5 Evidence approximation

In a maximum a-posteriori (MAP) approximation as addressed in Chapter 4, we treat model parameters Θ of a target model M as unknown but deterministic variables which are estimated by maximizing the posterior probability $p(\Theta|\mathbf{O})$ using some observation data $\mathbf{O}$. Prior distribution $p(\Theta|\Psi)$ with heuristically determined hyperparameters Ψ is introduced. This is known as the *point estimation* of a target model. However, from a full Bayesian perspective, we treat model parameters as random variables where the randomness is represented by a prior distribution $p(\Theta|\Psi)$. In contrast with point estimation in a MAP approximation based on heuristic hyperparameters, the *distribution estimation* is implemented for full Bayesian learning. According to the distribution estimation, we find the whole prior distribution $p(\Theta|\Psi)$, or equivalently estimate the corresponding hyperparameters Ψ from observations $\mathbf{O}$ in an empirical fashion. In this implementation, the *marginalization* of likelihood function $p(\mathbf{O}|\Theta)$ with respect to model parameters Θ should be calculated for model construction, as follows:[1]

$$p(\mathbf{O}|\Psi) = \int p(\mathbf{O}|\Theta)p(\Theta|\Psi)d\Theta. \tag{5.1}$$

Rather than trusting the point estimate in MAP approximation, the resulting *evidence function* $p(\mathbf{O}|\Psi)$ in *evidence framework* considers all possible values of model parameters when making a prediction of $\mathbf{O}$ as new observation data or training data depending on the task. In cases of complicated model structure and coupled latent variables, this evidence function is prone to be intractable and should be approximated to estimate the optimal hyperparameters Ψ. For the applications of speech and language processing, we focus on acoustic modeling and language modeling in accordance with the distribution estimation based on evidence approximation.

In what follows, Section 5.1 first presents the evidence framework and addresses the type-2 maximum likelihood estimation for general pattern recognition. The optimal hyperparameters are estimated based on this framework. The optimization of evidence function or marginal likelihood is then extended to *sparse Bayesian learning* for acoustic modeling based on Bayesian sensing hidden Markov models (Saon & Chien 2012*a*) in Section 5.2. In this section, the scheme of automatic relevance determination is introduced and illustrated. In addition, evidence approximation to a hierarchical Dirichlet language model (MacKay & Peto 1995) is detailed in Section 5.3. The optimal

[1] This chapter regards model structure M and parameters of prior distribution Ψ as the hyperparameters in a broad sense. Equation (5.1) formulates the likelihood function given Ψ, but the discussion can be applied to the case of using M.

https://doi.org/10.1017/CBO9781107295360.006 Published online by Cambridge University Press

hyperparameters are obtained for the acoustic model as well as the language model. These Bayesian models are beneficial for noise-robust speech recognition and large vocabulary continuous speech recognition.

5.1 Evidence framework

This section begins with a general discussion of the evidence framework. We first review the well-known Bayes theorem, as discussed in Section 2.1. The evidence function $p(\mathbf{O}|\Psi)$, given prior parameters Ψ, is introduced in the Bayes theorem, which relates posterior distribution $p(\Theta|\mathbf{O}, \Psi)$ to the likelihood function $p(\mathbf{O}|\Theta)$, prior distribution $p(\Theta|\Psi)$, and the evidence function $p(\mathbf{O}|\Psi)$ by

$$\begin{aligned} p(\Theta|\mathbf{O}, \Psi) &= \frac{p(\mathbf{O}|\Theta)p(\Theta|\Psi)}{p(\mathbf{O}|\Psi)} \\ &= \frac{p(\mathbf{O}|\Theta)p(\Theta|\Psi)}{\int p(\mathbf{O}|\Theta)p(\Theta|\Psi)d\Theta}. \end{aligned} \tag{5.2}$$

In words:

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}. \tag{5.3}$$

The evidence function is a marginal likelihood function which takes all values of model parameters Θ into account. It is precisely the normalization term that appears in the denominator in Bayes' theorem as shown in Eq. (5.2).

Although the evidence function has appeared in the previous sections (e.g., the MAP approximation in Eq. (4.2)), it has not been explicitly considered so far. However, the evidence function $p(\mathbf{O}|\Psi)$ can directly link the hyperparameters Ψ and observations $\mathbf{O}$ by marginalizing the model parameters Θ, and can be used to infer the hyperparameters Ψ in the Bayesian framework.

5.1.1 Bayesian model comparison

This section also considers the model structure M in the evidence framework. In MacKay (1992*a*), the evidence framework was proposed to conduct a Bayesian model comparison, where the best model or model structure M is selected according to the posterior distribution of M as

$$\hat{M} = \arg\max_M p(M|\mathbf{O}) = \arg\max_M p(\mathbf{O}|M)p(M). \tag{5.4}$$

Here $p(M)$ is a prior distribution of the model structure M. In the case that each model is equally probable or has uniform probability, i.e., $p(M) =$ constant, different models M are ranked according to the evidence function $p(\mathbf{O}|M)$, which is obtained by marginalizing the model parameters Θ based on the product and sum rules as

$$\begin{aligned} p(\mathbf{O}|M) &= \int p(\mathbf{O}, \Theta|M)d\Theta \\ &= \int p(\mathbf{O}|\Theta, M)p(\Theta|M)d\Theta. \end{aligned} \tag{5.5}$$

https://doi.org/10.1017/CBO9781107295360.006 Published online by Cambridge University Press

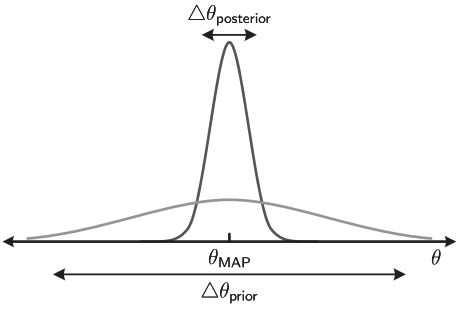

Figure 5.1 Evidence approximation. Adapted from Bishop (2006).

The marginalization of likelihood function over model parameters Θ or model structure is calculated to come out with a meaningful objective for model selection. However, we may obtain some insight into the model evidence by making a simple approximation to the integral over parameters Θ (MacKay 1992*a*, Bishop 2006). Consider the case of a model with a posterior distribution which is sharply peaked around the most probable value Θ_{MAP}, with the width $\triangle\Theta_{\text{posterior}}$. Then the posterior distribution is represented as $p(\Theta|\mathbf{O}) = \frac{1}{\triangle\Theta_{\text{posterior}}}$. Similarly, we also assume that the prior is flat with width $\triangle\Theta_{\text{prior}}$, and we have $p(\Theta) = \frac{1}{\triangle\Theta_{\text{prior}}}$. Therefore, by using the Bayes theorem, the evidence function in Eq. (5.5) can be approximated without solving the integral as:

$$\begin{aligned} p(\mathbf{O}) &= \int p(\mathbf{O}|\Theta)p(\Theta)d\Theta = \frac{p(\mathbf{O}|\Theta)p(\Theta)}{p(\Theta|\mathbf{O})} \\ &\approx p(\mathbf{O}|\Theta_{\text{MAP}})\frac{\triangle\Theta_{\text{posterior}}}{\triangle\Theta_{\text{prior}}}, \end{aligned} \tag{5.6}$$

where we omit the model structure index M for simplicity. The approximation to model evidence is illustrated by Figure 5.1. Without loss of generality, the notation Θ in Eq. (5.6) is treated as a single parameter. Taking the logarithm, we obtain

$$\log p(\mathbf{O}) \approx \log p(\mathbf{O}|\Theta^{\text{MAP}}) + \underbrace{\log\left(\frac{\triangle\Theta_{\text{posterior}}}{\triangle\Theta_{\text{prior}}}\right)}_{\text{Occam factor}}. \tag{5.7}$$

In this approximation, the first term represents the goodness-of-fit to the data $\mathbf{O}$ given the most probable parameter value Θ^{MAP}. The second term is known as the *Occam factor* (MacKay 1992*a*) which penalizes the model according to its complexity. Theoretically, we have the property $\triangle\Theta_{\text{posterior}} < \triangle\Theta_{\text{prior}}$. The Occam factor is negative and it increases in magnitude as the ratio $\frac{\triangle\Theta_{\text{posterior}}}{\triangle\Theta_{\text{prior}}}$ gets smaller. Thus, if parameters are finely tuned to the data in posterior distribution, then the penalty term is large. In practice, the model complexity or the Occam factor is multiplied by the number of adaptive parameters N in Θ. The optimal model complexity, as determined by the maximum evidence, is given by a trade-off between these two competing terms. A refined version of this evidence approximation could be further derived as:

https://doi.org/10.1017/CBO9781107295360.006 Published online by Cambridge University Press

$$\log p(\mathbf{O}) \approx \log p(\mathbf{O}|\Theta^{\text{MAP}}) + \underbrace{\log p(\Theta^{\text{MAP}}) + \frac{N}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{H}|}_{\text{Occam factor}}, \tag{5.8}$$

by using the *Laplace approximation*, which is detailed in Section 6.1. In Eq. (5.8), $\mathbf{H}$ is the Hessian matrix:

$$\mathbf{H} = -\nabla\nabla \log p(\mathbf{O}|\Theta^{\text{MAP}})p(\Theta^{\text{MAP}}) = -\nabla\nabla \log p(\Theta^{\text{MAP}}|\mathbf{O}), \tag{5.9}$$

which is the second derivative of the negative log posterior. The determinant of this matrix plays an important role in the Occam factor, which penalizes the model complexity.

