multivariate Gaussian distribution in Appendix C.6. Thus, $q(\mathbf{x}_1|\mathbf{O})$ is represented as a Gaussian distribution, and $\tilde{\boldsymbol{\mu}}_{\mathbf{x}_1}$ and $\tilde{\boldsymbol{\Sigma}}_{\mathbf{x}_1}$ are posterior hyperparameters obtained as follows:

$$\begin{aligned}\tilde{\boldsymbol{\mu}}_{\mathbf{x}_1} &\triangleq \left(\mathbb{E}[u_1]\mathbf{I}_{D_1}+\sum_{n=1}^{N}\mathbb{E}[v_n]\mathbf{U}_1^{\mathsf{T}}\mathbf{R}\mathbf{U}_1\right)^{-1} \\ &\quad\times\sum_{n=1}^{N}\mathbb{E}[v_n]\mathbf{U}_1^{\mathsf{T}}\mathbf{R}(\mathbf{o}_n-\mathbf{U}_2\mathbb{E}[\mathbf{x}_2]-\mathbf{m}) \\ \tilde{\boldsymbol{\Sigma}}_{\mathbf{x}_1} &\triangleq \left(\mathbb{E}[u_1]\mathbf{I}_{D_1}+\sum_{n=1}^{N}\mathbb{E}[v_n]\mathbf{U}_1^{\mathsf{T}}\mathbf{R}\mathbf{U}_1\right)^{-1}.\end{aligned} \tag{7.280}$$

Thus, the hyperparameters obtained are represented with prior hyperparameters Ψ and the expected values of $\mathbb{E}[u_1]$ and $\mathbb{E}[v_n]$.

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

- $q(\mathbf{x}_{2n}|\mathbf{O})$: this is similarly calculated by substituting the factors depending on $\mathbf{x}_{2n}$ in Eq. (7.273) into Eq. (7.275) as follows:

$$\begin{aligned}
&\log q(\mathbf{x}_{2n}|\mathbf{O}) \\
&\quad\propto \mathbb{E}_{(Z_{\backslash \mathbf{x}_{2n}})}[\log p(\mathbf{O}, Z)] \\
&\quad\propto \mathbb{E}_{(Z_{\backslash \mathbf{x}_{2n}})}\Bigg[\prod_{n'=1}^{N} \log \big(\mathcal{N}(\mathbf{o}_{n'}|\mathbf{m} + \mathbf{U}_1\mathbf{x}_1 \\
&\qquad\qquad +\mathbf{U}_2\mathbf{x}_{2n'}, v_{n'}^{-1}\mathbf{R}^{-1})\mathcal{N}(\mathbf{x}_{2n'}|\mathbf{0}, u_{2n'}^{-1}\mathbf{I}_{D_2})\big)\Bigg] \\
&\quad\propto \mathbb{E}_{(\mathbf{x}_1, v_n)}\left[\mathcal{N}(\mathbf{o}_n|\mathbf{m} + \mathbf{U}_1\mathbf{x}_1 + \mathbf{U}_2\mathbf{x}_{2n}, v_n^{-1}\mathbf{R}^{-1})\right] \\
&\qquad + \mathbb{E}_{(u_{2n})}\left[\mathcal{N}(\mathbf{x}_{2n}|\mathbf{0}, u_{2n}^{-1}\mathbf{I}_{D_2})\right].
\end{aligned} \tag{7.281}$$

The expectations are rewritten as follows:

$$\begin{aligned}
&\log q(\mathbf{x}_{2n}|\mathbf{O}) \\
&\quad\propto -\frac{1}{2}\mathbb{E}[v_n]\mathrm{tr}\left[\mathbf{U}_2^{\mathsf{T}}\mathbf{R}\mathbf{U}_2\mathbf{x}_{2n}\mathbf{x}_{2n}^{\mathsf{T}}\right] \\
&\qquad + 2\mathbb{E}[v_n]\mathrm{tr}\left[\mathbf{x}_{2n}^{\mathsf{T}}\mathbf{U}_2^{\mathsf{T}}\mathbf{R}(\mathbf{o}_n - \mathbf{U}_2\mathbb{E}[\mathbf{x}_1] - \mathbf{m})\right] \\
&\qquad - \frac{1}{2}\mathbb{E}[u_{2n}]\mathrm{tr}\left[\mathbf{x}_{2n}\mathbf{x}_{2n}^{\mathsf{T}}\right] \\
&\quad\propto -\frac{1}{2}\mathrm{tr}\left[\left(\mathbb{E}[u_{2n}]\mathbf{I}_{D_2} + \mathbb{E}[v_n]\mathbf{U}_2^{\mathsf{T}}\mathbf{R}\mathbf{U}_2\right)\mathbf{x}_{2n}\mathbf{x}_{2n}^{\mathsf{T}}\right] \\
&\qquad + \mathrm{tr}\left[\mathbf{x}_{2n}^{\mathsf{T}}\mathbb{E}[v_n]\mathbf{U}_2^{\mathsf{T}}\mathbf{R}(\mathbf{o}_n - \mathbf{U}_2\mathbb{E}[\mathbf{x}_1] - \mathbf{m})\right] \\
&\quad\propto \log\mathcal{N}(\mathbf{x}_{2n}|\tilde{\boldsymbol{\mu}}_{\mathbf{x}_{2n}}, \tilde{\boldsymbol{\Sigma}}_{\mathbf{x}_{2n}}).
\end{aligned} \tag{7.282}$$

Thus, $q(\mathbf{x}_{2n}|\mathbf{O})$ is also represented as a Gaussian distribution, and $\tilde{\boldsymbol{\mu}}_{\mathbf{x}_{2n}}$ and $\tilde{\boldsymbol{\Sigma}}_{\mathbf{x}_{2n}}$ are posterior hyperparameters obtained as follows:

$$\begin{aligned}
\tilde{\boldsymbol{\mu}}_{\mathbf{x}_{2n}} &\triangleq \left(\mathbb{E}[u_{2n}]\mathbf{I}_{D_2} + \mathbb{E}[v_n]\mathbf{U}_2^{\mathsf{T}}\mathbf{R}\mathbf{U}_2\right)^{-1}\mathbb{E}[v_n]\mathbf{U}_2^{\mathsf{T}}\mathbf{R}(\mathbf{o}_n - \mathbf{U}_2\mathbb{E}[\mathbf{x}_1] - \mathbf{m}), \\
\tilde{\boldsymbol{\Sigma}}_{\mathbf{x}_{2n}} &\triangleq \left(\mathbb{E}[u_{2n}]\mathbf{I}_{D_2} + \mathbb{E}[v_n]\mathbf{U}_2^{\mathsf{T}}\mathbf{R}\mathbf{U}_2\right)^{-1}.
\end{aligned} \tag{7.283}$$

Note that the hyperparameters obtained are represented with prior hyperparameters Ψ and the expected values of $\mathbb{E}[u_{2n}]$, $\mathbb{E}[v_n]$, and $\mathbb{E}[\mathbf{x}_1]$. Compared with Eq. (7.280), Eq. (7.283) has a similar functional form, but it is computed recording by recording, while Eq. (7.280) is computed with accumulation over every recording n.

- $q(u_1|\mathbf{O})$: this is also similarly calculated by substituting the factors depending on u_1 in Eq. (7.273) into Eq. (7.275). However, compared with the previous two cases, the gamma and Gaussian distributions appear in the formulation as follows:

$$\begin{aligned}
&\log q(u_1|\mathbf{O}) \\
&\quad\propto \mathbb{E}_{(Z_{\backslash u_1})}[\log p(\mathbf{O}, Z)] \\
&\quad\propto \mathbb{E}_{(Z_{\backslash u_1})}\left[\mathcal{N}(\mathbf{x}_1|\mathbf{0}, u_1^{-1}\mathbf{I}_{D_1})\mathrm{Gam}\left(u_1\middle|\frac{\phi_1}{2}, \frac{\phi_1}{2}\right)\right] \\
&\quad\propto \mathbb{E}_{(\mathbf{x}_1)}[\log\mathcal{N}(\mathbf{x}_1|\mathbf{0}, u_1^{-1}\mathbf{I}_{D_1})] + \log\left(\mathrm{Gam}\left(u_1\middle|\frac{\phi_1}{2}, \frac{\phi_1}{2}\right)\right).
\end{aligned} \tag{7.284}$$

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

By using the definition of the gamma distribution in Appendix C.11, this equation can be rewritten as follows:

$$\begin{aligned}
&\log q(u_1|\mathbf{O}) \\
&\propto \log\left(|u_1 \mathbf{I}_{D_1}|^{\frac{1}{2}}\right) - \frac{\mathbb{E}[\mathbf{x}_1^{\mathsf{T}}\mathbf{x}_1]}{2} u_1 + \log\left(u_1^{\frac{\phi_1}{2}-1}\right) - \frac{\phi_1}{2} u_1 \\
&\propto \log\left(u_1^{\frac{\phi_1+D_1}{2}-1}\right) - \frac{\phi_1 + \mathbb{E}[\mathbf{x}_1^{\mathsf{T}}\mathbf{x}_1]}{2} u_1 \\
&\propto \log\left(\mathrm{Gam}\left(u_1 \middle| \frac{\tilde{\phi}_1}{2}, \frac{\tilde{r}_1}{2}\right)\right).
\end{aligned} \tag{7.285}$$

Thus, $q(u_1|\mathbf{O})$ is represented as a gamma distribution, and $\tilde{\phi}_1$ and $\tilde{r}_1$ are posterior hyperparameters obtained as follows:

$$\begin{aligned}
\tilde{\phi}_1 &\triangleq \phi_1 + D_1, \\
\tilde{r}_1 &\triangleq \phi_1 + \mathbb{E}[\mathbf{x}_1^{\mathsf{T}}\mathbf{x}_1].
\end{aligned} \tag{7.286}$$

These posterior hyperparameters are obtained with their original prior hyperparameter ϕ_1 and the second-order expectation value of $\mathbb{E}[\mathbf{x}_1^{\mathsf{T}}\mathbf{x}_1]$.

- $q(u_{2n}|\mathbf{O})$: this is similarly calculated by substituting the factors depending on u_{2n} in Eq. (7.273) into Eq. (7.275):

$$\begin{aligned}
&\log q(u_{2n}|\mathbf{O}) \\
&\propto \mathbb{E}_{(Z_{\backslash u_{2n}})}[\log p(\mathbf{O}, Z)] \\
&\propto \mathbb{E}_{(Z_{\backslash u_{2n}})}\left[\prod_{n'=1}^{N} \mathcal{N}(\mathbf{x}_{2n'}|\mathbf{0}, u_{2n'}^{-1}\mathbf{I}_{D_2})\mathrm{Gam}\left(u_{2n'} \middle| \frac{\phi_2}{2}, \frac{\phi_2}{2}\right)\right] \\
&\propto \mathbb{E}_{(\mathbf{x}_{2n})}[\log \mathcal{N}(\mathbf{x}_{2n}|\mathbf{0}, u_{2n}^{-1}\mathbf{I}_{D_2})] + \log\left(\mathrm{Gam}\left(u_{2n} \middle| \frac{\phi_2}{2}, \frac{\phi_2}{2}\right)\right) \\
&\propto \log\left(\mathrm{Gam}\left(u_{2n} \middle| \frac{\tilde{r}_{2n}}{2}, \frac{\tilde{\phi}_{2n}}{2}\right)\right).
\end{aligned} \tag{7.287}$$

$q(u_{2n}|\mathbf{O})$ is represented as a gamma distribution, and $\tilde{\phi}_{2n}$ and $\tilde{r}_{2n}$ are posterior hyperparameters, obtained as follows:

$$\begin{aligned}
\tilde{\phi}_{2n} &\triangleq \phi_2 + D_2, \\
\tilde{r}_{2n} &\triangleq \phi_2 + \mathbb{E}[\mathbf{x}_{2n}^{\mathsf{T}}\mathbf{x}_{2n}].
\end{aligned} \tag{7.288}$$

These posterior hyperparameters are also obtained with their original prior hyperparameter ϕ_2 and the second-order expectation value of $\mathbb{E}[\mathbf{x}_{2n}^{\mathsf{T}}\mathbf{x}_{2n}]$, and these are very similar to the posterior hyperparameters of $q(u_1|\mathbf{O})$ in Eq. (7.286).