We can attain further insight into Bayesian model comparison and understand how the marginal likelihood is favorable to the models with intermediate complexity by considering Figure 5.2 (Bishop 2006). Here, the horizontal axis is a one-dimensional representation of the data space $\mathbf{O}$. We consider three models M_1, M_2, and M_3 in which M_1 is the simplest and M_3 is the most complex. When generating a particular data set from a specific model, we first choose the values of the parameters from their prior distribution $p(\Theta)$. Then, given these parameter values, we sample the data from $p(\mathbf{O}|\Theta)$. A simple or *shallow* model has little variability and so will generate data sets using $p(\mathbf{O})$, which is confined to a relatively small region in space $\mathbf{O}$. In contrast, a complex or *deep* model can generate a variety of data sets, and so its distribution $p(\mathbf{O})$ is spread over a large region of the data space $\mathbf{O}$. In this example, for the particular observed data set $\mathbf{O}_M$, the model M_2 with intermediate complexity has the largest evidence.

5.1.2 Type-2 maximum likelihood estimation

The complexity of a model M is generally defined by the scope of data set $\mathbf{O}$ that model M could predict. This scope is not only determined by the model size (e.g., model structure, model order, or number of parameters N), but is also affected by the hyperparameters Ψ of the parameters Θ which are used to generate the observations $\mathbf{O}$. Instead of point estimation of model parameters Θ in MAP approximation based on heuristically determined hyperparameters, the evidence approximation conducts distribution estimation which determines the optimal prior distribution $\hat{p}(\Theta|\Psi)$ as a whole, or equivalently

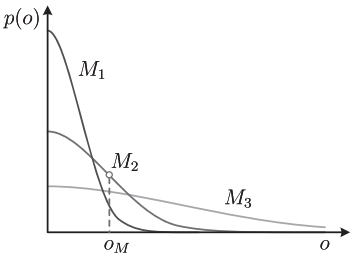

Figure 5.2 Model comparison based on the maximum evidence. Adapted from Bishop (2006).

https://doi.org/10.1017/CBO9781107295360.006 Published online by Cambridge University Press

infers the optimal hyperparameters $\hat{\Psi}$ corresponding to the prior distribution according to

$$\hat{\Psi} = \Psi^{\mathrm{ML2}} = \arg\max_{\Psi} p(\mathbf{O}|\Psi) = \arg\max_{\Psi} \int p(\mathbf{O}|\Theta, \Psi) p(\Theta|\Psi) d\Theta. \tag{5.10}$$

Differently from conventional ML estimation for model parameters Θ, the type-2 maximum likelihood (ML2) estimation aims to search the optimal hyperparameters Ψ^{ML2} from the observation data $\mathbf{O}$. In the literature, a point estimation based on ML or MAP conducts the so-called *level-1 inference*, while the distribution estimation based on ML2 undertakes the *level-2 inference* (MacKay 1995, Kwok 2000). The level-3 inference is performed to rank different models M according to the posterior probability $p(M|\mathbf{O})$ or the evidence function $p(\mathbf{O}|M)$ with equally probable model M. Three levels of inference can be iterated.

The evidence framework or the ML2 estimation has been developed for different learning machines including linear regression/classification networks (MacKay 1992*b*, Bishop 2006), feed-forward neural network (NN) (MacKay 1992*c*), support vector machine (SVM) (Kwok 2000), and hidden Markov model (Zhang, Liu, Chien *et al*. 2009). For the cases of linear regression and neural network regression models, the optimal hyperparameters of weight parameters and modeling errors are estimated by maximizing the evidence function of training data which is marginalized with respect to the weight parameters. In Zhang *et al*. (2009), the optimal hyperparameters were derived for the mixture model of exponential family distributions and then realized to build the robust HMMs for noisy speech recognition. Practically, these hyperparameters Ψ are interpreted as the *regularization parameter* λ, which plays a crucial role in the regularized regression models. The regularized models are developed to prevent the *over-fitting problem* in conventional models based on ML estimation or least-squares regression. In what follows, the evidence framework is illustrated to be closely connected to the *regularization theory*, which has been developed to regularize model structure and deal with the over-fitting problem when building generative models for speech recognition and other information systems.

5.1.3 Regularization in regression model

Model regularization is a fundamental issue in pattern recognition. It aims to estimate the smoothed parameters or construct a generalized model which has good prediction capability for future unseen data. The gap or mismatch between training data and test data can be compensated by tackling this issue. The over-fitting problem or the ill-posed problem is also resolved by following the regularization theory. In the regularized least-squares (RLS) model $f(\cdot)$, the over-fitting problem is avoided by incorporating a regularization term $E_w(\mathbf{w})$ into the training objective, and this penalizes too complex a model. Here, model parameters Θ are rewritten by an N-dimensional weight vector $\mathbf{w}$. The regularization term is determined by the weight parameters $\mathbf{w}$ and the corresponding

https://doi.org/10.1017/CBO9781107295360.006 Published online by Cambridge University Press

model structure M. The simplest form of this regularizer is given by a sum-of-squares of the parameter elements:

$$E_w(\mathbf{w}) = \frac{1}{2}\mathbf{w}^\mathsf{T}\mathbf{w} = \frac{1}{2}\|\mathbf{w}\|^2 = \frac{1}{2}\sum_{j=0}^{N-1} w_j^2. \tag{5.11}$$

The sum-of-squares error function is calculated by introducing training samples $\mathbf{O} = \{\mathbf{o}_t | t = 1, \cdots, T\}$ given by model parameters $\mathbf{w}$, i.e.

$$E_o(\mathbf{w}) = \frac{1}{2}\sum_{t=1}^{T}(f(\mathbf{o}_t, \mathbf{w}) - y_t)^2, \tag{5.12}$$

where y_t is the target value of the observation $\mathbf{o}_t$ at time t. Accordingly, the training objective for RLS parameters $\mathbf{w}^{\mathsf{RLS}}$ is yielded as a regularized least-squares function which is formed by

$$\mathbf{w}^{\mathsf{RLS}} = \arg\min_{\mathbf{w}} \{E_o(\mathbf{w}) + \lambda E_w(\mathbf{w})\}. \tag{5.13}$$

Notably, a regularization parameter λ is introduced in Eq. (5.13) to balance the trade-off between the sum-of-squares error function $E_o(\mathbf{w})$ and the model complexity penalty function $E_w(\mathbf{w})$. Minimization of the training objective in Eq. (5.13) eventually obtains a set of parameters $\mathbf{w}$ which works simultaneously towards the goals of fitting the data and reducing the norm of the solution. Regularization theory is beneficial for model selection. Regularization parameter λ is generally selected by applying the *cross validation* method using a small set of validation data which is outside the training data $\mathbf{O}$.

Nevertheless, it is more attractive to pursue Bayesian interpretation of model regularization. Considering the same regression problem, but now under a probabilistic framework, we assume that the modeling error $f(\mathbf{o}_t, \mathbf{w}) - y_t$ has a Gaussian distribution:

$$p(y_t|\mathbf{o}_t, \mathbf{w}, \beta) = \mathcal{N}(f(\mathbf{o}_t, w) - y_t)|0, \beta^{-1}), \tag{5.14}$$

and the parameter vector $\mathbf{w}$ comes from a Gaussian distribution

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}), \tag{5.15}$$

where $\mathbf{I}$ is the $N \times N$ identity matrix and β and α are the precision parameters for modeling error and parameter vector, respectively. Here, the hyperparameters $\Psi = \{\alpha, \beta\}$ consist of α and β. The MAP estimate of model parameters $\mathbf{w}^{\mathsf{MAP}}$ is obtained by maximizing the posterior distribution

$$p(\mathbf{w}|\mathbf{o}_t, y_t, \alpha, \beta) \propto p(y_t|\mathbf{o}_t, \mathbf{w}, \beta)p(\mathbf{w}|\alpha), \tag{5.16}$$

or equivalently minimizing the negative log posterior distribution,

$$\frac{\beta}{2}\sum_{t=1}^{T}\{f(\mathbf{o}_t, \mathbf{w}) - y_t\}^2 + \frac{\alpha}{2}\mathbf{w}^\mathsf{T}\mathbf{w}, \tag{5.17}$$

by using training samples $\mathbf{O} = \{\mathbf{o}_t | t = 1, \cdots, T\}$. It is interesting to see that maximizing the posterior distribution in Eq. (5.17) is equivalent to minimizing the regularized least-squares error function in Eq. (5.13) with a regularization parameter $\lambda = \alpha/\beta$.

https://doi.org/10.1017/CBO9781107295360.006 Published online by Cambridge University Press

Readers may refer to MacKay (1992*c*) and Bishop (2006) for detailed solution to optimal hyperparameters $\Psi^{\mathsf{ML2}} = \{\alpha^{\mathsf{ML2}}, \beta^{\mathsf{ML2}}\}$ of a linear regression model with regression function

$$f(\mathbf{o}_t, \mathbf{w}) = w_0 + \sum_{j=1}^{N-1} w_j \phi_j(\mathbf{o}_t) = \mathbf{w}^{\mathsf{T}} \boldsymbol{\phi}(\mathbf{o}_t). \tag{5.18}$$

A linear combination of fixed non-linear functions of the input variable $\mathbf{o}_t$ is considered. Here, $\boldsymbol{\phi} = [\phi_0, \cdots, \phi_{N-1}]^{\mathsf{T}}$ denotes the basis functions and $\phi_0(\mathbf{o}_t) = 1$ is assigned. For the case of a neural network regression model, ML2 estimation of hyperparameters is addressed in Section 6.4.

The comparison among RLS estimation, MAP estimation, and ML2 estimation is further investigated below. According to RLS estimation, level-1 inference is performed to find model parameters $\mathbf{w}^{\mathsf{RLS}}$ by using training data $\mathbf{O}$ while the hyperparameter λ is estimated in level-2 inference via a cross validation scheme by using additional *validation data*. However, Bayesian inference is implemented to calculate model parameters $\mathbf{w}^{\mathsf{MAP}}$ in level-1 inference based on MAP estimation. In level-2 inference, the hyperparameters $\Psi^{\mathsf{ML2}} = \{\alpha^{\mathsf{ML2}}, \beta^{\mathsf{ML2}}\}$ are inferred using ML2 estimation. By applying MAP and ML2 methods, the same training data $\mathbf{O}$ are used to estimate parameters $\mathbf{w}^{\mathsf{MAP}}$ and hyperparameters Ψ^{ML2}, respectively. It is desirable that *no validation data are required* by using the Bayesian approach. Besides, RLS and MAP methods fulfil the point estimation and assume that the estimates $\hat{\mathbf{w}} = \mathbf{w}_{\mathsf{RLS}}$ and $\hat{\mathbf{w}} = \mathbf{w}^{\mathsf{MAP}}$ are true values for prediction of new data $\mathbf{O}$ in a test phase according to likelihood function $p(\mathbf{O}|\hat{\mathbf{w}})$. Instead of relying on single parameter values $\hat{\mathbf{w}}$ in the RLS or MAP method, the ML2 method implements the distribution estimation and directly infers the optimal hyperparameters $\hat{\Psi} = \Psi^{\mathsf{ML2}}$ by maximizing the predictive distribution or the marginal likelihood of training data $p(\mathbf{O}|\Psi) = \int p(\mathbf{O}|\mathbf{w}, \Psi) p(\mathbf{w}|\Psi) d\mathbf{w}$, where all possible values of parameters $\mathbf{w}$ are considered. In a test phase, the same marginal distribution $p(\mathbf{O}|\hat{\Psi})$, given the estimated hyperparameters $\hat{\Psi}$, is applied for prediction of new data $\mathbf{O}$.

5.1.4 Evidence framework for HMM and SVM

Although the evidence framework is only addressed for a linear regression model $f(\cdot)$, extensions to the classification models including HMM and SVM have been proposed in Zhang *et al*. (2009) and Kwok (2000), respectively. When ML2 estimation is applied for the HMM framework, we estimate the hyperparameters of continuous-density HMM parameters including initial state probabilities $\{\pi_j\}$, state transition probabilities $\{a_{ij}\}$, mixture weights $\{\omega_{jk}\}$, mean vectors $\{\boldsymbol{\mu}_{jk}\}$, and covariance matrices $\{\boldsymbol{\Sigma}_{jk}\}$. For the probability parameters $\{\pi_j\}$, $\{a_{ij}\}$ and $\{\omega_{jk}\}$ in multinomial distributions, ML2 estimation is performed to find the corresponding hyperparameters Ψ which are the parameters of *Dirichlet priors* for multinomial or discrete variables of states j, state pairs (i, j), and mixture components k, respectively. For the remaining Gaussian parameters $\{\boldsymbol{\mu}_{jk}\}$ and $\{\boldsymbol{\Sigma}_{jk}\}$ of continuous feature vectors $\{\mathbf{o}_t\}$, ML2 estimation aims to calculate the corresponding hyperparameters Ψ which are the parameters of *Gaussian–Wishart priors* for Gaussian mean vectors $\{\boldsymbol{\mu}_{jk}\}$ and precision (or inverse covariance) matrices $\{\boldsymbol{\Sigma}_{jk}^{-1}\}$. In general,

https://doi.org/10.1017/CBO9781107295360.006 Published online by Cambridge University Press

Dirichlet distribution is known as the *conjugate prior* for multinomial variables while Gaussian–Wishart distribution is seen as the conjugate prior for Gaussian variables. By following this guideline, the closed-form solution to the integral in marginal likelihood does exist, so that the optimization of marginal distribution with respect to individual hyperparameters has an analytical solution. These hyperparameters characterize the uncertainties of HMM parameters which could be applied for robust speech recognition according to the Bayesian predictive classification as addressed in Eq. (3.16) and Section 6.3. This approach is different from conventional BPC based on the hyperparameters which are heuristically determined or calculated in an ensemble way (Huo & Lee 2000, Chien & Liao 2001).

The support vector machine (SVM) approach is based on the idea of structural risk minimization, which shows that the generalization error is bounded by the sum of the training set error and the Vapnik-Chervonenkis (VC) dimension of the learning machine (Abu-Mostafa 1989, Vapnik 1995). By minimizing this upper bound, generalization to future data is improved. Generalization error is related not to the number of inputs, but to the margin with which it separates the data. SVM has been successfully applied in many classification problems including speech recognition (Ganapathiraju, Hamaker & Picone 2004). Although SVM is a nonparametric method, the probabilistic framework and Bayesian perspective have been introduced to deal with the selection of two tuning parameters or hyperparameters, including:

- a regularization parameter λ, which determines the trade-off between minimizing the training errors and minimizing the model complexity;
- a kernel parameter, which implicitly defines the high dimensional feature space to be used.

Conventionally, these hyperparameters are empirically selected by hand or via cross validation. The evidence framework has been applied to find the optimal regularization parameter for SVM (Kwok 2000). Next, we address the detailed estimation of hyperparameters for two practical solutions to speech recognition. One is developed for sparse Bayesian acoustic modeling while the other is proposed for hierarchical Dirichlet language modeling.

5.2 Bayesian sensing HMMs

Speech recognition systems are usually constructed by collecting large amounts of training data and estimating a large number of model parameters to achieve the desired recognition accuracy on test data. A large set of context-dependent Gaussian components (several hundred thousand components is usually the norm) is trained to build context-dependent phone models. GMMs with Gaussian mean vectors and diagonal covariance matrices may not be an accurate representation of high dimensional acoustic features. The Gaussian components may be overdetermined. The mismatch between training data and test conditions may not be carefully compensated. The uncertainty

https://doi.org/10.1017/CBO9781107295360.006 Published online by Cambridge University Press

of estimated HMM parameters may not be properly characterized. A Bayesian learning approach is introduced to tackle these issues based on the *basis representation*. ML2 estimation is conducted to estimate the automatic relevance determination (ARD) parameter (MacKay 1995, Tipping 2001) which is the state-dependent hyperparameter of weight parameter in basis representation. Sparse Bayesian learning is performed by using the ARD parameter.