- $q(v_n|\mathbf{O})$: finally, this is also calculated by substituting the factors depending on v_n in Eq. (7.273) into Eq. (7.275) as follows:

$$\begin{aligned}
&\log q(v_n|\mathbf{O}) \\
&\propto \mathbb{E}_{(Z_{\backslash v_n})}[\log p(\mathbf{O}, Z)]
\end{aligned}$$

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

$$\begin{aligned}
&\propto \mathbb{E}_{(Z_{\backslash v_n})}\left[\prod_{n'=1}^{N}\mathcal{N}(\mathbf{o}_{n'}|\mathbf{m}+\mathbf{U}_1\mathbf{x}_1+\mathbf{U}_2\mathbf{x}_{2n'}, v_{n'}^{-1}\mathbf{R}^{-1})\right.\\
&\qquad\qquad\qquad\left.\times \mathrm{Gam}\left(v_{n'}\middle|\frac{\phi_v}{2},\frac{\phi_v}{2}\right)\right]\\
&\propto \mathbb{E}_{(\mathbf{x}_1,\mathbf{x}_{2n})}\left[\mathcal{N}(\mathbf{o}_n|\mathbf{m}+\mathbf{U}_1\mathbf{x}_1+\mathbf{U}_2\mathbf{x}_{2n}, v_n^{-1}\mathbf{R}^{-1})\right]\\
&\quad+\log\left(\mathrm{Gam}\left(v_n\middle|\frac{\phi_v}{2},\frac{\phi_v}{2}\right)\right)\\
&\propto \log\left(\mathrm{Gam}\left(v_n\middle|\frac{\tilde{r}_{vn}}{2},\frac{\tilde{\phi}_{vn}}{2}\right)\right).
\end{aligned}\tag{7.289}$$

$q(v_n|\mathbf{O})$ is represented as a gamma distribution, and $\tilde{\phi}_{vn}$ and $\tilde{r}_{vn}$ are posterior hyperparameters obtained as follows:

$$\begin{aligned}
\tilde{\phi}_{vn} &\triangleq \phi_v + D,\\
\tilde{r}_{vn} &\triangleq \phi_v + \mathbb{E}[\boldsymbol{\varepsilon}_n^{\mathsf{T}}\mathbf{R}\boldsymbol{\varepsilon}_n],
\end{aligned}\tag{7.290}$$

where $\boldsymbol{\varepsilon}_n$ is a residual vector appearing in the basic equation of the joint factor analysis in Eq. (7.265), and is represented as:

$$\boldsymbol{\varepsilon}_n = \mathbf{o}_n - (\mathbf{m}+\mathbf{U}_1\mathbf{x}_1+\mathbf{U}_2\mathbf{x}_{2n})\,. \tag{7.291}$$

$\mathbb{E}[\boldsymbol{\varepsilon}_n^{\mathsf{T}}\mathbf{R}\boldsymbol{\varepsilon}_n]$ is an expectation over both $\mathbf{x}_1$ and $\mathbf{x}_{2n}$, and defined as follows:

$$\begin{aligned}
&\mathbb{E}[\boldsymbol{\varepsilon}_n^{\mathsf{T}}\mathbf{R}\boldsymbol{\varepsilon}_n]\\
&\quad\triangleq \mathbb{E}_{\mathbf{x}_1,\mathbf{x}_{2n}}[(\mathbf{o}_n-\mathbf{m}-\mathbf{U}_1\mathbf{x}_1-\mathbf{U}_2\mathbf{x}_{2n})^{\mathsf{T}}\mathbf{R}(\mathbf{o}_n-\mathbf{m}-\mathbf{U}_1\mathbf{x}_1-\mathbf{U}_2\mathbf{x}_{2n})]\\
&\quad= (\mathbf{o}_n-\mathbf{m})^{\mathsf{T}}\mathbf{R}(\mathbf{o}_n-\mathbf{m}) + \mathrm{tr}\left[\mathbf{U}_1^{\mathsf{T}}\mathbf{R}\mathbf{U}_1\mathbb{E}\left[\mathbf{x}_1\mathbf{x}_1^{\mathsf{T}}\right]\right] + \mathrm{tr}\left[\mathbf{U}_2^{\mathsf{T}}\mathbf{R}\mathbf{U}_2\mathbb{E}\left[\mathbf{x}_{2n}\mathbf{x}_{2n}^{\mathsf{T}}\right]\right]\\
&\qquad- 2(\mathbf{o}_n-\mathbf{m})^{\mathsf{T}}\mathbf{R}\mathbf{U}_1\mathbb{E}\left[\mathbf{x}_1\right] - 2(\mathbf{o}_n-\mathbf{m})^{\mathsf{T}}\mathbf{R}\mathbf{U}_2\mathbb{E}\left[\mathbf{x}_{2n}\right]\\
&\qquad+ 2\mathrm{tr}\left[\mathbf{U}_1^{\mathsf{T}}\mathbf{R}\mathbf{U}_2\mathbb{E}\left[\mathbf{x}_{2n}\right]\mathbb{E}\left[\mathbf{x}_1^{\mathsf{T}}\right]\right].
\end{aligned}\tag{7.292}$$

This value is computed by the first- and second-order expectation of $\mathbf{x}_1$ and $\mathbf{x}_{2n}$.

Thus, we can provide the VB posterior distributions of all hidden variables analytically. Note that these equations are iteratively performed to obtain the sub-optimal posterior distributions. That is, all the posterior distribution calculations need the expectation values of hidden variables Z, which can be computed using the posterior distributions obtained with the previous iteration. We provide the expectation values of $\mathbf{x}_1$ and $\mathbf{x}_{2n}$, which are easily obtained by reference to the expectation formulas of a multivariate Gaussian distribution in Appendix C.6 and posterior hyperparameters in Eqs. (7.280) and (7.283):

$$\begin{aligned}
\mathbb{E}\left[\mathbf{x}_1\right] &= \int \mathbf{x}_1 q(\mathbf{x}_1|\mathbf{O})d\mathbf{x}_1 = \tilde{\boldsymbol{\mu}}_{\mathbf{x}_1},\\
\mathbb{E}\left[\mathbf{x}_1\mathbf{x}_1^{\mathsf{T}}\right] &= \int \mathbf{x}_1\mathbf{x}_1^{\mathsf{T}} q(\mathbf{x}_1|\mathbf{O})d\mathbf{x}_1 = \tilde{\boldsymbol{\Sigma}}_{\mathbf{x}_1},
\end{aligned}$$

$$
\begin{aligned}
\mathbb{E}\left[\mathbf{x}_{2n}\right] &= \int \mathbf{x}_{2n} q(\mathbf{x}_{2n}|\mathbf{O}) d\mathbf{x}_{2n} = \widetilde{\boldsymbol{\mu}}_{\mathbf{x}_{2n}}, \\
\mathbb{E}\left[\mathbf{x}_{2n}\mathbf{x}_{2n}^{\mathsf{T}}\right] &= \int \mathbf{x}_{2n}\mathbf{x}_{2n}^{\mathsf{T}} q(\mathbf{x}_{2n}|\mathbf{O}) d\mathbf{x}_{2n} = \widetilde{\boldsymbol{\Sigma}}_{\mathbf{x}_{2n}}.
\end{aligned}
\tag{7.293}
$$

Similarly, we provide the expectation values of u_1, u_{2n}, and v_n, which are also easily obtained by reference to the expectation formulas of gamma distribution in Appendix C.11 and posterior hyperparameters in Eqs. (7.286), (7.288), and (7.290):

$$
\begin{aligned}
\mathbb{E}\left[u_1\right] &= \int u_1 q(u_1|\mathbf{O}) du_1 = \frac{\widetilde{\phi}_1}{\widetilde{r}_1}, \\
\mathbb{E}\left[u_{2n}\right] &= \int u_{2n} q(u_{2n}|\mathbf{O}) du_{2n} = \frac{\widetilde{\phi}_{2n}}{\widetilde{r}_{2n}}, \\
\mathbb{E}\left[v_n\right] &= \int v_n q(v_n|\mathbf{O}) dv_n = \frac{\widetilde{\phi}_{vn}}{\widetilde{r}_{vn}}.
\end{aligned}
\tag{7.294}
$$

Finally, we summarize the analytical results of the posterior distributions $q(Z|\mathbf{O})$, as follows:

$$
\begin{aligned}
& q(Z|\mathbf{O}) \\
&= q(\mathbf{x}_1|\mathbf{O}) q(u_1|\mathbf{O}) \prod_{n=1}^{N} q(\mathbf{x}_{2n}|\mathbf{O}) q(u_{2n}|\mathbf{O}) q(v_n|\mathbf{O}) \\
&= \mathcal{N}(\mathbf{x}_1|\widetilde{\boldsymbol{\mu}}_{\mathbf{x}_1}, \widetilde{\boldsymbol{\Sigma}}_{\mathbf{x}_1}) \mathrm{Gam}\left(u_1 \middle| \frac{\tilde{\phi}_1}{2}, \frac{\tilde{r}_1}{2}\right) \\
&\quad \times \prod_{n=1}^{N} \mathcal{N}(\mathbf{x}_{2n}|\widetilde{\boldsymbol{\mu}}_{\mathbf{x}_{2n}}, \widetilde{\boldsymbol{\Sigma}}_{\mathbf{x}_{2n}}) \mathrm{Gam}\left(u_{2n} \middle| \frac{\tilde{r}_{2n}}{2}, \frac{\tilde{\phi}_{2n}}{2}\right) \mathrm{Gam}\left(v_n \middle| \frac{\tilde{r}_{vn}}{2}, \frac{\tilde{\phi}_{vn}}{2}\right),
\end{aligned}
\tag{7.295}
$$

where posterior hyperparameters are represented as:

$$
\begin{cases}
\widetilde{\boldsymbol{\mu}}_{\mathbf{x}_1} & \triangleq \left(\mathbb{E}[u_1]\mathbf{I}_{D_1} + \sum_{n=1}^{N} \mathbb{E}[v_n]\mathbf{U}_1^{\mathsf{T}}\mathbf{R}\mathbf{U}_1\right)^{-1}, \\
& \qquad \times \sum_{n=1}^{N} \mathbb{E}[v_n]\mathbf{U}_1^{\mathsf{T}}\mathbf{R}(\mathbf{o}_n - \mathbf{U}_2\mathbb{E}[\mathbf{x}_2] - \mathbf{m}), \\
\widetilde{\boldsymbol{\Sigma}}_{\mathbf{x}_1} & \triangleq \left(\mathbb{E}[u_1]\mathbf{I}_{D_1} + \sum_{n=1}^{N} \mathbb{E}[v_n]\mathbf{U}_1^{\mathsf{T}}\mathbf{R}\mathbf{U}_1\right)^{-1}, \\
\widetilde{\boldsymbol{\mu}}_{\mathbf{x}_{2n}} & \triangleq \left(\mathbb{E}[u_{2n}]\mathbf{I}_{D_2} + \mathbb{E}[v_n]\mathbf{U}_2^{\mathsf{T}}\mathbf{R}\mathbf{U}_2\right)^{-1} \mathbb{E}[v_n]\mathbf{U}_2^{\mathsf{T}}\mathbf{R}(\mathbf{o}_n - \mathbf{U}_2\mathbb{E}[\mathbf{x}_1] - \mathbf{m}), \\
\widetilde{\boldsymbol{\Sigma}}_{\mathbf{x}_{2n}} & \triangleq \left(\mathbb{E}[u_{2n}]\mathbf{I}_{D_2} + \mathbb{E}[v_n]\mathbf{U}_2^{\mathsf{T}}\mathbf{R}\mathbf{U}_2\right)^{-1},
\end{cases}
\tag{7.296}
$$

$$
\begin{cases}
\tilde{\phi}_1 & \triangleq \phi_1 + D_1, \\
\tilde{r}_1 & \triangleq \phi_1 + \mathbb{E}[\mathbf{x}_1^{\mathsf{T}}\mathbf{x}_1], \\
\tilde{\phi}_{2n} & \triangleq \phi_2 + D_2, \\
\tilde{r}_{2n} & \triangleq \phi_2 + \mathbb{E}[\mathbf{x}_{2n}^{\mathsf{T}}\mathbf{x}_{2n}], \\
\tilde{\phi}_{vn} & \triangleq \phi_v + D, \\
\tilde{r}_{vn} & \triangleq \phi_v + \mathbb{E}[\boldsymbol{\varepsilon}_n^{\mathsf{T}}\mathbf{R}\boldsymbol{\varepsilon}_n].
\end{cases}
\tag{7.297}
$$

Once we obtain the VB posterior distributions, we can also calculate the variational lower bound, which is discussed in the next section.