5.2.1 Basis representation

An acoustic feature vector $\mathbf{o}$ can be viewed as lying in a vector space spanned by a set of basis vectors. Such a basis representation has been popular for regression problems in machine learning and for signal recovery in the signal processing literature. This approach is now increasingly important for acoustic feature representation. Compressive sensing and sparse representation are popular topics in the signal processing community. The basic idea of compressive sensing is to encode a feature vector $\mathbf{o} \in \mathbb{R}^D$ based on a set of over-determined dictionary or basis vectors $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_N]$ via

$$\mathbf{o} = w_1\boldsymbol{\phi}_1 + \cdots + w_N\boldsymbol{\phi}_N = \boldsymbol{\Phi}\mathbf{w}, \tag{5.19}$$

where the sensing weights $\mathbf{w} = [w_1, \cdots, w_N]^\mathsf{T}$ are sparse and the basis vectors $\boldsymbol{\Phi}$ are formed by training samples. A relatively small set of relevant basis vectors is used for sparse representation based on this exemplar-based method. The sparse solution to $\mathbf{w}$ can be derived by optimizing the ℓ_1-regularized objective function (Sainath, Ramabhadran, Picheny *et al*. 2011). However, the exemplar-based method is a memory-based method, which is time-consuming with high memory cost. It is also important to integrate HMMs into sparse representation of continuous speech frames $\mathbf{O} = \{\mathbf{o}_t | t = 1, \cdots, T\}$.

5.2.2 Model construction

Bayesian sensing HMMs (BS-HMMs) (Saon & Chien 2012*a*) are developed by incorporating Markov chains into the basis representation of continuous speech. A Bayesian sensing framework is presented for speech recognition. The underlying aspect of BS-HMMs is to measure an observed feature vector $\mathbf{o}_t$ of a speech sentence $\mathbf{O}$ based on a compact set of state-dependent dictionary $\boldsymbol{\Phi}_j = \left[\boldsymbol{\phi}_{j1}, \cdots, \boldsymbol{\phi}_{jN}\right]$ at state j. For each frame, the reconstruction error between measurement $\mathbf{o}_t$ and its representation $\boldsymbol{\Phi}_j\mathbf{w}_t$, where $\mathbf{w}_t = [w_{t1}, \cdots, w_{tN}]^\mathsf{T}$, is assumed to be Gaussian distributed with zero mean and a state-dependent covariance matrix or inverse precision matrix $\boldsymbol{\Sigma}_j = \mathbf{R}_j^{-1}$. The state likelihood function with time-dependent sensing weights $\mathbf{w}_t$ is defined by

$$p(\mathbf{o}_t|\Theta_j) \triangleq \mathcal{N}(\mathbf{o}_t|\boldsymbol{\Phi}_j\mathbf{w}_t, \mathbf{R}_j^{-1}). \tag{5.20}$$

The Bayesian perspective in BS-HMMs has its origin from the relevance vector machine (RVM) (Tipping 2001). Figure 5.3 illustrates the graphical model based on BS-HMMs. The RVM is known as a sparse Bayesian learning approach for regression and classification problems. We would like to apply RVM to conduct sparse basis representation and combine it with HMMs to characterize the dynamics in the time domain. Therefore,

https://doi.org/10.1017/CBO9781107295360.006 Published online by Cambridge University Press

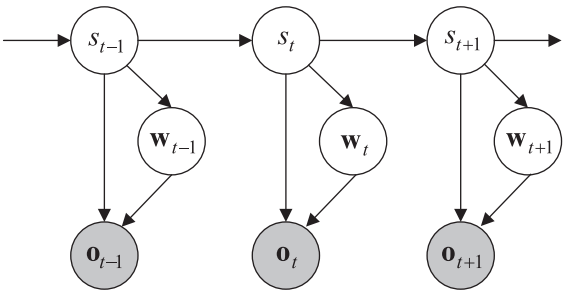

Figure 5.3 Graphical model of BS-HMM.

BS-HMM parameters are obtained by $\Theta = \{\pi_j, a_{ij}, \mathbf{\Phi}_j, \mathbf{R}_j\}$. Obviously, similarly to conventional GMM-based HMMs, we can extend BS-HMM to deal with a *mixture model* for basis representation where the mixture weight ω_{jk}, the basis vectors $\mathbf{\Phi}_{jk}$, and the precision matrix $\mathbf{R}_{jk}$ of individual mixture component k are incorporated. In what follows, we neglect the extension to a mixture model and exclude the time-dependent weight parameters $\mathbf{w}_t$ from the parameter set Θ.

5.2.3 Automatic relevance determination

However, the sensing weights $\mathbf{w}_t$ play a crucial role in basis representation, and so we introduce Bayesian compressive sensing (Ji, Xue & Carin 2008) for acoustic modeling. The idea of Bayesian learning in BS-HMMs is to yield the *distribution estimates* of the speech feature vectors $\mathbf{o}_t$ due to the variations of sensing weights $\mathbf{w}_t$ in basis representation. A Gaussian prior with zero mean and state-dependent diagonal precision matrix $\mathcal{A}_j = \text{diag}\{\alpha_{jn}\}$ is introduced to characterize the weight vector, i.e.

$$\begin{aligned} p\left(\mathbf{w}_t|\mathcal{A}_j\right) &= \mathcal{N}(\mathbf{w}_t|\mathbf{0}, \text{diag}\{\alpha_{jn}^{-1}\}) \\ &= \prod_{n=1}^{N} \mathcal{N}(w_{tn}|0, \alpha_{jn}^{-1}). \end{aligned} \tag{5.21}$$

Considering a *hierarchical prior* model where precision parameter α_{jn} is represented by a gamma prior with parameters a and b in Appendix C.11:

$$p(\alpha_{jn}) = \text{Gam}(\alpha_{jn}|a, b) = \frac{1}{\Gamma(a)} b^a \alpha_{jn}^{a-1} \exp(-b\alpha_{jn}). \tag{5.22}$$

The marginal prior distribution is derived as a Student's t-distribution as defined in Appendix C.16:

$$\begin{aligned} p(w_{tn}|a, b) &= \int_0^\infty \mathcal{N}(w_{tn}|0, \alpha_{jn}^{-1}) \text{Gam}(\alpha_{jn}|a, b) d\alpha_{jn} \\ &= \frac{b^a}{\Gamma(a)} \left(\frac{1}{2\pi}\right)^{1/2} \left(b + \frac{w_{tn}^2}{2}\right)^{-a-1/2} \Gamma(a + 1/2) \\ &= \text{St}\left(w_{tn} \middle| 0, \frac{b}{a}, 2a\right), \end{aligned} \tag{5.23}$$

https://doi.org/10.1017/CBO9781107295360.006 Published online by Cambridge University Press

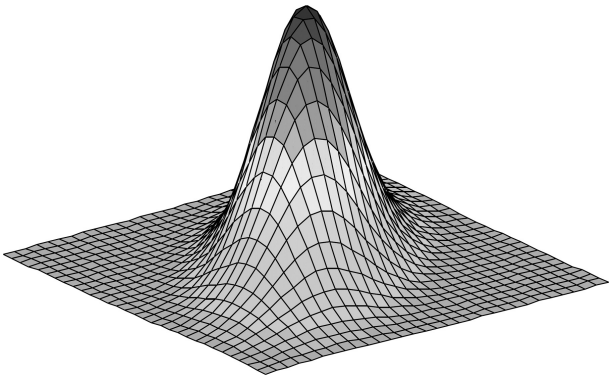

Figure 5.4 An example of two-dimensional Gaussian distribution.

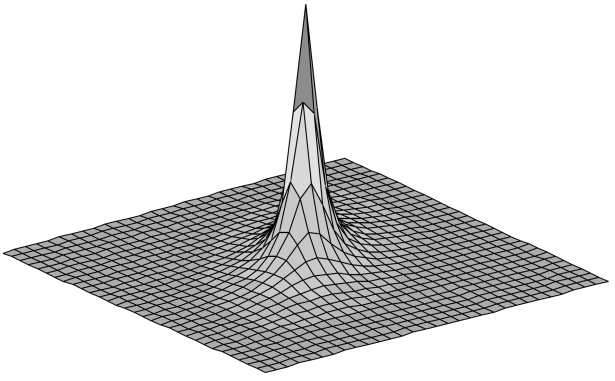

Figure 5.5 An example of two-dimensional Student's t-distribution.

which is known as a *sparse prior* distribution, since this distribution has a heavy tail and steep peak. Student's t-distribution, illustrated in Figure 5.5, is a heavy tailed distribution and is more robust to outliers than a Gaussian distribution (Figure 5.4) (Bishop 2006).

Correspondingly, if the precision parameter α_{jn} in $\mathcal{N}(w_{tn}|0, \alpha_{jn}^{-1})$ is large, the weight parameter w_{tn} is likely to be zero, $w_{tn} \rightarrow 0$, which implies that the associated basis vector $\boldsymbol{\phi}_{jn}$ is irrelevant to Bayesian basis representation of a target observation $\mathbf{o}_t$. The physical meaning of automatic relevance determination (ARD) is then reflected by the precision parameter α_{jn} in state-dependent hyperparameters, $\Psi = \{\mathcal{A}_j = \text{diag}\{\alpha_{jn}\}\}$. We simply call α_{jn} the ARD parameter. According to the ARD scheme, only relevant basis vectors are selected to represent sequence data $\mathbf{O}$. Sparse Bayesian learning (SBL) can be realized by using an ARD scheme (Tipping 2001). In the implementation, the values of state-dependent ARD parameters $\{\alpha_{j1}, \cdots, \alpha_{jN}\}$ can be used to rank or select salient basis vectors $\{\boldsymbol{\phi}_{j1}, \cdots, \boldsymbol{\phi}_{jN}\}$ which are relevant to a target HMM state j. The larger the estimated value α_{jn}, the more likely it is that the basis vector $\boldsymbol{\phi}_{jn}$ should be pruned from the parameter set. The compressed model can be achieved by applying this property (Saon & Chien 2011). The ARD parameter serves as a *compression factor* for

https://doi.org/10.1017/CBO9781107295360.006 Published online by Cambridge University Press

model complexity control. One can initially train a large model and then prune it to a smaller size by removing basis elements which correspond to the larger ARD values. Although we appear to be utilizing a non-sparse Gaussian prior over the weights, in truth the hierarchical formulation implies that the real weight prior is clearly recognized as encouraging sparsity. Considering this sparse Bayesian basis representation, BS-HMM parameters and hyperparameters are formed by $\{\Theta, \Psi\} = \{\pi_j, a_{ij}, \mathbf{\Phi}_j, \mathbf{R}_j, \mathcal{A}_j | j = 1, \cdots, J\}$ consisting of initial state probability π_j, state transition probability a_{ij}, basis vectors $\mathbf{\Phi}_j$, and precision matrices of reconstruction errors $\mathbf{R}_j$ and sensing weights $\mathcal{A}_j$. Level-1 inference and level-2 inference are done simultaneously in BS-HMMs.

5.2.4 Model inference

We estimate BS-HMM parameters and hyperparameters from the observed speech data $\mathbf{O}$ according to the type-2 ML (ML2) estimation:

$$\{\pi_j^{\mathsf{ML2}}, a_{ij}^{\mathsf{ML2}}, \mathbf{\Phi}_j^{\mathsf{ML2}}, \mathbf{R}_j^{\mathsf{ML2}}, \mathcal{A}_j^{\mathsf{ML2}}\} = \arg \max_{\{\pi_j, a_{ij}, \mathbf{\Phi}_j, \mathbf{R}_j, \mathcal{A}_j\}} p(\mathbf{O}|\{\pi_j, a_{ij}, \mathbf{\Phi}_j, \mathbf{R}_j, \mathcal{A}_j\}). \quad (5.24)$$

Without loss of generality, we view Eq. (5.24) as the ML2 estimation because the marginal likelihood with respect to sensing weights $\mathbf{w}_t$ is calculated at each frame t in likelihood function $p(\mathbf{O}|\{\pi_j, a_{ij}, \mathbf{\Phi}_j, \mathbf{R}_j, \mathcal{A}_j\})$. However, the marginalization over π_j, a_{ij}, $\mathbf{\Phi}_j$, and $\mathbf{R}_j$ is not considered. Since the optimization procedure is affected by an incomplete data problem, the EM algorithm (Dempster *et al*. 1976) is applied to find the optimal solution to $\{\pi_j, a_{ij}, \mathbf{\Phi}_j, \mathbf{R}_j, \mathcal{A}_j\}$. In E-step, an auxiliary function is calculated by averaging the log likelihood function of the new estimates $\{\Theta', \Psi'\}$, given the old estimates $\{\Theta, \Psi\}$ over all latent variables $\{S, V\}$:

$$\begin{aligned} Q(\Theta', \Psi'|\Theta, \Psi) &= \mathbb{E}_{(S,V)}[\log p(\mathbf{O}, S, V|\Theta', \Psi')|\mathbf{O}, \Theta, \Psi] \\ &= \sum_S \sum_V p(S, V|\mathbf{O}, \Theta, \Psi) \log p(\mathbf{O}, S, V|\Theta', \Psi'). \end{aligned} \quad (5.25)$$

In the M-step, we maximize the auxiliary function with respect to new parameters and hyperparameters $\{\Theta', \Psi'\}$,

$$\{\Theta', \Psi'\} = \arg \max_{\{\Theta', \Psi'\}} Q(\Theta', \Psi'|\Theta, \Psi), \quad (5.26)$$

to find optimal parameters and hyperparameters. The auxiliary function is expanded by

$$\begin{aligned} &\sum_S \sum_V p(S, V|\mathbf{O}, \Theta, \Psi) \left[\sum_{t=1}^{T} (\log a'_{s_{t-1}s_t} + \log p(\mathbf{o}_t|\Theta'_{s_t}, \Psi'_{s_t})) \right] \\ &= \sum_j \sum_{t=1}^{T} \gamma_t(j) \left[\log a'_{s_{t-1}j} + \log \int p(\mathbf{o}_t|\mathbf{w}_t, \mathbf{\Phi}'_j, \mathbf{R}'_j) p(\mathbf{w}_t|\mathcal{A}'_j) d\mathbf{w}_t \right], \end{aligned} \quad (5.27)$$

where $\gamma_t(j) = p(s_t = j|\mathbf{O}, \Theta, \Psi)$ is the posterior probability of being in state j at time t given the current parameters and hyperparameters $\{\Theta, \Psi\}$ generating measurements $\mathbf{O}$.

https://doi.org/10.1017/CBO9781107295360.006 Published online by Cambridge University Press

We tacitly use the convention $a_{s_0 s_1} = \pi_{s_1}$. Since the estimation of initial state probability π_j and state transition probability a_{ij} is the same as that in HMMs, we neglect the estimation of $\{\pi_j, a_{ij}\}$ hereafter.

5.2.5 Evidence function or marginal likelihood

The key issue in the E-step is to calculate the frame-based evidence function or marginal likelihood $p(\mathbf{o}_t|\Theta'_{s_t}, \Psi'_{s_t}) = p(\mathbf{o}_t|\mathbf{\Phi}'_j, \mathbf{R}'_j, \mathcal{A}'_j)$, which is marginalized over sensing weights $\mathbf{w}_t$ at state $s_t = j$ and is proportional to