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

7.5.4 Variational lower bound

As discussed in Section 4.6, speaker verification can be performed by using the likelihood ratio (Eq. (4.129)):

$$\frac{p(\mathbf{O}|H_0)}{p(\mathbf{O}|H_1)}, \tag{7.298}$$

where H_0 means that $\mathbf{O}$ is from the hypothesized speaker, while H_1 means that $\mathbf{O}$ is *not* from the hypothesized speaker. Instead of using the likelihood of $p(\mathbf{O})$ where we neglect the hypothesis index H, using the lower bound, we can treat speaker verification in a Bayesian sense. That is, we use the variational lower bound equation (in Eq. (7.5)):

$$p(\mathbf{O}) = \int \log p(\mathbf{O}, Z) dZ \geq \mathcal{F}[q(Z|\mathbf{O})]. \tag{7.299}$$

Then we use $\mathcal{F}[q(Z|\mathbf{O})]$ instead of $p(\mathbf{O})$. From the definition of the variational lower bound in Eq. (7.4), Eq. (7.299) can be decomposed into the following terms as follows:

$$\begin{aligned}\mathcal{F}[q(Z|\mathbf{O})] &\triangleq \mathbb{E}_{(Z)}\left[\log \frac{p(\mathbf{O}, Z)}{q(Z|\mathbf{O})}\right] = \mathbb{E}_{(Z)}\left[\log \frac{p(\mathbf{O}|Z)p(Z)}{q(Z|\mathbf{O})}\right] \\ &= \mathbb{E}_{(Z)}\left[\log p(\mathbf{O}|Z)\right] - \mathrm{KL}(q(Z|\mathbf{O})\|p(Z)),\end{aligned} \tag{7.300}$$

where the second term is the Kullback–Leibler divergence between the variational posterior and prior distributions.

Let us focus on the first term of the variational lower bound. The conditional likelihood $p(\mathbf{O}|Z)$ is represented as the following Gaussian by using Eq. (7.267):

$$p(\mathbf{O}|Z) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{o}_n|\mathbf{m} + \mathbf{U}_1\mathbf{x}_1 + \mathbf{U}_2\mathbf{x}_{2n}, \nu_n^{-1}\mathbf{R}^{-1}). \tag{7.301}$$

By using Eq. (7.301), $\mathbb{E}_{(Z)}\left[\log p(\mathbf{O}|Z)\right]$ is represented as follows:

$$\begin{aligned}&\mathbb{E}_{(Z)}\left[\log p(\mathbf{O}|Z)\right] \\ &\quad= \mathbb{E}_{(Z)}\left[\log \prod_{n=1}^{N} \mathcal{N}(\mathbf{o}_n|\mathbf{m} + \mathbf{U}_1\mathbf{x}_1 + \mathbf{U}_2\mathbf{x}_{2n}, \nu_n^{-1}\mathbf{R}^{-1})\right] \\ &\quad= \sum_{n=1}^{N} \mathbb{E}_{(\mathbf{x}_1, \mathbf{x}_{2n}, \nu_n)}\left[\log \mathcal{N}(\mathbf{o}_n|\mathbf{m} + \mathbf{U}_1\mathbf{x}_1 + \mathbf{U}_2\mathbf{x}_{2n}, \nu_n^{-1}\mathbf{R}^{-1})\right].\end{aligned} \tag{7.302}$$

This expectation is calculated by using the definition of multivariate Gaussian distribution in Appendix C.6 and the residual vector $\boldsymbol{\varepsilon}_n$ in Eq. (7.291), as follows:

$$\begin{aligned}&\mathbb{E}_{(\mathbf{x}_1, \mathbf{x}_{2n}, \nu_n)}\left[\log \mathcal{N}(\mathbf{o}_n|\mathbf{m} + \mathbf{U}_1\mathbf{x}_1 + \mathbf{U}_2\mathbf{x}_{2n}, \nu_n^{-1}\mathbf{R}^{-1})\right] \\ &\quad= \mathbb{E}_{(\mathbf{x}_1, \mathbf{x}_{2n}, \nu_n)}\left[\log C_{\mathcal{N}}(\nu_n^{-1}\mathbf{R}^{-1})) - \frac{\nu_n}{2}\boldsymbol{\varepsilon}_n^{\mathsf{T}}\mathbf{R}\boldsymbol{\varepsilon}_n\right] \\ &\quad= \frac{D}{2}\mathbb{E}[\log \nu_n] - \frac{D}{2}\log(2\pi) + \frac{1}{2}\log|\mathbf{R}| - \frac{1}{2}\mathbb{E}[\nu_n]\mathbb{E}[\boldsymbol{\varepsilon}_n^{\mathsf{T}}\mathbf{R}\boldsymbol{\varepsilon}_n].\end{aligned} \tag{7.303}$$

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

Therefore, Eq. (7.302) is calculated as:

$$\mathbb{E}_{(Z)}\left[\log p(\mathbf{O}|Z)\right] = \sum_{n=1}^{N}\left(\frac{D}{2}\mathbb{E}[\log v_n] - \frac{D}{2}\log(2\pi) + \frac{1}{2}\log|\mathbf{R}| - \frac{1}{2}\mathbb{E}[v_n]\mathbb{E}[\boldsymbol{\varepsilon}_n^{\mathsf{T}}\mathbf{R}\boldsymbol{\varepsilon}_n]\right). \tag{7.304}$$

Now, let us focus on the KL divergence in Eq. (7.301), which is decomposed into the following terms based on the factorization forms of Eqs. (7.272) and (7.274):

$$\mathrm{KL}(q(Z|\mathbf{O})\|p(Z)) = \mathrm{KL}(q(\mathbf{x}_1, u_1|\mathbf{O})\|p(\mathbf{x}_1, u_1)) + \sum_{n=1}^{N}\mathrm{KL}(q(\mathbf{x}_{2n}, u_{2n}|\mathbf{O})\|p(\mathbf{x}_{2n}, u_{2n})) + \sum_{n=1}^{N}\mathrm{KL}(q(v_n|\mathbf{O})\|p(v_n)). \tag{7.305}$$

The KL divergence is decomposed into the three KL divergence terms. Now, consider $\mathrm{KL}(q(\mathbf{x}_1, u_1|\mathbf{O})\|p(\mathbf{x}_1, u_1))$, which can be further factorized as follows:

$$\mathrm{KL}(q(\mathbf{x}_1, u_1|\mathbf{O})\|p(\mathbf{x}_1, u_1)) = \mathbb{E}_{q(u_1|\mathbf{O})}\left[\mathrm{KL}(q(\mathbf{x}_1|\mathbf{O})\|p(\mathbf{x}_1|u_1))\right] + \mathrm{KL}(q(u_1|\mathbf{O})\|p(u_1)). \tag{7.306}$$

Thus, we need to compute the KL divergences of Gaussian and gamma distributions, respectively. We use the following formulas, which are the analytical result of the KL divergence between Gaussian and gamma distributions:

$$\mathrm{KL}(\mathcal{N}(\mathbf{x}|\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})\|\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})) = -\frac{D}{2} - \frac{1}{2}\log|\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{\Sigma}}| + \mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\Sigma} + (\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu})(\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu})^{\mathsf{T}}\right)\right], \tag{7.307}$$

and

$$\mathrm{KL}(\mathrm{Gam}(y|\tilde{\alpha}, \tilde{\beta})\|\mathrm{Gam}(y|\alpha, \beta)) = \log\frac{\Gamma(\alpha)}{\Gamma(\tilde{\alpha})} + \tilde{\alpha}\log\tilde{\beta} - \alpha\log\beta + (\tilde{\alpha} - \alpha)(\psi(\tilde{\alpha}) - \log\tilde{\beta}) + \tilde{\alpha}\frac{\beta - \tilde{\beta}}{\tilde{\beta}}. \tag{7.308}$$

Therefore, by using Eq. (7.307), $\mathbb{E}_{q(u_1|\mathbf{O})}\left[\mathrm{KL}(q(\mathbf{x}_1|\mathbf{O})\|p(\mathbf{x}_1|u_1))\right]$ can be rewritten as follows:

$$\begin{aligned}
&\mathbb{E}_{q(u_1|\mathbf{O})}\left[\mathrm{KL}(q(\mathbf{x}_1|\mathbf{O})\|p(\mathbf{x}_1|u_1))\right] \\
&= \mathbb{E}_{q(u_1|\mathbf{O})}\left[\mathrm{KL}(\mathcal{N}(\mathbf{x}_1|\tilde{\boldsymbol{\mu}}_{\mathbf{x}_1}, \tilde{\boldsymbol{\Sigma}}_{\mathbf{x}_1}\|\mathcal{N}(\mathbf{x}_1|\mathbf{0}, u_1^{-1}\mathbf{I}_{D_1}))\right] \\
&= \mathbb{E}_{q(u_1|\mathbf{O})}\left[-\frac{D_1}{2} - \frac{1}{2}\log|u_1\tilde{\boldsymbol{\Sigma}}_{\mathbf{x}_1}| + \mathrm{tr}\left[u_1\left(u_1^{-1}\mathbf{I}_{D_1} + (\tilde{\boldsymbol{\mu}}_{\mathbf{x}_1} - \boldsymbol{\mu})(\tilde{\boldsymbol{\mu}}_{\mathbf{x}_1} - \boldsymbol{\mu})^{\mathsf{T}}\right)\right]\right] \\
&= -\frac{D_1}{2} - \frac{D_1}{2}\mathbb{E}[\log u_1] - \frac{1}{2}\log|\tilde{\boldsymbol{\Sigma}}_{\mathbf{x}_1}| + \frac{1}{2}\mathbb{E}[\log u_1]\mathbb{E}[\mathbf{x}_1^{\mathsf{T}}\mathbf{x}_1].
\end{aligned} \tag{7.309}$$

Similarly, other terms can be obtained by using VB analytically.

Thus, we can obtain the variational lower bound for the joint factor analysis. This can be used as an objective function of the likelihood test, as discussed before, and is also used to optimize the hyperparameters Ψ and Φ based on the evidence approximation, as discussed in Chapter 5.

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

7.6 Latent Dirichlet allocation

Latent Dirichlet allocation (LDA) (Blei *et al.* 2003) is known as the popular machine learning approach which was proposed to build the *latent topic model* for information retrieval, document modeling, text categorization, and collaborative filtering. LDA has been successfully applied for image modeling, music retrieval, speech recognition, and many others. In general, LDA is an extended study from a topic model based on probabilistic latent semantic analysis (PLSA) (Hofmann 1999*b*, Hofmann 2001), which was addressed in Section 3.7.3. PLSA extracts the latent semantic information and estimates the topic parameters according to maximum likelihood (ML) estimation, which suffers from the over-trained problem. In addition, PLSA could not represent the unseen words and documents. The number of PLSA parameters increases remarkably with the number of collected documents. To compensate these weaknesses, LDA incorporates the Dirichlet priors to characterize the topic mixture probabilities. The marginal likelihood over all possible values of topic mixture probabilities is calculated and maximized so as to construct the LDA-based topic model for document representation. Unseen documents are generalized by using the LDA parameters, which are estimated according to the variational Bayes inference procedure. The model complexity is controlled as the training documents become larger. PLSA and LDA extract the topic information at document level, and this could be combined in a language model for speech recognition. In what follows, we first address the construction of an LDA model from a set of training documents. The optimization objective is formulated for model training. Then the variational Bayes (VB) inference is detailed. The variational distribution over multiple latent variables is introduced to find a VB solution to the LDA model.