$$
\begin{aligned}
&\int |\mathbf{R}'_j|^{1/2} \exp\left[-\frac{1}{2}(\mathbf{o}_t - \mathbf{\Phi}'_j\mathbf{w}_t)^\mathsf{T}\mathbf{R}'_j(\mathbf{o}_t - \mathbf{\Phi}'_j\mathbf{w}_t)\right] \\
&\qquad \times |\mathcal{A}'_j|^{1/2} \exp\left[-\frac{1}{2}\mathbf{w}_t^\mathsf{T}\mathcal{A}'_j\mathbf{w}_t\right] d\mathbf{w}_t \\
&= |\mathbf{R}'_j|^{1/2}|\mathcal{A}'_j|^{1/2} \int \exp\Bigg[-\frac{1}{2}\Big(\mathbf{o}_t^\mathsf{T}\mathbf{R}'_j\mathbf{o}_t \\
&\qquad - 2\mathbf{o}_t^\mathsf{T}\mathbf{R}'_j\mathbf{\Phi}'_j\mathbf{w}_t + \mathbf{w}_t^\mathsf{T}\left((\mathbf{\Phi}'_j)^\mathsf{T}\mathbf{R}'_j\mathbf{\Phi}'_j + \mathcal{A}'_j\right)\mathbf{w}_t\Big)\Bigg] d\mathbf{w}_t \\
&= |\mathbf{R}'_j|^{1/2}|\mathcal{A}'_j|^{1/2} \exp\left[-\frac{1}{2}(\mathbf{o}_t^\mathsf{T}\mathbf{R}'_j\mathbf{o}_t)\right] \\
&\qquad \times \int \exp\left[-\frac{1}{2}\begin{pmatrix} \mathbf{w}_t^\mathsf{T}(\mathbf{\Sigma}'_j)^{-1}\mathbf{w}_t - 2(\mathbf{o}_t^\mathsf{T}\mathbf{R}'_j\mathbf{\Phi}'_j\mathbf{\Sigma}'_j)(\mathbf{\Sigma}'_j)^{-1}\mathbf{w}_t \\ +(\mathbf{o}_t^\mathsf{T}\mathbf{R}'_j\mathbf{\Phi}'_j\mathbf{\Sigma}'_j)(\mathbf{\Sigma}'_j)^{-1}(\mathbf{\Sigma}'_j(\mathbf{\Phi}'_j)^\mathsf{T}\mathbf{R}'_j\mathbf{o}_t) \\ -(\mathbf{o}_t^\mathsf{T}\mathbf{R}'_j\mathbf{\Phi}'_j\mathbf{\Sigma}'_j)(\mathbf{\Sigma}'_j)^{-1}(\mathbf{\Sigma}'_j(\mathbf{\Phi}'_j)^\mathsf{T}\mathbf{R}'_j\mathbf{o}_t) \end{pmatrix}\right] d\mathbf{w}_t \\
&\propto |\mathbf{R}'_j|^{1/2}|\mathcal{A}'_j|^{1/2}|\mathbf{\Sigma}'_j|^{1/2} \exp\left[-\frac{1}{2}\mathbf{o}_t^\mathsf{T}(\mathbf{R}'_j - \mathbf{R}'_j\mathbf{\Phi}'_j\mathbf{\Sigma}'_j(\mathbf{\Phi}'_j)^\mathsf{T}\mathbf{R}'_j)\mathbf{o}_t\right] \\
&= |\mathbf{R}'_j|^{1/2}|\mathcal{A}'_j|^{1/2}|\mathbf{\Sigma}'_j|^{1/2} \exp\left[-\frac{1}{2}(\mathbf{o}_t^\mathsf{T}\mathbf{R}'_j\mathbf{o}_t - (\mathbf{m}'_{tj})^\mathsf{T}(\mathbf{\Sigma}'_j)^{-1}\mathbf{m}'_{tj})\right].
\end{aligned} \tag{5.28}
$$

In Eq. (5.28), the notations

$$(\mathbf{\Sigma}'_j)^{-1} \triangleq (\mathbf{\Phi}'_j)^\mathsf{T}\mathbf{R}'_j\mathbf{\Phi}'_j + \mathcal{A}'_j, \tag{5.29}$$

$$\mathbf{m}'_{tj} \triangleq \mathbf{\Sigma}'_j(\mathbf{\Phi}'_j)^\mathsf{T}\mathbf{R}'_j\mathbf{o}_t, \tag{5.30}$$

are introduced, and the integral of a Gaussian distribution

$$\mathcal{N}(\mathbf{w}_t|\mathbf{\Sigma}'_j(\mathbf{\Phi}'_j)^\mathsf{T}\mathbf{R}'_j\mathbf{o}_t, \mathbf{\Sigma}'_j) \tag{5.31}$$

is manipulated. By applying the Woodbury matrix inversion (Eq. (B.20))

$$(\mathbf{A} + \mathbf{UCV})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}(\mathbf{C}^{-1} + \mathbf{VA}^{-1}\mathbf{U})^{-1}\mathbf{VA}^{-1}, \tag{5.32}$$

given the dimensionally compatible matrices $\mathbf{A}$, $\mathbf{U}$, $\mathbf{C}$, and $\mathbf{V}$, the marginal likelihood $p(\mathbf{o}_t|\mathbf{\Phi}'_j, \mathbf{R}'_j, \mathcal{A}'_j)$ is derived as

$$
\begin{aligned}
&\mathcal{N}(\mathbf{o}_t|\mathbf{0}, (\mathbf{R}'_j - \mathbf{R}'_j\mathbf{\Phi}'_j\left((\mathbf{\Phi}'_j)^\mathsf{T}\mathbf{R}'_j\mathbf{\Phi}'_j + \mathcal{A}'_j\right)^{-1}(\mathbf{\Phi}'_j)^\mathsf{T}\mathbf{R}'_j)^{-1}) \\
&\qquad = \mathcal{N}(\mathbf{o}_t|\mathbf{0}, (\mathbf{R}'_j)^{-1} + \mathbf{\Phi}'_j(\mathcal{A}'_j)^{-1}(\mathbf{\Phi}'_i)^\mathsf{T}),
\end{aligned} \tag{5.33}
$$

https://doi.org/10.1017/CBO9781107295360.006 Published online by Cambridge University Press

which is a Gaussian likelihood function with zero mean. The equality of the determinant

$$|\mathbf{R}'_j - \mathbf{R}'_j\mathbf{\Phi}'_j\mathbf{\Sigma}'_j(\mathbf{\Phi}'_j)^{\mathsf{T}}\mathbf{R}'_j| = |\mathbf{R}'_j||\mathcal{A}'_j||\mathbf{\Sigma}'_j| \tag{5.34}$$

is held. This implies that frame discrimination among different states is done solely on the basis of the covariance matrix. Apparently, the covariance matrix $(\mathbf{R}'_j)^{-1} + \mathbf{\Phi}'_j(\mathcal{A}'_j)^{-1}(\mathbf{\Phi}'_i)^{\mathsf{T}}$ is positive definite, so that $p(\mathbf{o}_t|\mathbf{\Phi}'_j, \mathbf{R}'_j, \mathcal{A}'_j)$ is a valid probability density function. For diagonal $\mathbf{R}'_j$, the marginal likelihood is seen as a new Gaussian distribution with a factor analyzed covariance matrix $(\mathbf{R}'_j)^{-1} + \mathbf{\Phi}'_j(\mathcal{A}'_j)^{-1}(\mathbf{\Phi}'_j)^{\mathsf{T}}$, where the factor loading matrix $\mathbf{\Phi}'_j(\mathcal{A}'_j)^{-1/2}$ is seen as a rank-N correction to $(\mathbf{R}'_j)^{-1}$ (Saon & Chien 2011).

5.2.6 Maximum a-posteriori sensing weights

In BS-HMMs, we can determine the maximum a-posteriori (MAP) estimate of Bayesian sensing weights $\mathbf{w}_t^{\mathsf{MAP}}$ for each observation $\mathbf{o}_t$ from

$$\begin{aligned}\mathbf{w}_t^{\mathsf{MAP}} &= \arg\max_{\mathbf{w}_t}\ p(\mathbf{w}_t|\mathbf{o}_t, \Theta', \Psi') \\ &= \arg\max_{\mathbf{w}_t}\ p(\mathbf{o}_t|\mathbf{w}_t, \mathbf{\Phi}'_j, \mathbf{R}'_j)p(\mathbf{w}_t|\mathcal{A}'_j) \\ &= \mathbf{\Sigma}'_j(\mathbf{\Phi}'_j)^{\mathsf{T}}\mathbf{R}'_j\mathbf{o}_t \triangleq \mathbf{m}_{tj},\end{aligned} \tag{5.35}$$

which is seen as a weighted product in vector space of observation $\mathbf{o}_t$ and transposed basis vectors $(\mathbf{\Phi}'_j)^{\mathsf{T}}$. The notations $\mathbf{m}'_{tj}$ (or equivalently $\mathbf{w}_t^{\mathsf{MAP}}$) and $\mathbf{\Sigma}'_j$ are the mean vector and the covariance matrix of the posterior distribution $p(\mathbf{w}_t|\mathbf{o}_t, \Theta', \Psi')$, respectively. The precision matrix for $\mathbf{w}_t$ is modified from $\mathcal{A}'_j$ of a-priori density $p(\mathbf{w}_t)$ to $(\mathbf{\Phi}'_j)^{\mathsf{T}}\mathbf{R}'_j\mathbf{\Phi}'_j + \mathcal{A}'_j$ of the a-posteriori distribution $p(\mathbf{w}_t|\mathbf{o}_t)$. The difference term $(\mathbf{\Phi}'_j)^{\mathsf{T}}\mathbf{R}'_j\mathbf{\Phi}'_j$ comes from the likelihood function $p(\mathbf{o}_t|\mathbf{w}_t)$, and is caused by the measurement $\mathbf{\Phi}'_j\mathbf{w}_t$ for observation $\mathbf{o}_t$ at frame t represented by new basis vectors $\mathbf{\Phi}'_j$ of state j. This is meaningful because Bayesian learning performs subjective inference, naturally increasing the model precision.

5.2.7 Optimal parameters and hyperparameters

By substituting Eq. (5.28) into Eq. (5.27), the optimal BS-HMM parameters and hyperparameters are estimated by maximizing the expanded auxiliary function with respect to individual parameters and hyperparameters $\{\mathbf{\Phi}'_j, \mathbf{R}'_j, \mathcal{A}'_j\}$. The auxiliary function in a BS-HMM state is simplified to

$$\begin{aligned}&Q(\mathbf{\Phi}'_j, \mathbf{R}'_j, \mathcal{A}'_j|\mathbf{\Phi}_j, \mathbf{R}_j, \mathcal{A}_j) \\ &\quad= \sum_{t=1}^{T}\gamma_t(j)\log\int p(\mathbf{o}_t|\mathbf{w}_t, \mathbf{\Phi}'_j, \mathbf{R}'_j)p(\mathbf{w}_t|\mathcal{A}'_j)d\mathbf{w}_t \\ &\quad\propto \sum_{t=1}^{\mathsf{T}}\gamma_t(j)\Big[\log|\mathbf{R}'_j| + \log|\mathcal{A}'_j| \\ &\qquad\qquad + \log|\mathbf{\Sigma}_j| - \mathbf{o}_t^{\mathsf{T}}\mathbf{R}'_j\mathbf{o}_t + (\mathbf{m}'_{tj})^{\mathsf{T}}(\mathbf{\Sigma}'_j)^{-1}\mathbf{m}'_{tj}\Big].\end{aligned} \tag{5.36}$$

https://doi.org/10.1017/CBO9781107295360.006 Published online by Cambridge University Press

Let us first consider the maximization of Eq. (5.28) with respect to $N \times N$ hyperparameter matrix $\mathcal{A}'_j$. We take the gradient of Eq. (5.28) with respect to $\mathcal{A}'_j$ and set it to zero to obtain

$$\sum_{t=1}^{T} \gamma_t(j) \left[(\mathcal{A}'_j)^{-1} - \mathbf{\Sigma}'_j - \underbrace{\mathbf{\Sigma}'_j (\mathbf{\Phi}'_j)^\mathsf{T} \mathbf{R}'_j \mathbf{o}_t}_{\mathbf{m}'_{tj}} \cdot \underbrace{\mathbf{o}_t^\mathsf{T} \mathbf{R}'_j \mathbf{\Phi}'_j \mathbf{\Sigma}'_j}_{(\mathbf{m}'_{tj})^\mathsf{T}} \right] = 0. \tag{5.37}$$

The value of $\mathcal{A}'_j$ that maximizes the auxiliary function satisfies

$$\begin{aligned} (\mathcal{A}_j^{\mathsf{ML2}})^{-1} &= \mathbf{\Sigma}'_j + \frac{\sum_{t=1}^{T} \gamma_t(j) \mathbf{m}'_{tj} (\mathbf{m}'_{tj})^\mathsf{T}}{\sum_{t=1}^{T} \gamma_t(j)} \\ &\triangleq \mathbf{F}^{\mathsf{ML2}}(\mathcal{A}'_j). \end{aligned} \tag{5.38}$$

Notably, Eq. (5.38) is an implicit solution to $\mathcal{A}'_j$ because $\mathbf{F}^a$ is a function of $\mathcal{A}'_j$. The hyperparameter $(\mathcal{A}_j^{\mathsf{ML2}})^{-1}$ of sensing weights is obtained by adding the covariance matrix $\mathbf{\Sigma}'_j$ of posterior $p(\mathbf{w}_t | \mathbf{o}_t, \Theta', \Psi')$ and the weighted autocorrelation of MAP sensing weights $\{\mathbf{m}'_{tj} \triangleq \mathbf{w}_t^{\mathsf{MAP}}\}$.

To find an ML2 estimate of basis vectors $\mathbf{\Phi}'_j$, we maximize Eq. (5.36) by taking the gradient of the terms related to $\mathbf{\Phi}'_j$ and setting it to zero, which leads to

$$\begin{aligned} \frac{\partial}{\partial \mathbf{\Phi}'_j} \Bigg[\sum_{t=1}^{T} \gamma_t(j) \Bigg[&- \log |(\mathbf{\Phi}'_j)^\mathsf{T} \mathbf{R}'_j \mathbf{\Phi}'_j + \mathcal{A}'_j| \\ &+ \mathbf{o}_t^\mathsf{T} \mathbf{R}'_j \mathbf{\Phi}'_j \left((\mathbf{\Phi}'_j)^\mathsf{T} \mathbf{R}'_j \mathbf{\Phi}'_j + \mathcal{A}'_j \right)^{-1} (\mathbf{\Phi}'_j)^\mathsf{T} \mathbf{R}'_j \mathbf{o}_t \Bigg] \Bigg] = 0, \end{aligned} \tag{5.39}$$

where the gradients of the two terms are derived as the $D \times N$ matrices given by

$$\begin{aligned} &\frac{\partial}{\partial \mathbf{\Phi}'_j} \log |(\mathbf{\Phi}'_j)^\mathsf{T} \mathbf{R}'_j \mathbf{\Phi}'_j + \mathcal{A}'_j| \\ &\qquad = 2 \mathbf{R}'_j \mathbf{\Phi}'_j \left((\mathbf{\Phi}'_j)^\mathsf{T} \mathbf{R}'_j \mathbf{\Phi}'_j + \mathcal{A}'_j \right)^{-1} = 2 \mathbf{R}'_j \mathbf{\Phi}'_j \mathbf{\Sigma}'_j, \end{aligned} \tag{5.40}$$

$$\begin{aligned} &\frac{\partial}{\partial \mathbf{\Phi}'_j} \left[\mathbf{o}_t^\mathsf{T} \mathbf{R}'_j \mathbf{\Phi}'_j \left((\mathbf{\Phi}'_j)^\mathsf{T} \mathbf{R}'_j \mathbf{\Phi}'_j + \mathcal{A}'_j \right)^{-1} (\mathbf{\Phi}'_j)^\mathsf{T} \mathbf{R}'_j \mathbf{o}_t \right] \\ &\qquad = \frac{\partial}{\partial \mathbf{\Phi}'_j} \mathrm{tr} \{ \mathbf{\Sigma}'_j (\mathbf{\Phi}'_j)^\mathsf{T} \mathbf{R}'_j \mathbf{o}_t \mathbf{o}_t^\mathsf{T} \mathbf{R}'_j \mathbf{\Phi}'_j \} \\ &\qquad = -2 \mathbf{R}'_j \mathbf{\Phi}'_j \mathbf{\Sigma}'_j (\mathbf{\Phi}'_j)^\mathsf{T} \mathbf{R}'_j \mathbf{o}_t \mathbf{o}_t^\mathsf{T} \mathbf{R}'_j \mathbf{\Phi}'_j \mathbf{\Sigma}'_j + 2 \mathbf{R}'_j \mathbf{o}_t \mathbf{o}_t^\mathsf{T} \mathbf{R}'_j \mathbf{\Phi}'_j \mathbf{\Sigma}'_j. \end{aligned} \tag{5.41}$$

The optimal solution $\mathbf{\Phi}_j^{\mathsf{ML2}}$, which is a $D \times N$ matrix, satisfies

$$\begin{aligned} &\sum_{t=1}^{T} \gamma_t(j) \left[-\mathbf{R}'_j \mathbf{\Phi}'_j \mathbf{\Sigma}'_j - \mathbf{R}'_j \mathbf{\Phi}'_j \mathbf{\Sigma}'_j (\mathbf{\Phi}'_j)^\mathsf{T} \mathbf{R}'_j \mathbf{o}_t \mathbf{o}_t^\mathsf{T} \mathbf{R}'_j \mathbf{\Phi}'_j \mathbf{\Sigma}'_j + \mathbf{R}'_j \mathbf{o}_t \mathbf{o}_t^\mathsf{T} \mathbf{R}'_j \mathbf{\Phi}'_j \mathbf{\Sigma}'_j \right] \\ &\qquad = \sum_{t=1}^{T} \gamma_t(j) \mathbf{R}'_j \left[-\mathbf{\Phi}'_j \mathbf{\Sigma}'_j - \mathbf{\Phi}'_j \mathbf{m}'_{tj} (\mathbf{m}'_{tj})^\mathsf{T} + \mathbf{o}_t (\mathbf{m}'_{tj})^\mathsf{T} \right] = 0. \end{aligned} \tag{5.42}$$

https://doi.org/10.1017/CBO9781107295360.006 Published online by Cambridge University Press