7.6.1 Model construction

LDA provides a powerful mechanism to discover latent topic structure from a bag of M documents with a bag of words $\mathbf{w} = \{w_1, \cdots, w_N\}$, where $w_n \in \mathcal{V}$. The text corpus is denoted by $\mathcal{D} = \{\mathbf{w}_1, \cdots, \mathbf{w}_M\}$. Figure 7.2 is a graphical representation of LDA. Using LDA, each of the n words w_n is represented by a multinomial distribution conditioned on the topic z_n:

$$w_n|z_n, \boldsymbol{\beta} \sim \text{Multi}(\boldsymbol{\beta}), \tag{7.310}$$

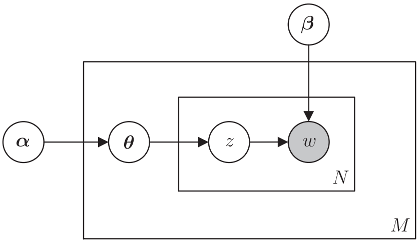

Figure 7.2 Representation of latent Dirichlet allocation.

where $\boldsymbol{\beta} \in \mathbb{R}^{|\mathcal{V}| \times K}$ denotes the multinomial parameters. The topic z_n of word w_n is also generated by a multinomial distribution given parameters $\boldsymbol{\theta} \in \mathbb{R}^K$:

$$z_n|\boldsymbol{\theta} \sim \text{Multi}(\boldsymbol{\theta}), \tag{7.311}$$

where $\theta_k \geq 0$, $\sum_{k=1}^K \theta_k = 1$ and K topics are assumed. Importantly, the latent topics of each document are treated as random variables. We assume that the multinomial parameters $\boldsymbol{\theta}$ of K topics are drawn from a Dirichlet distribution,

$$\begin{aligned}\boldsymbol{\theta}|\boldsymbol{\alpha} &\sim \text{Dir}(\boldsymbol{\alpha}) \\ &= \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \theta_1^{\alpha_1 - 1} \cdots \theta_K^{\alpha_K - 1},\end{aligned} \tag{7.312}$$

where $\Gamma(\cdot)$ is the gamma function and $\boldsymbol{\alpha}$ is a K-vector parameter with component $\alpha_k > 0$. This K-dimensional Dirichlet random variable $\boldsymbol{\theta}$ lies in the $(K-1)$-simplex. Thus, LDA parameters $\{\boldsymbol{\alpha}, \boldsymbol{\beta}\}$ contain the K-dimensional Dirichlet parameters $\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_K]^\mathsf{T}$ for K topic mixtures $\boldsymbol{\theta}$ and the topic-dependent unigram parameters $\boldsymbol{\beta} = \{\beta_{kv}\} = \{p(w = v|k)\}$. The parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are estimated by maximizing the marginal likelihood of a text corpus $\mathcal{D}$ or an N-word document $\mathbf{w}$ over the topic mixtures $\boldsymbol{\theta}$ and the topic labels $\mathbf{z} = \{z_n\}$:

$$\{\boldsymbol{\alpha}^{\text{ML2}}, \boldsymbol{\beta}^{\text{ML2}}\} = \arg\max_{\{\boldsymbol{\alpha}, \boldsymbol{\beta}\}} p(\mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}), \tag{7.313}$$

where

$$p(\mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \int p(\boldsymbol{\theta}|\boldsymbol{\alpha}) \left(\prod_{n=1}^N \sum_{k=1}^K p(z_n = k|\boldsymbol{\theta}) \times p(w_n = v|z_n = k, \boldsymbol{\beta}) \right) d\boldsymbol{\theta}. \tag{7.314}$$

The marginal likelihood is calculated by integrating over continuous variable $\boldsymbol{\theta}$ and discrete variable $\mathbf{z}$. Note that LDA parameters $\{\boldsymbol{\alpha}^{\text{ML2}}, \boldsymbol{\beta}^{\text{ML2}}\}$ are estimated according to the type-2 maximum likelihood method, as addressed in Section 5.1.2, because the likelihood function considers all possible values of topic mixtures in different topics. Latent variables in LDA include topic mixtures and topic labels $\{\boldsymbol{\theta}, \mathbf{z}\}$. Strictly speaking, the LDA model parameters Θ contain $\{\boldsymbol{\theta}, \boldsymbol{\beta}\}$ while the hyperparameters Ψ should contain $\boldsymbol{\alpha}$. But, using LDA, only the parameters $\boldsymbol{\theta}$ are integrated out. Without loss of generality, we treat both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ as LDA parameters to be optimized from training data $\mathbf{w}$.

However, in model inference, we should apply the EM algorithm which involves a calculation of posterior distribution of latent variables $\{\boldsymbol{\theta}, \mathbf{z}\}$ given a document $\mathbf{w}$:

$$p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta})}{p(\mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta})}. \tag{7.315}$$

Unfortunately, this distribution is intractable because the normalization term

$$p(\mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \int \left(\prod_{k=1}^K \theta_k^{\alpha_k - 1} \right) \times \left(\prod_{n=1}^N \sum_{k=1}^K \prod_{v=1}^{|\mathcal{V}|} (\theta_k \beta_{kv})^{w_n^v} \right) d\boldsymbol{\theta} \tag{7.316}$$

is an intractable function due to the *coupling* between $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ in the summation over latent topics. Here, the superscript v in w_n^v denotes the component index, i.e., $w_n^v = 1$

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

and $w_n^j = 0$ for $j \neq v$. Next, we address the variational Bayes (VB) inference procedure, which is divided into three steps, namely finding the lower bound, finding the variational parameters, and finding the model parameters.

7.6.2 VB inference: lower bound

Although the posterior distribution is intractable for exact inference, a variety of approximate inference algorithms can be used for LDA including a Laplace approximation, variational approximation, and Markov chain Monte Carlo. In this section, a simple convexity-based variational inference is introduced to implement an LDA model by using Jensen's inequality. A *variational distribution*,

$$q(\boldsymbol{\theta}, \mathbf{z}|\boldsymbol{\gamma}, \boldsymbol{\phi}) = q(\boldsymbol{\theta}|\boldsymbol{\gamma}) \prod_{n=1}^{N} q(z_n|\phi_n), \tag{7.317}$$

is used as a surrogate for the posterior distribution $p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})$, where the *variational parameters* $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$ are estimated via an optimization. According to Jensen's inequality using the convex function $-\log(\cdot)$, we have

$$\begin{aligned}
\log p(\mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}) &= \log \int \sum_{\mathbf{z}} p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}) d\boldsymbol{\theta} \\
&= \log \int \sum_{\mathbf{z}} \frac{p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}) q(\boldsymbol{\theta}, \mathbf{z})}{q(\boldsymbol{\theta}, \mathbf{z})} d\boldsymbol{\theta} \\
&\geq \int \sum_{\mathbf{z}} q(\boldsymbol{\theta}, \mathbf{z}) \log p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}) d\boldsymbol{\theta} \\
&\quad - \int \sum_{\mathbf{z}} q(\boldsymbol{\theta}, \mathbf{z}) \log q(\boldsymbol{\theta}, \mathbf{z}) d\boldsymbol{\theta} \\
&= \mathbb{E}_{(\boldsymbol{\theta}, \mathbf{z})}[\log p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta})] - \mathbb{E}_{(\boldsymbol{\theta}, \mathbf{z})}[\log q(\boldsymbol{\theta}, \mathbf{z})] \\
&\triangleq \mathcal{L}[\boldsymbol{\gamma}, \boldsymbol{\phi}; \boldsymbol{\alpha}, \boldsymbol{\beta}] \triangleq \mathcal{F}[q(\boldsymbol{\theta}|\boldsymbol{\gamma}), q(\mathbf{z}|\boldsymbol{\phi})],
\end{aligned} \tag{7.318}$$

where $\mathbb{E}[\cdot]$ denotes the expectation operation and variational parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$ are omitted for simplicity. Jensen's inequality provides us with a lower bound $\mathcal{L}[\boldsymbol{\gamma}, \boldsymbol{\phi}; \boldsymbol{\alpha}, \boldsymbol{\beta}]$ on the logarithm of marginal likelihood, given an arbitrary variational distribution $q(\boldsymbol{\theta}, \mathbf{z}|\boldsymbol{\gamma}, \boldsymbol{\phi})$. It can be easily verified that the difference between the left-hand-side and the right-hand-side of Eq. (7.318) is the KL divergence between the variational posterior distribution and the true posterior distribution. We have

$$\log p(\mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\phi}; \boldsymbol{\alpha}, \boldsymbol{\beta}) + \mathrm{KL}(q(\boldsymbol{\theta}, \mathbf{z}|\boldsymbol{\gamma}, \boldsymbol{\phi}) \| p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})). \tag{7.319}$$

Therefore, maximizing the lower bound $\mathcal{L}[\boldsymbol{\gamma}, \boldsymbol{\phi}; \boldsymbol{\alpha}, \boldsymbol{\beta}]$ with respect to $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$ is equivalent to finding the optimal variational distribution $q(\boldsymbol{\theta}, \mathbf{z}|\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\phi}})$ with variational parameters $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\phi}}$, which is closest to the true posterior distribution $p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})$. To do so, the lower bound is expanded by

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\phi}; \boldsymbol{\alpha}, \boldsymbol{\beta}) = {} & \mathbb{E}_{(\boldsymbol{\theta})}[\log p(\boldsymbol{\theta}|\boldsymbol{\alpha})] + \mathbb{E}_{(\boldsymbol{\theta}, \mathbf{z})}[\log p(\mathbf{z}|\boldsymbol{\theta})] \\
& + \mathbb{E}_{(\mathbf{z})}[\log p(\mathbf{w}|\mathbf{z}, \boldsymbol{\beta})] - \mathbb{E}_{(\boldsymbol{\theta})}[\log q(\boldsymbol{\theta})] - \mathbb{E}_{(\mathbf{z})}[\log q(\mathbf{z})].
\end{aligned} \tag{7.320}$$

Using the fact that the expectation of the sufficient statistics is equivalent to the derivative of the log normalization factor with respect to the natural parameter, we obtain (Blei *et al.* 2003)

$$\mathbb{E}_{(\boldsymbol{\theta})}[\log \theta_k|\boldsymbol{\alpha}] = \Psi(\alpha_k) - \Psi\left(\sum_{i=1}^{K}\alpha_i\right), \tag{7.321}$$

where Ψ is the first derivative of the log gamma function, also called the di-gamma function, as used in Eqs. (5.82) and (7.81). The lower bound is further expanded in terms of variational parameters $\{\boldsymbol{\gamma}, \boldsymbol{\phi}\}$ and model parameters $\{\boldsymbol{\alpha}, \boldsymbol{\beta}\}$ by

$$\begin{aligned}\mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\phi}; \boldsymbol{\alpha}, \boldsymbol{\beta}) = \log \Gamma\left(\sum_{k=1}^{K}\alpha_k\right) &- \sum_{k=1}^{K}\log\Gamma(\alpha_k)\\ &+ \sum_{k=1}^{K}(\alpha_k - 1)\left(\Psi(\gamma_k) - \Psi\left(\sum_{i=1}^{K}\gamma_i\right)\right)\\ &+ \sum_{n=1}^{N}\sum_{k=1}^{K}\phi_{nk}\left(\Psi(\gamma_k) - \Psi\left(\sum_{i=1}^{K}\gamma_i\right)\right)\\ &+ \sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{v=1}^{|\mathcal{V}|}\phi_{nk}w_n^v\log\beta_{kv} - \log\Gamma\left(\sum_{k=1}^{K}\gamma_k\right)\\ &+ \sum_{k=1}^{K}\log\Gamma(\gamma_k) - \sum_{k=1}^{K}(\gamma_k - 1)\left(\Psi(\gamma_k) - \Psi\left(\sum_{i=1}^{K}\gamma_i\right)\right)\\ &- \sum_{n=1}^{N}\sum_{k=1}^{K}\phi_{nk}\log\phi_{nk}.\end{aligned} \tag{7.322}$$

Typically, finding the lower bound of marginal likelihood in VB inference is equivalent to performing the VB E-step. However, we need to estimate the optimal variational parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$ to finalize the VB E-step.

7.6.3 VB inference: variational parameters

The variational Dirichlet parameters $\boldsymbol{\gamma}$ and variational multinomial parameters $\boldsymbol{\phi}$ are estimated by maximizing the lower bound in Eq. (7.322). The terms related to $\boldsymbol{\gamma}$ are collected and arranged thus:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\gamma}) = \sum_{k=1}^{K}\left(\Psi(\gamma_k) - \Psi\left(\sum_{i=1}^{K}\gamma_i\right)\right)\left(\alpha_k + \sum_{n=1}^{N}\phi_{nk} - \gamma_k\right)\\ - \log\Gamma\left(\sum_{k=1}^{K}\gamma_k\right) + \sum_{k=1}^{K}\log\Gamma(\gamma_k).\end{aligned} \tag{7.323}$$

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

Differentiating with respect to the individual parameter γ_k, we have

$$\begin{aligned}\frac{\partial \mathcal{L}(\boldsymbol{\gamma})}{\partial \gamma_k} &= \Psi'(\gamma_k)\left(\alpha_k + \sum_{n=1}^{N}\phi_{nk} - \gamma_k\right)\\ &\quad - \Psi'\left(\sum_{i=1}^{K}\gamma_i\right)\sum_{i=1}^{K}\left(\alpha_i + \sum_{n=1}^{N}\phi_{ni} - \gamma_i\right). \end{aligned} \tag{7.324}$$

Setting this equation to zero yields the optimal variational parameters

$$\hat{\gamma}_k = \alpha_k + \sum_{n=1}^{N}\phi_{nk} \qquad 1 \le k \le K. \tag{7.325}$$

Note that the variational Dirichlet parameters $\hat{\boldsymbol{\gamma}} = \{\hat{\gamma}_k\}$ are seen as the surrogate of the Dirichlet model parameters $\boldsymbol{\alpha}$, which sufficiently reflect the topic mixture probabilities $\boldsymbol{\theta}$.

On the other hand, when optimizing the lower bound $\mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\phi}; \boldsymbol{\alpha}, \boldsymbol{\beta})$ with respect to variational multinomial parameters $\boldsymbol{\phi}$, a constrained maximization problem should be tackled under the constraint

$$\sum_{k=1}^{K}\phi_{nk} = 1. \tag{7.326}$$

Therefore, we collect the terms in the lower bound related to the individual variational parameter ϕ_{nk} and form the Lagrangian with the Lagrange multiplier λ_n:

$$\begin{aligned}\mathcal{L}(\phi_{nk}) &= \phi_{nk}\left(\Psi(\gamma_k) - \Psi\left(\sum_{i=1}^{K}\gamma_i\right)\right)\\ &\quad + \phi_{nk}\log\beta_{kv} - \phi_{nk}\log\phi_{nk} + \lambda_n\left(\sum_{i=1}^{K}\phi_{ni} - 1\right), \end{aligned} \tag{7.327}$$

where the unique word v is selected for w_n such that $w_n^v = 1$ and $w_n^j = 0$ for $j \neq v$. Differentiating this Lagrangian with respect to ϕ_{nk} yields

$$\frac{\partial \mathcal{L}(\phi_{nk})}{\partial \phi_{nk}} = \Psi(\gamma_k) - \Psi\left(\sum_{i=1}^{K}\gamma_i\right) + \log\beta_{kv} - \log\phi_{nk} - 1 + \lambda_n. \tag{7.328}$$

We set this differentiation to zero and derive the variational parameter ϕ_{nk} which is written as a function of Lagrange multiplier λ_n. By substituting the derived variational parameter ϕ_{nk} into the constraint given in Eq. (7.326), we can estimate the multiplier and then obtain the optimal variational parameters:

$$\begin{aligned}\hat{\phi}_{nk} &= \frac{\beta_{kv}\exp\left(\Psi(\gamma_k) - \Psi\left(\sum_{i=1}^{K}\gamma_i\right)\right)}{\sum_{j=1}^{K}\beta_{jv}\exp\left(\Psi(\gamma_j) - \Psi\left(\sum_{i=1}^{K}\gamma_i\right)\right)}\\ &\quad 1 \le n \le N, \quad 1 \le k \le K. \end{aligned} \tag{7.329}$$

7.6.4 VB inference: model parameters

After finding the optimal variational parameters $\{\boldsymbol{\gamma}, \boldsymbol{\phi}\}$, we fix these parameters and substitute the optimal variational distribution $q(\boldsymbol{\theta}, \mathbf{z}|\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\phi}})$ to calculate the updated lower bound $\mathcal{L}(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\phi}}; \boldsymbol{\alpha}, \boldsymbol{\beta})$. The VB M-step treats the variational lower bound $\mathcal{L}(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\phi}}; \boldsymbol{\alpha}, \boldsymbol{\beta})$ as a surrogate of the intractable marginal log likelihood $\log p(\mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta})$, and estimates the LDA parameters by

$$\{\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}\} = \arg \max_{\{\boldsymbol{\alpha}, \boldsymbol{\beta}\}} \mathcal{L}(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\phi}}; \boldsymbol{\alpha}, \boldsymbol{\beta}). \tag{7.330}$$

First, we deal with the optimization over the conditional multinomial distributions $\boldsymbol{\beta} = \{\beta_{kv}\} = \{p(w_n = v|z_n = k)\}$. The lower bound is hereafter calculated from a text corpus $\mathcal{D} = \{\mathbf{w}_1, \cdots, \mathbf{w}_M\}$. The terms containing model parameters $\boldsymbol{\beta}$ are collected and the constraints

$$\sum_{v=1}^{|\mathcal{V}|} \beta_{kv} = 1 \tag{7.331}$$

are imposed, so as to form the Lagrangian with Lagrange multipliers $\{\lambda_k\}$:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta}) = &\sum_{d=1}^{M} \sum_{n=1}^{N_d} \sum_{k=1}^{K} \sum_{v=1}^{|\mathcal{V}|} \hat{\phi}_{dnk} w_{dn}^{v} \log \beta_{kv} \\ &+ \sum_{k=1}^{K} \lambda_k \left(\sum_{v=1}^{|\mathcal{V}|} \beta_{kv} - 1 \right).\end{aligned} \tag{7.332}$$

We differentiate this Lagrangian with respect to individual β_{kv} and set it to zero to find the optimal conditional multinomial distributions:

$$\begin{aligned}\hat{\beta}_{kv} = &\frac{\sum_{d=1}^{M} \sum_{n=1}^{N_d} \hat{\phi}_{dnk} w_{dn}^{v}}{\sum_{m=1}^{|\mathcal{V}|} \sum_{d=1}^{M} \sum_{n=1}^{N_d} \hat{\phi}_{dnk} w_{dn}^{m}} \\ &1 \leq k \leq K, \quad 1 \leq v \leq |\mathcal{V}|.\end{aligned} \tag{7.333}$$

To deal with the optimization over the Dirichlet parameters $\boldsymbol{\alpha} = \{\alpha_k\}$, we collect the terms in the lower bound which contain $\boldsymbol{\alpha}$ and give

$$\begin{aligned}\mathcal{L}(\boldsymbol{\alpha}) = \sum_{d=1}^{M} \Bigg[&\log \Gamma \left(\sum_{k=1}^{K} \alpha_k \right) - \sum_{k=1}^{K} \log \Gamma(\alpha_k) \\ &+ \sum_{k=1}^{K} (\alpha_k - 1) \left(\Psi(\hat{\gamma}_{dk}) - \Psi \left(\sum_{i=1}^{K} \hat{\gamma}_{di} \right) \right) \Bigg].\end{aligned} \tag{7.334}$$

Differentiating with respect to individual α_k gives

$$\begin{aligned}\frac{\partial \mathcal{L}(\boldsymbol{\alpha})}{\partial \alpha_k} = &M \left(\Psi \left(\sum_{i=1}^{K} \alpha_i \right) - \Psi(\alpha_k) \right) \\ &+ \sum_{d=1}^{M} \left(\Psi(\hat{\gamma}_{dk}) - \Psi \left(\sum_{i=1}^{K} \hat{\gamma}_{di} \right) \right).\end{aligned} \tag{7.335}$$

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

There is no closed-form solution to the optimal Dirichlet parameter α_k, since the right-hand-side of Eq. (7.335) depends on α_i where $i \neq k$. We should use an iterative algorithm to find the $K \times 1$ optimal parameter vector $\boldsymbol{\alpha}$. Here, the Newton–Raphson optimization is applied to find the optimal parameters by iteratively performing

$$\boldsymbol{\alpha}^{(\tau+1)} = \boldsymbol{\alpha}^{(\tau)} - (\mathbf{H}(\boldsymbol{\alpha}^{(\tau)}))^{-1} \nabla \mathcal{L}(\boldsymbol{\alpha}^{(\tau)}), \tag{7.336}$$

where τ is the iteration index, $\nabla \mathcal{L}$ is the $K \times 1$ gradient vector and $\mathbf{H}$ is the $K \times K$ Hessian matrix consisting of the second-order differentiations in different entries:

$$\frac{\partial \mathcal{L}(\boldsymbol{\alpha})}{\partial \alpha_k \alpha_j} = \delta(k,j) M \Psi'(\alpha_k) - \Psi' \left(\sum_{i=1}^{K} \alpha_i \right), \tag{7.337}$$

where $\delta(k,j)$ denotes a Kronecker delta function. The inverse of Hessian matrix $(\mathbf{H}(\boldsymbol{\alpha}^{(\tau)}))^{-1}$ can be obtained by applying the Woodbury matrix inversion. As a result, LDA model parameters $\{\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}\}$ are estimated in a VB M-step. The VB inference based on an EM algorithm is accordingly completed by maximizing the variational lower bound of marginal likelihood $\mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\phi}; \boldsymbol{\alpha}, \boldsymbol{\beta})$ through performing a VB E-step for updating variational parameters $\{\boldsymbol{\gamma}, \boldsymbol{\phi}\}$ and a VB M-step for estimating model parameters $\{\boldsymbol{\alpha}, \boldsymbol{\beta}\}$. The increase of the lower bound is assured by VB–EM iterations.

7.7 Latent topic language model

LDA is established as a latent topic model which is designed for document representation by using a bag of words in the form of document $\mathbf{w}$ or text corpus $\mathcal{D}$. LDA has been successfully extended for language modeling and applied for continuous speech recognition in Tam & Schultz (2005) and Chien & Chueh (2011). However, before the LDA language model, the topic model based on PLSA was constructed and merged in the n-gram language model (Gildea & Hofmann 1999), as addressed in Section 3.7.3. But, the PLSA-based topic model suffers from the weaknesses of poor generalization and redundant model complexity. The performance of the resulting PLSA language model is limited. Compared to the PLSA language model, we are more interested in the LDA language model and its application in speech recognition. In the literature, the LDA-based topic model is incorporated into the n-gram language model based on an indirect method (Tam & Schultz 2005) and a direct method (Chien & Chueh 2011), which are introduced in Section 7.7.1 and Section 7.7.2, respectively.

7.7.1 LDA language model

LDA is generally *indirect* for characterizing the n-gram regularities of a current word w_i given its history words w_{i-n+1}^{i-1}. The word index n in the document model differs from i in the language model. A document $\mathbf{w}$ has N words while a sentence has T words. The hierarchical Dirichlet language model in Section 5.3 was presented as an alternative to language model smoothing (MacKay & Peto 1995). In Yaman, Chien & Lee (2007),

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

the hierarchical Dirichlet priors were estimated by a maximum a-posteriori method for language model adaptation. These two methods have adopted the Dirichlet priors to explore the structure of a language model from lower-order n-gram to higher-order n-gram. There was no topic-based language model involved. In Wallach (2006), the LDA bi-gram was proposed by considering a bag of bi-gram events from the collected documents in construction of an LDA. This LDA bi-gram was neither derived nor specifically employed for speech recognition. The basic LDA model ignores the word order and is not in accordance with sentence generation in speech recognition. In Tam & Schultz (2005, 2006), the LDA model parameters $\{\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}\}$ were calculated from training documents $\mathcal{D}$ and then used to estimate the online topic probabilities or the variational Dirichlet parameters $\hat{\boldsymbol{\gamma}} = \{\hat{\gamma}_k\}$. The online parameters $\hat{\boldsymbol{\gamma}}$ were estimated by treating all history words in a sentence w_1^{i-1} (Tam & Schultz 2005), or even the transcription of a whole sentence w_1^T (Tam & Schultz 2006) as a single *document* $\mathbf{w}$.

Similarly to the PLSA n-gram as illustrated in Section 3.7.3, the LDA n-gram is formed as a soft-clustering model or a topic mixture model, which is calculated by combining the topic probabilities $p(z_i = k|w_1^{i-1})$ driven by history words w_1^{i-1} and the topic-dependent unigrams $\boldsymbol{\beta} = \{\beta_{kv}\} = \{p(w_i = v|z_i = k)\}$ of current word w_i. The combination is marginalized over different topics:

$$\begin{aligned} p_{\text{LDA}}(w_i = v|w_1^{i-1}) &= \sum_{k=1}^{K} p(w_i = v|z_i = k)p(z_i = k|w_1^{i-1}) \\ &\approx \sum_{k=1}^{K} \beta_{kv} \frac{\hat{\gamma}_k}{\sum_{j=1}^{K} \hat{\gamma}_j}. \end{aligned} \tag{7.338}$$

Here, the topic multinomial distributions can be driven either by the history words $p(z_i = k|w_1^{i-1})$ or by the whole sentence words $p(z_i = k|w_1^T)$. These multinomial distributions are approximated and proportional to the variational Dirichlet parameters $\hat{\boldsymbol{\gamma}}$ as calculated in Eq. (7.325). In this implementation, the pre-trained model parameters $\{\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}\}$ and the online estimated variational parameters $\{\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\phi}}\}$ should be calculated to determine the LDA n-gram $p_{\text{LDA}}(w_i|w_1^{i-1})$ in an online fashion. In Tam & Schultz (2005), the LDA language model was improved by further adapting the standard ML-based n-gram model using the LDA n-gram according to a linear interpolation scheme with a parameter $0 < \lambda < 1$:

$$\hat{p}(w_i|w_1^{i-1}) = \lambda p_{\text{ML}}(w_i|w_{i-n+1}^{i-1}) + (1 - \lambda)p_{\text{LDA}}(w_i|w_1^{i-1}). \tag{7.339}$$

In Tam & Schultz (2006), the language model adaptation based on the unigram rescaling was implemented by

$$\hat{p}(w_i|w_1^{i-1}) = p_{\text{ML}}(w_i|w_{i-n+1}^{i-1}) \frac{p_{\text{LDA}}(w_i|w_1^{i-1})}{p_{\text{ML}}(w_i)}. \tag{7.340}$$

Typically, the model parameters $\{\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}\}$ in the LDA language model are inferred at the document level, catching the long-distance topic information but only *indirectly* representing the n-gram events.

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

7.7.2 Dirichlet class language model

LDA (Blei *et al.* 2003) builds a hierarchical Bayesian topic model and extracts the latent topics or clusters from a collection of M documents $\mathcal{D} = \{\mathbf{w}_m\}$. The bag-of-words scheme is adopted without considering the word sequence, and so it is not directly designed for speech recognition where the performance is seriously affected by the prior probability of a word string $W = w_1^T \triangleq \{w_1, \cdots, w_T\}$ or the language model $p(W)$. In Bengio, Ducharme, Vincent *et al.* (2003), the neural network language model was proposed to deal with the data sparseness problem in language modeling by projecting the ordered history vector into a continuous space and then calculating the n-gram language model based on the multilayer perceptron. MLP involves the error back-propagation training algorithm based on the least-squares estimation, which is vulnerable to the overfitting problem (Bishop 2006). Considering the topic modeling in LDA and the continuous representation of history word sequence w_{i-n+1}^{i-1} in a neural network language model, the Dirichlet class language model (DCLM) (Chien & Chueh 2011) is presented to build a *direct* LDA language model for speech recognition. For a vocabulary with $|\mathcal{V}|$ words, the $n-1$ history words w_{i-n+1}^{i-1} are first represented by a $(n-1)|\mathcal{V}| \times 1$ vector $\mathbf{h}_{i-n+1}^{i-1}$ consisting of $n-1$ block subvectors. Each block is represented by the 1-to-$|\mathcal{V}|$ coding scheme with a $|\mathcal{V}|$ dimensional vector where the vth word of vocabulary is encoded by setting the vth entry of the vector to be one and all the other entries to be zero. The order of history words is considered in $\mathbf{h}_{i-n+1}^{i-1}$. Figure 7.3 shows the system architecture of calculating a DCLM $p_{\text{DC}}(w_i = v|w_{i-n+1}^{i-1})$. A global projection is involved to project the ordered history vector $\mathbf{h}_{i-n+1}^{i-1}$ into a latent topic or class space where the projection $\mathbf{g}(\mathbf{h}_{i-n+1}^{i-1})$ could be either a linear function or a non-linear function. In the case of a linear function, a projection matrix $\mathbf{A} = [\mathbf{a}_1, \cdots, \mathbf{a}_C]$ consisting of

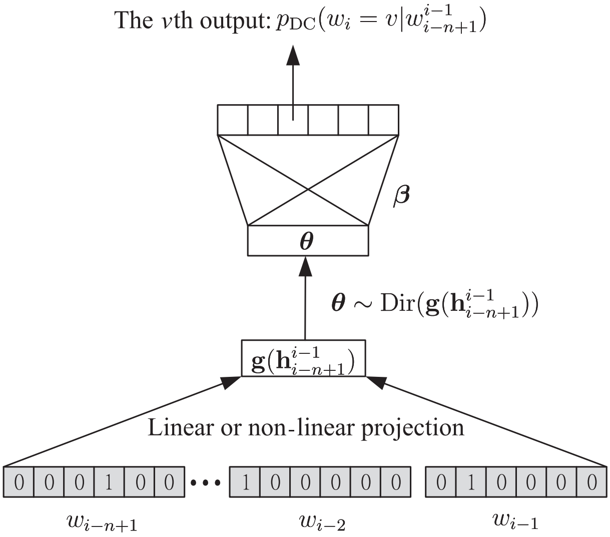
The vth output: $p_{\text{DC}}(w_i = v|w_{i-n+1}^{i-1})$
$\boldsymbol{\beta}$
$\boldsymbol{\theta}$
$\boldsymbol{\theta} \sim \text{Dir}(\mathbf{g}(\mathbf{h}_{i-n+1}^{i-1}))$
$\mathbf{g}(\mathbf{h}_{i-n+1}^{i-1})$
Linear or non-linear projection
0 0 0 1 0 0
1 0 0 0 0 0
0 1 0 0 0 0
w_{i-n+1}
w_{i-2}
w_{i-1}

Figure 7.3 System architecture for Dirichlet class language model.

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

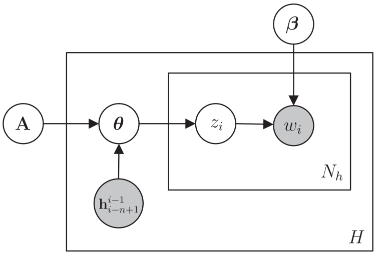

Figure 7.4 Representation of a Dirichlet class language model.

C basis vectors $\{\mathbf{a}_c\}$ is involved. The class structure of all n-gram events from training corpus $\mathcal{D} = \{w_i, w_{i-n+1}^{i-1}\}$ is represented by Dirichlet distributions. The class uncertainty is compensated by marginalizing the likelihood function over the Dirichlet priors. The latent structure in the DCLM reflects the class of an n-gram event rather than the topic in an LDA model. A DCLM is regarded as a kind of class-based language model, which is inferred by the variational Bayes (VB) procedure by maximizing the variational lower bound of a marginal likelihood of training data.

7.7.3 Model construction

A DCLM acts as a Bayesian class-based language model, which involves the prior distribution of the class variable. Figure 7.4 is a graphical representation of a DCLM from a training set of current words and history words $\mathcal{D} = \{w_i, w_{i-n+1}^{i-1}\} = \{w_i, \mathbf{h}_{i-n+1}^{i-1}\}$. The training corpus has H history events in $\{w_{i-n+1}^{i-1}\}$. Each history event $h = w_{i-n+1}^{i-1}$ has N_h possible predicted words $\{w_i\}$. Note that the $(n-1)|\mathcal{V}|$-dimensional discrete history vector is projected to a C-dimensional continuous class space using the class-dependent linear discriminant function

$$g_c(\mathbf{h}_{i-n+1}^{i-1}) = \mathbf{a}_c^\intercal \mathbf{h}_{i-n+1}^{i-1}. \tag{7.341}$$

This function is used as the hyperparameter of a Dirichlet prior for the class mixture probability θ_c or equivalently the class posterior probability of latent variable $z_i = c$ given history vector $\mathbf{h}_{i-n+1}^{i-1}$:

$$\theta_c \triangleq p(z_i = c|\boldsymbol{\theta}, \mathbf{h}_{i-n+1}^{i-1}). \tag{7.342}$$

Thus, the uncertainty of class posterior probabilities $\mathbf{g}(\mathbf{h}_{i-n+1}^{i-1}) = \{g_c(\mathbf{h}_{i-n+1}^{i-1})\}$ is characterized by a Dirichlet prior distribution:

$$\begin{aligned}\boldsymbol{\theta} = [\theta_1, \cdots, \theta_C]^\intercal | \mathbf{h}_{i-n+1}^{i-1}, \mathbf{A} &\sim \text{Dir}(\mathbf{g}(\mathbf{h}_{i-n+1}^{i-1})) \\ &= \text{Dir}(\mathbf{A}^\intercal \mathbf{h}_{i-n+1}^{i-1}),\end{aligned} \tag{7.343}$$

subject to the constraint $g_c(\mathbf{h}_{i-n+1}^{i-1}) > 0$ or $\mathbf{a}_c > 0$. Each word $w_i = v$ is generated by the conditional probability $\beta_{cv} \triangleq p(w_i = v|z_i = c)$ given a latent class $z_i = c$.

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

The n-gram probability obtained using DCLM is calculated by marginalizing the joint likelihood over the latent variables including class mixtures $\boldsymbol{\theta}$ and C latent classes:

$$\begin{aligned}
p_{\mathrm{DC}}(w_i = v|\mathbf{h}_{i-n+1}^{i-1}, \mathbf{A}, \boldsymbol{\beta}) &= \sum_{c=1}^{C} p(w_i = v|z_i = c, \boldsymbol{\beta})p(z_i = c|\mathbf{h}_{i-n+1}^{i-1}, \mathbf{A}) \\
&= \sum_{c=1}^{C} p(w_i = v|z_i = c, \boldsymbol{\beta}) \\
&\quad \times \int p(\boldsymbol{\theta}|\mathbf{h}_{i-n+1}^{i-1}, \mathbf{A})p(z_i = c|\boldsymbol{\theta})d\boldsymbol{\theta} \\
&= \sum_{c=1}^{C} \beta_{cv}\mathbb{E}_{(\boldsymbol{\theta})}[p(z_i = c|\boldsymbol{\theta}, \mathbf{h}_{i-n+1}^{i-1}, \mathbf{A})] \\
&= \sum_{c=1}^{C} \beta_{cv}\mathbb{E}_{(\boldsymbol{\theta})}[\theta_c|\mathbf{h}_{i-n+1}^{i-1}, \mathbf{A}] \\
&= \sum_{c=1}^{C} \beta_{cv} \frac{\mathbf{a}_c^{\mathsf{T}}\mathbf{h}_{i-n+1}^{i-1}}{\sum_{j=1}^{C} \mathbf{a}_j^{\mathsf{T}}\mathbf{h}_{i-n+1}^{i-1}}.
\end{aligned} \tag{7.344}$$

In Eq. (7.344), the integral is calculated over a multinomial variable $p(z_i = c|\boldsymbol{\theta}, \mathbf{h}_{i-n+1}^{i-1}, \mathbf{A}))$ with Dirichlet prior distribution $p(\boldsymbol{\theta}|\mathbf{h}_{i-n+1}^{i-1}, \mathbf{A})$ and is equivalent to the distribution mean. This integral is seen as a marginalization over different classes c, and is also obtained as a class mixture model with the class-dependent unigram probabilities $\{\beta_{cv}|1 \leq c \leq C\}$ weighted by the class mixture probabilities $\{p(z_i = c|\mathbf{h}_{i-n+1}^{i-1}, \mathbf{A})|1 \leq c \leq C\}$ driven by the ordered history vector $\mathbf{h}_{i-n+1}^{i-1}$. However, we should estimate DCLM parameters $\{\mathbf{A}, \boldsymbol{\beta}\}$ and substitute them into Eq. (7.344) to calculate the Dirichlet class (DC) n-gram probabilities.

7.7.4 VB inference: lower bound

Similarly to LDA, DCLM parameters $\{\mathbf{A}, \boldsymbol{\beta}\}$ are estimated according to the type-2 maximum likelihood method where the marginal likelihood over latent variables is maximized. As seen in Figure 7.4, the latent variables in DCLM include the continuous values of class mixture probabilities $\boldsymbol{\theta} = \{\theta_c\}$ and the discrete values of class labels $\mathbf{z} = \{z_i\}$. The optimization problem is formulated as

$$\{\mathbf{A}^{\mathrm{ML2}}, \boldsymbol{\beta}^{\mathrm{ML2}}\} = \arg\max_{\mathbf{A},\boldsymbol{\beta}} \log p(\mathcal{D}|\mathbf{A}, \boldsymbol{\beta}), \tag{7.345}$$

where

$$\begin{aligned}
\log p(\mathcal{D}|\mathbf{A}, \boldsymbol{\beta}) &= \sum_{\{w_i,\mathbf{h}_{i-n+1}^{i-1}\}\in\mathcal{D}} \log p(w_i|\mathbf{h}_{i-n+1}^{i-1}, \mathbf{A}, \boldsymbol{\beta}) \\
&= \sum_{\mathbf{h}_{i-n+1}^{i-1}} \log p(\mathbf{w}_h|\mathbf{h}_{i-n+1}^{i-1}, \mathbf{A}, \boldsymbol{\beta})
\end{aligned}$$

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

$$= \sum_{\mathbf{h}_{i-n+1}^{i-1}} \log \left(\prod_{i=1}^{N_h} p(w_i|\mathbf{h}_{i-n+1}^{i-1}, \mathbf{A}, \boldsymbol{\beta}) \right)$$

$$= \sum_{\mathbf{h}_{i-n+1}^{i-1}} \log \Bigg(\int p(\theta|\mathbf{h}_{i-n+1}^{i-1}, \mathbf{A})$$

$$\times \left(\prod_{i=1}^{N_h} \sum_{c=1}^{C} p(w_i = v|z_i = c, \boldsymbol{\beta}) p(z_i = c|\boldsymbol{\theta}) \right) d\boldsymbol{\theta} \Bigg). \tag{7.346}$$

The log marginal likelihood is calculated by integrating over the continuous $\boldsymbol{\theta}$ and summing the discrete $\{z_i = c|1 \leq c \leq C\}$. However, directly optimizing Eq. (7.346) is intractable because of the coupling between $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ in the summation over latent classes. The posterior distribution

$$p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{w}_h, \mathbf{h}_{i-n+1}^{i-1}, \mathbf{A}, \boldsymbol{\beta}) = \frac{p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}_h|\mathbf{h}_{i-n+1}^{i-1}, \mathbf{A}, \boldsymbol{\beta})}{p(\mathbf{w}_h|\mathbf{h}_{i-n+1}^{i-1}, \mathbf{A}, \boldsymbol{\beta})} \tag{7.347}$$

becomes intractable for model inference. This posterior distribution is calculated from the n-gram events with the N_h predicted words occurring after the fixed history words $h \triangleq w_{i-n+1}^{i-1}$,

$$\mathbf{w}_h \triangleq \{w_i|c(hw_i) > 0\}, \tag{7.348}$$

where $c(\cdot)$ denotes the count of an n-gram event. The variational Bayes (VB) inference procedure is involved to construct the variational DCLM where the lower bound of marginal likelihood in Eq. (7.346) serves as the surrogate to be maximized to find the optimal $\{\mathbf{A}^{\text{ML2}}, \boldsymbol{\beta}^{\text{ML2}}\}$.

To do so, a decomposed variational distribution,

$$q(\boldsymbol{\theta}, \mathbf{z}|\boldsymbol{\gamma}_h, \boldsymbol{\phi}_h) = q(\boldsymbol{\theta}|\boldsymbol{\gamma}_h) \prod_{i=1}^{N_h} q(z_i|\boldsymbol{\phi}_{h,i}), \tag{7.349}$$

is introduced to approximate the true posterior distribution $p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{w}_h, \mathbf{h}_{i-n+1}^{i-1}, \mathbf{A}, \boldsymbol{\beta})$. A graphical representation of the variational DCLM is illustrated in Figure 7.5. Here, $\boldsymbol{\gamma}_h = \{\gamma_{h,c}\}$ and $\boldsymbol{\phi}_h = \{\phi_{h,ic}\}$ denote the history-dependent variational Dirichlet and multinomial parameters, respectively. By referring to Section 7.6.2, the lower bound on the log marginal likelihood is derived by applying the Jensen's inequality and is expanded by

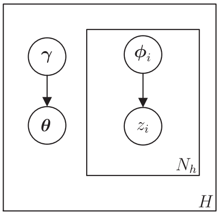

Figure 7.5 Representation of a variational Dirichlet class language model.

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\gamma}_h, \boldsymbol{\phi}_h; \mathbf{A}, \boldsymbol{\beta}) &= \sum_{\mathbf{h}_{i-n+1}^{i-1}} \Big\{ \mathbb{E}_{(\boldsymbol{\theta})}[\log p(\boldsymbol{\theta}|\mathbf{h}_{i-n+1}^{i-1}, \mathbf{A})] + \mathbb{E}_{(\boldsymbol{\theta},\mathbf{z})}[\log p(\mathbf{z}|\boldsymbol{\theta})] \\
&\quad + \mathbb{E}_{(\mathbf{z})}[\log p(\mathbf{w}_h|\mathbf{z}, \boldsymbol{\beta})] - \mathbb{E}_{(\boldsymbol{\theta})}[\log q(\boldsymbol{\theta}|\boldsymbol{\gamma}_h)] - \mathbb{E}_{(\mathbf{z})}[\log q(\mathbf{z}|\boldsymbol{\phi}_h)] \Big\} \\
&= \sum_{\mathbf{h}_{i-n+1}^{i-1}} \Bigg\{ \log \Gamma \left(\sum_{c=1}^{C} \mathbf{a}_c^\top \mathbf{h}_{i-n+1}^{i-1} \right) - \sum_{c=1}^{C} \log \Gamma \left(\mathbf{a}_c^\top \mathbf{h}_{i-n+1}^{i-1} \right) \\
&\quad + \sum_{c=1}^{C} \left(\mathbf{a}_c^\top \mathbf{h}_{i-n+1}^{i-1} - 1 \right) \left(\Psi(\gamma_{h,c}) - \Psi \left(\sum_{j=1}^{C} \gamma_{h,j} \right) \right) \\
&\quad + \sum_{i=1}^{N_h} \sum_{c=1}^{C} \phi_{h,ic} \left(\Psi(\gamma_{h,c}) - \Psi \left(\sum_{j=1}^{C} \gamma_{h,j} \right) \right) \\
&\quad + \sum_{i=1}^{N_h} \sum_{c=1}^{C} \sum_{v=1}^{|\mathcal{V}|} \phi_{h,ic} w_i^v \log \beta_{cv} - \log \Gamma \left(\sum_{c=1}^{C} \gamma_{h,c} \right) + \sum_{c=1}^{C} \log \Gamma(\gamma_{h,c}) \\
&\quad - \sum_{c=1}^{C} (\gamma_{h,c} - 1) \left(\Psi(\gamma_{h,c}) - \Psi \left(\sum_{j=1}^{C} \gamma_{h,j} \right) \right) - \sum_{i=1}^{N_h} \sum_{c=1}^{C} \phi_{h,ic} \log \phi_{h,ic} \Bigg\}.
\end{aligned}
\tag{7.350}
$$

We have applied Eq. (7.321) to find the variational lower bound for DCLM.

7.7.5 VB inference: parameter estimation

A VB–EM algorithm has been developed for inference of DCLM parameters. In this VB–EM procedure, we first conduct the VB E-step to find the optimal expectation function or lower bound $\mathcal{L}(\hat{\boldsymbol{\gamma}}_h, \hat{\boldsymbol{\phi}}_h; \mathbf{A}, \boldsymbol{\beta})$ in Eq. (7.350), or equivalently to estimate the optimal variational Dirichlet parameters $\hat{\boldsymbol{\gamma}}_h$ and variational multinomial parameters $\hat{\boldsymbol{\phi}}_h$. Similarly to Section 7.6.3, we collect the terms in Eq. (7.350) that are related to $\boldsymbol{\gamma}_h$ and maximize these terms, expressed by $\mathcal{L}(\boldsymbol{\gamma}_h)$, with respect to $\boldsymbol{\gamma}_h$ to find the optimal variational Dirichlet parameters:

$$
\hat{\gamma}_{h,c} = \mathbf{a}_c^\top \mathbf{h}_{i-n+1}^{i-1} + \sum_{i=1}^{N_h} \phi_{h,ic} \qquad 1 \leq c \leq C. \tag{7.351}
$$

In addition, when maximizing the lower bound $\mathcal{L}(\boldsymbol{\gamma}_h, \boldsymbol{\phi}_h; \mathbf{A}, \boldsymbol{\beta})$ with respect to the variational multinomial parameters $\boldsymbol{\phi}_h$, we need to collect all terms related to $\boldsymbol{\phi}_h$ and solve a constrained optimization problem subject to a constraint for multinomial distributions:

$$
\sum_{c=1}^{C} \phi_{h,ic} = 1. \tag{7.352}
$$

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

The Lagrangian $\mathcal{L}(\boldsymbol{\phi})$ given the history-dependent Lagrange multipliers $\{\lambda_{h,i}\}$ is arranged and maximized so as to find the optimal variational multinomial distributions:

$$\hat{\phi}_{h,ic} = \frac{\beta_{cv} \exp\left(\Psi(\gamma_{h,c}) - \Psi\left(\sum_{j=1}^{C} \gamma_{h,j}\right)\right)}{\sum_{l=1}^{C} \beta_{lv} \exp\left(\Psi(\gamma_{h,l}) - \Psi\left(\sum_{j=1}^{C} \gamma_{h,j}\right)\right)}, \qquad (7.353)$$
$$1 \le i \le T, \quad 1 \le c \le C$$

where the unique word v is selected for w_i such that $w_i^v = 1$ and $w_i^l = 0$ for $l \neq v$. The variational lower bound is updated with the optimal variational parameters $\mathcal{L}(\hat{\boldsymbol{\gamma}}_h, \hat{\boldsymbol{\phi}}_h; \mathbf{A}, \boldsymbol{\beta})$.

On the other hand, in a VB M-step, we fix the variational parameters $\hat{\boldsymbol{\gamma}}_h, \hat{\boldsymbol{\phi}}_h$ and optimize the updated lower bound to estimate the DCLM model parameters:

$$\{\hat{\mathbf{A}}, \hat{\boldsymbol{\beta}}\} = \arg\max_{\{\mathbf{A},\boldsymbol{\beta}\}} \mathcal{L}(\hat{\boldsymbol{\gamma}}_h, \hat{\boldsymbol{\phi}}_h; \mathbf{A}, \boldsymbol{\beta}). \qquad (7.354)$$

In estimating the conditional multinomial distributions $\boldsymbol{\beta} = \{\beta_{cv}\} = \{p(w_i = v|z_i = c)\}$, the terms containing model parameters $\boldsymbol{\beta}$ are collected and the constraints

$$\sum_{v=1}^{|\mathcal{V}|} \beta_{cv} = 1 \qquad (7.355)$$

are imposed to arrange the Lagrangian $\mathcal{L}(\boldsymbol{\beta})$ with C Lagrange multipliers $\{\lambda_c\}$. By solving

$$\frac{\partial \mathcal{L}(\boldsymbol{\beta})}{\partial \beta_{cv}} = 0 \qquad (7.356)$$

and considering the constraints, we estimate the optimal conditional multinomial distributions

$$\hat{\beta}_{cv} = \frac{\sum_{\mathbf{h}_{i-n+1}^{i-1}} \sum_{i=1}^{N_h} \hat{\phi}_{h,ic} w_i^v}{\sum_{l=1}^{|\mathcal{V}|} \sum_{\mathbf{h}_{i-n+1}^{i-1}} \sum_{i=1}^{N_h} \hat{\phi}_{h,ic} w_i^l} \qquad (7.357)$$
$$1 \le c \le C, \quad 1 \le v \le |\mathcal{V}|$$

by substituting the updated variational multinomial distributions $\hat{\boldsymbol{\phi}}_h = \{\hat{\phi}_{h,ic}\}$. However, there is no closed-form solution to optimal DCLM parameter $\mathbf{A} = [\mathbf{a}_1, \cdots, \mathbf{a}_C]$. This parameter is used to project the ordered history vector $\mathbf{h}_{i-n+1}^{i-1}$ into latent class space. The projected parameter is treated as the Dirichlet class parameter. We may apply the Newton–Raphson algorithm or simply adopt the gradient descent algorithm

$$\mathbf{a}_c^{(\tau+1)} = \mathbf{a}_c^{(\tau)} - \eta \nabla \mathcal{L}(\mathbf{a}_c^{(\tau)}) \quad 1 \le c \le C \qquad (7.358)$$

to derive the optimal Dirichlet class parameter $\hat{\mathbf{A}}$ by using the gradient

$$\nabla_{\mathbf{a}_c}\mathcal{L}(\hat{\boldsymbol{\gamma}},\hat{\boldsymbol{\phi}};\mathbf{A},\boldsymbol{\beta}) = \sum_{\mathbf{h}_{i-n+1}^{i-1}}\left\{\Psi\left(\sum_{j=1}^{C}\mathbf{a}_j^\top\mathbf{h}_{i-n+1}^{i-1}\right)\right.$$
$$\left.-\Psi\left(\mathbf{a}_c^\top\mathbf{h}_{i-n+1}^{i-1}\right)+\Psi(\hat{\gamma}_{h,c})-\Psi\left(\sum_{j=1}^{C}\hat{\gamma}_{h,j}\right)\right\}\cdot\mathbf{h}_{i-n+1}^{i-1}. \tag{7.359}$$

Given the estimated model parameters $\{\hat{\mathbf{A}},\hat{\boldsymbol{\beta}}\}$, the DCLM n-gram $p_{\text{DC}}(w_i = v|\mathbf{h}_{i-n+1}^{i-1},\mathbf{A},\boldsymbol{\beta})$ in Eq. (7.344) is implemented. The DCLM n-gram could be improved by interpolating with the modified Kneser–Ney (MKN) language model, as discussed in Eq. (3.251). However, the performance of DCLM is limited due to the modeling of Dirichlet classes only inside the n-gram window.

7.7.6 Cache Dirichlet class language model

In general, the long-distance information outside the n-gram window is not captured. This weakness can be compensated in the cache DCLM (Chien & Chueh 2011). The cache DCLM treats all history words w_1^{i-1} as cache memory and incorporates their class information $z_1^{i-1}=\{z_1,\cdots,z_{i-1}\}$ into language modeling as

$$\begin{aligned}
&p(w_i=v|\mathbf{h}_{i-n+1}^{i-1},\mathbf{A},\boldsymbol{\beta},w_1^{i-1})\\
&=\sum_{z_1^{i-1}}p(z_1^{i-1}|w_1^{i-1})p(w_i|z_1^{i-1},\mathbf{h}_{i-n+1}^{i-1},\mathbf{A},\boldsymbol{\beta})\\
&=\sum_{z_1^{i-1}}p(z_1^{i-1}|w_1^{i-1})\sum_{c=1}^{C}p(w_i=v|z_i=c,\boldsymbol{\beta})\\
&\times\int p(\boldsymbol{\theta}|\mathbf{h}_{i-n+1}^{i-1},\mathbf{A},z_1^{i-1})p(z_i=c|\boldsymbol{\theta})d\boldsymbol{\theta},
\end{aligned}\tag{7.360}$$

where the marginalization over latent classes c and class mixtures $\boldsymbol{\theta}$ is performed. We have the posterior distribution

$$p(\boldsymbol{\theta}|\mathbf{h}_{i-n+1}^{i-1},\mathbf{A},z_1^{i-1})=\frac{p(\boldsymbol{\theta}|\mathbf{h}_{i-n+1}^{i-1},\mathbf{A})p(z_1^{i-1}|\boldsymbol{\theta},\mathbf{h}_{i-n+1}^{i-1},\mathbf{A})}{p(z_1^{i-1}|\mathbf{h}_{i-n+1}^{i-1},\mathbf{A})}, \tag{7.361}$$

where

$$\begin{aligned}
p(z_1^{i-1}|\boldsymbol{\theta},\mathbf{h}_{i-n+1}^{i-1},\mathbf{A})&=p(z_1^{i-1}|\boldsymbol{\theta})\\
&=\prod_{c=1}^{C}\theta_c^{\sum_{j=1}^{i-1}\delta(z_j,c)}.
\end{aligned}\tag{7.362}$$

The denominator term of this posterior distribution is independent of w_i and could be ignored in calculating the cache DCLM in Eq. (7.360). For practical purposes, the summation over z_1^{i-1} is simplified by adopting a single best class sequence $\hat{z}_1^{i-1}$, where the

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

best class $\hat{z}_{i-1}$ of word w_{i-1} is detected from the previous $n-1$ words w_{i-n}^{i-2} and the best classes $\hat{z}_1^{i-2}$ corresponding to previous $i-1$ words, namely

$$\begin{aligned}\hat{z}_{i-1} &= \arg\max_{z_{i-1}} p(w_{i-1} = v, z_{i-1} | \hat{z}_1^{i-2}, w_{i-n}^{i-2}, \mathbf{A}, \boldsymbol{\beta}) \\ &= \arg\max_{c} \beta_{cv} \frac{\mathbf{a}_c^\top \mathbf{h}_{i-n}^{i-2} + \sum_{j=1}^{i-2} \delta(\hat{z}_j, c)}{\sum_{l=1}^{C} \left[\mathbf{a}_l^\top \mathbf{h}_{i-n}^{i-2} + \sum_{j=1}^{i-2} \delta(\hat{z}_j, l)\right]}. \end{aligned} \tag{7.363}$$

A detailed derivation is given in Chien & Chueh (2011).

As a result, the recursive detection from $\hat{z}_1$ to $\hat{z}_{i-1}$ is done and applied to approximate the cache DCLM as

$$\begin{aligned} & p(w_i = v | \mathbf{h}_{i-n+1}^{i-1}, \mathbf{A}, \boldsymbol{\beta}, w_1^{i-1}) \\ &\approx \sum_{c=1}^{C} p(w_i = v | \hat{z}_i = c, \boldsymbol{\beta}) \int p(\boldsymbol{\theta} | \mathbf{h}_{i-n+1}^{i-1}, \mathbf{A}) \\ &\quad \times \prod_{m=1}^{C} \theta_m^{\sum_{j=1}^{i-1} \delta(\hat{z}_j, m)} p(z_i = c | \boldsymbol{\theta}) d\boldsymbol{\theta} \\ &\approx \sum_{c=1}^{C} \beta_{cv} \frac{\mathbf{a}_c^\top \mathbf{h}_{i-n+1}^{i-1} + \rho \sum_{j=1}^{i-1} \tau^{i-j-1} \delta(\hat{z}_j, c)}{\sum_{l=1}^{C} \left[\mathbf{a}_l^\top \mathbf{h}_{i-n+1}^{i-1} + \rho \sum_{j=1}^{i-1} \tau^{i-j-1} \delta(\hat{z}_j, l)\right]}. \end{aligned} \tag{7.364}$$

Here, the product of the Dirichlet distribution $p(\boldsymbol{\theta} | \mathbf{h}_{i-n+1}^{i-1}, \mathbf{A})$ and the multinomial distribution $\prod_{m=1}^{C} \theta_m^{\sum_{j=1}^{i-1} \delta(\hat{z}_j, m)}$ is a new Dirichlet distribution. Taking the integral in Eq. (7.364) is equivalent to finding the mean of the new Dirichlet distribution. In this cache DCLM, we introduce a weighting factor $0 < \rho \leq 1$ to balance two terms in the numerator and the denominator and a forgetting factor $o < \tau \leq 1$ to discount the distant class information. A graphical representation of the cache DCLM is given in Figure 7.6. The best classes $\{\hat{z}_1, \cdots, z_{\hat{i}-1}\}$ corresponding to all history words $\{w_1, \cdots, w_{i-1}\}$ are recursively detected and then merged in prediction of the next word w_i.

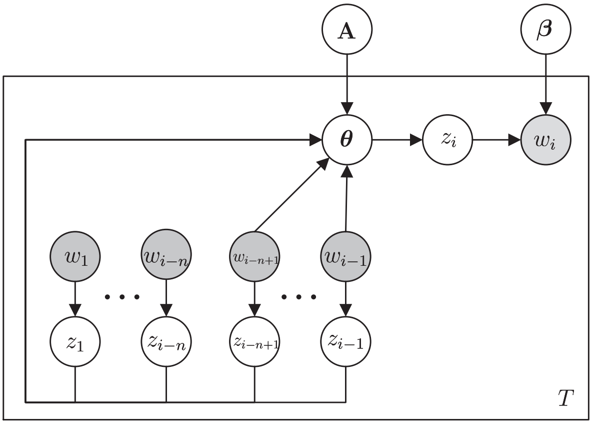

Figure 7.6 Representation of a cache Dirichlet class language model.

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

Table 7.3 Comparison of frequently used words of the latent variables extracted by LDA LM and DCLM.

	Topic/class	Frequently used words in latent topics or classes
LDA LM	Family	toy, kids, theater, event, season, shoe, teen, children's, plays, films, sports, magazines, Christmas, bowling, husband, anniversary, girls, festival, couple, parents,wife, friends
	Election	candidates, race, voters, challenger, democrat, state's, selection, county, front, delegates, elections, reverend, republicans, polls, conventions, label, politician, ballots
	War	troops, killed, Iraqi, attack, ship, violence, fighting, soldiers, mines, Iranian, independence, marines, revolution, died, nation, protect, armed, democracy, violent, commander
DCLM	Quantity	five, seven, two, eight, cents, six, one, nine, four, three, zero, million, point, percent, years, megabyte, minutes, milligrams, bushels, miles, marks, pounds, yen, dollars
	Business	exchange, prices, futures, index, market, sales, revenue, earnings, trading, plans, development, business, funds, organization, traders, ownership, holdings, investment
	In+	addition, the, fact, American, October, recent, contrast, Europe, June, Tokyo, July, March, turn, other, my, Washington, order, Chicago, case, China, general, which

7.7.7 System performance

The *Wall Street Journal* (*WSJ*) corpus was utilized to evaluate different language models for continuous speech recognition (CSR). The SI-84 training set was adopted to estimate the HMM parameters. The feature vector consisted of 12 MFCCs and one log energy and their first and second derivatives. Triphone models were built for 39 phones and one background silence. Each triphone model had three states with eight Gaussian components. The 1987-1989 *WSJ* corpus with 38M words was used to train the baseline backoff trigrams. A total of 86K documents and 3M trigram histories were used. The 20K non-verbalized punctuation, closed vocabularies were adopted. A total of 333 test utterances were sampled from the November 1992 ARPA CSR benchmark test data. In the implementation, the baseline tri-gram was used to generate 100-best lists. Various language models were interpolated with a baseline tri-gram using an interpolation weight, and were employed for N-best rescoring. The neural network LM (NNLM) (Bengio *et al.* 2003), class-based LM (Brown *et al.* 1992), PLSA LM (Gildea & Hofmann 1999), LDA LM (Tam & Schultz 2005, Tam & Schultz 2006), DCLM, and cache DCLM (Chien & Chueh 2011) were evaluated in the comparison.

Table 7.3 lists some examples of latent topics and classes, and the corresponding frequently used words. The three topics and classes were selected from LDA LM with $K = 100$ and DCLM with $C = 100$, respectively. The frequently used words were identified according to the likelihood of the words given target topics or classes. We can see that the frequently used words within a topic or class are semantically close to

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

Table 7.4 Comparison of word error rates for different methods with different sizes of training data and numbers of classes.

	Size of training data			
	6M	12M	18M	38M
Baseline LM	39.2%	21.3%	15.8%	12.9%
NNLM	35.5%	19.6%	15.0%	12.4%
Class-based LM	35.5%	19.7%	15.0%	12.4%
PLSA LM	36.0%	19.8%	15.0%	12.3%
LDA LM	35.9%	19.7%	14.7%	12.2%
DCLM (*C*=200)	35.9%	19.6%	14.6%	12.0%
Cache DCLM (*C*=200)	34.2%	19.3%	14.5%	11.9%
DCLM (*C*=500)	35.2%	19.2%	14.3%	11.7%
Cache DCLM (*C*=500)	33.9%	19.0%	14.2%	11.6%

each other. For some cases, the topically related words from LDA LM appear independently. These topics are not suitable for generating natural language. In contrast, DCLM extracts the class information from n-gram events and performs the history clustering based on sentence generation. For example, the latent class “In+” denotes the category of words that follow the preposition “in.” The words “fact,” “addition,” and ”June” usually follow the word “in,” and appear as frequent words of the same class. The word order is reflected in the clustering procedure. Table 7.4 reports the word error rates for baseline LM, NNLM, class-based LM, PLSA LM, LDA LM, DCLM and cache DCLM with $C = 200$ and $C = 500$ under different sizes of training data. The issue of small sample size is examined. It is consistent that the word error rate is reduced when the amount of training data is increased. In the case of sparse training data (6M), the topic-based and class-based methods work well. In particular, the cache DCLM with $C = 200$ achieved an error rate reduction of 12.9% which outperforms the other related methods. When the number of classes is increased to $C = 500$, the improvement is obvious for the case of large training data (38M).

7.8 Summary

This chapter has introduced various applications of VB to speech and language processing, including CDHMM-based acoustic modeling, acoustic model adaptation, latent topic models, and latent topic language models. Compared with the previous inference approximations based on MAP, evidence, and asymptotic approximations, VB deals with Bayesian inference based on a distribution estimation without considering the asymptotic property, which often provides better solutions in terms of consistently using the Bayesian manner. In addition, the inference algorithm obtained can be regarded as an extension of the conventional EM type iterative algorithm. This makes implementation easier as we can utilize existing source codes based on the ML and MAP–EM algorithms. One of the difficulties in VB is that obtaining the analytical solutions of VB

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press

posterior distributions and variational lower bound is hard due to the complicated expectation and integral calculations. However, the solutions provided in this chapter (general formulas and specific solutions for acoustic and language modeling issues) would cover most of the mathematical analysis in the other VB applications in speech and language processing. Actually, there have been various other aspects of VB including speech feature extraction (Kwon, Lee & Chan 2002, Valente & Wellekens 2004*a*), voice activity detection (Cournapeau, Watanabe, Nakamura *et al*. 2010), speech/speaker GMM (Valente 2006, Pettersen 2008), speaker diarization (Valente, Motlicek & Vijayasenan 2010, Ishiguro, Yamada, Araki *et al*. 2012), and statistical speech synthesis (Hashimoto, Nankaku & Tokuda 2009, Hashimoto, Zen, Nankaku *et al*. 2009). Readers who are interested in these topics could follow these studies and develop new techniques with the solutions provided in this chapter.

Although VB provides a fully Bayesian treatment, VB still cannot overcome the problem of local optimum solutions based on the EM style algorithm, and VB also cannot provide analytical solutions for non-exponential family distributions. The next chapter deals with MCMC-based Bayesian approaches, which potentially overcome these problems with a fully Bayesian treatment.

https://doi.org/10.1017/CBO9781107295360.008 Published online by Cambridge University Press