Similarly to the solution to $\mathcal{A}'_j$, the ML2 estimate can be expressed in an implicit form written by a function $\mathbf{F}^{\phi}(\mathbf{\Phi}'_j)$, namely

$$\begin{aligned}\mathbf{\Phi}_j^{\mathsf{ML2}} &= \left[\sum_{t=1}^{T}\gamma_t(j)\mathbf{o}_t(\mathbf{m}'_{tj})^{\intercal}\right]\left[\sum_{t=1}^{T}\gamma_t(j)(\mathbf{\Sigma}'_j+\mathbf{m}'_{tj}(\mathbf{m}'_{tj})^{\intercal})\right]^{-1}\\ &= \frac{\sum_{t=1}^{T}\gamma_t(j)\mathbf{o}_t(\mathbf{m}'_{tj})^{\intercal}}{\sum_{t=1}^{T}\gamma_t(j)}\cdot\mathcal{A}_j^{\mathsf{ML2}}\\ &\triangleq \mathbf{F}^{\mathsf{ML2}}(\mathbf{\Phi}'_j).\end{aligned} \tag{5.43}$$

This solution is viewed as a weighted operation in the outer product space of observations $\{\mathbf{o}_t\}$ and MAP sensing weights $\{\mathbf{m}'_{tj} \triangleq \mathbf{w}_t^{\mathsf{MAP}}\}$. The posterior probabilities $\{\gamma_t(j)\}$ and the ML2 hyperparameters $\mathcal{A}_j^{\mathsf{ML2}}$ serve as the weights and the rotation operator of the weighted average, respectively.

To find the ML2 estimate of precision matrix $\mathbf{R}'_j$, we maximize Eq. (5.36) with respect to $\mathbf{R}'_j$ and obtain

$$\begin{aligned}\sum_{t=1}^{T}\gamma_t(j)\Big[&(\mathbf{R}'_j)^{-1}-\mathbf{\Phi}'_j\mathbf{\Sigma}'_j(\mathbf{\Phi}'_j)^{\intercal}-\mathbf{o}_t\mathbf{o}_t^{\intercal}\\ &+\frac{\partial}{\partial\mathbf{R}'_j}\left(\mathbf{o}_t^{\intercal}\mathbf{R}'_j\mathbf{\Phi}'_j\mathbf{\Sigma}'_j(\mathbf{\Phi}'_j)^{\intercal}\mathbf{R}'_j\mathbf{o}_t\right)\Big]=0,\end{aligned} \tag{5.44}$$

where

$$\begin{aligned}\frac{\partial}{\partial\mathbf{R}'_j}\left(\mathbf{o}_t^{\intercal}\mathbf{R}'_j\mathbf{\Phi}'_j\mathbf{\Sigma}'_j(\mathbf{\Phi}'_j)^{\intercal}\mathbf{R}'_j\mathbf{o}_t\right) &= \frac{\partial}{\partial\mathbf{R}'_j}\mathrm{tr}\{\mathbf{\Sigma}'_j(\mathbf{\Phi}'_j)^{\intercal}\mathbf{R}'_j\mathbf{o}_t\mathbf{o}_t^{\intercal}\mathbf{R}'_j\mathbf{\Phi}'_j\}\\ &= -\mathbf{\Phi}'_j\mathbf{\Sigma}'_j(\mathbf{\Phi}'_j)^{\intercal}\mathbf{R}'_j\mathbf{o}_t\mathbf{o}_t^{\intercal}\mathbf{R}'_j\mathbf{\Phi}'_j\mathbf{\Sigma}'_j(\mathbf{\Phi}'_j)^{\intercal}\\ &\quad+\mathbf{\Phi}'_j\mathbf{\Sigma}'_j(\mathbf{\Phi}'_j)^{\intercal}\mathbf{R}'_j\mathbf{o}_t\mathbf{o}_t^{\intercal}+\mathbf{o}_t\mathbf{o}_t^{\intercal}\mathbf{R}'_j\mathbf{\Phi}'_j\mathbf{\Sigma}'_j(\mathbf{\Phi}'_j)^{\intercal}.\end{aligned} \tag{5.45}$$

We derive the ML2 solution as

$$\begin{aligned}(\mathbf{R}_j^{\mathsf{ML2}})^{-1} &= \frac{\sum_{t=1}^{T}\gamma_t(j)\begin{bmatrix}\mathbf{\Phi}'_j\mathbf{\Sigma}'_j(\mathbf{\Phi}'_j)^{\intercal}+\mathbf{o}_t\mathbf{o}_t^{\intercal}\\ +\mathbf{\Phi}'_j\mathbf{\Sigma}'_j(\mathbf{\Phi}'_j)^{\intercal}\mathbf{R}'_j\mathbf{o}_t\mathbf{o}_t^{\intercal}\mathbf{R}'_j\mathbf{\Phi}'_j\mathbf{\Sigma}'_j(\mathbf{\Phi}'_j)^{\intercal}\\ -\mathbf{\Phi}'_j\mathbf{\Sigma}'_j(\mathbf{\Phi}'_j)^{\intercal}\mathbf{R}'_j\mathbf{o}_t\mathbf{o}_t^{\intercal}-\mathbf{o}_t\mathbf{o}_t^{\intercal}\mathbf{R}'_j\mathbf{\Phi}'_j\mathbf{\Sigma}'_j(\mathbf{\Phi}'_j)^{\intercal}\end{bmatrix}}{\sum_{t=1}^{T}\gamma_t(j)}\\ &= \mathbf{\Phi}'_j\mathbf{\Sigma}'_j(\mathbf{\Phi}'_j)^{\intercal}+\frac{\sum_{t=1}^{T}\gamma_t(j)(\mathbf{o}_t-\mathbf{\Phi}'_j\mathbf{m}'_{tj})(\mathbf{o}_t-\mathbf{\Phi}'_j\mathbf{m}'_{tj})^{\intercal}}{\sum_{t=1}^{T}\gamma_t(j)}\\ &\triangleq \mathbf{F}^{\mathsf{ML2}}(\mathbf{R}'_j),\end{aligned} \tag{5.46}$$

which is also an implicit solution to $\mathbf{R}'_j$ since the right-hand-side (RHS) of Eq. (5.46) depends on $\mathbf{R}'_j$. Note that the RHS of Eq. (5.46) is symmetric positive definite. The first term is a scaled covariance matrix $\mathbf{\Sigma}'_j$ of the posterior distribution $p(\mathbf{w}_t|\mathbf{o}_t,\Theta',\Psi')$, which is doubly transformed by $\mathbf{\Phi}'_j$. The second term is interpreted as a covariance matrix weighted by posterior probabilities $\gamma_t(j)$, and is calculated using observations $\{\mathbf{o}_t\}$ and

https://doi.org/10.1017/CBO9781107295360.006 Published online by Cambridge University Press

"mean" vectors $\mathbf{\Phi}'_j\mathbf{m}'_{tj}$. This corresponds to performing *Bayesian sensing*, again using the MAP estimates $\{\mathbf{m}'_{tj} \triangleq \mathbf{w}_t^{\mathsf{MAP}}\}$.

We can see that type-2 ML (ML2) estimates of BS-HMM parameters and hyper-parameters are consistently formulated as the *implicit solutions*, which are beneficial for efficient implementation and good convergence in parameter estimation. Differently from conventional basis representation, where basis vectors and sensing weights are found separately, BS-HMMs provide a multivariate Bayesian approach to hybrid estimation of the compact basis vectors and the precision matrices of sensing weights under a consistent objective function. No training examples are stored for memory-based implementation.

To improve LVCSR performance based on BS-HMMs, there have been several extensions developed for acoustic modeling. As mentioned in Section 5.2.2, the mixture model of BS-HMMs can be extended by considering multiple sets of basis vectors per state. A mixture model of basis vectors is included for acoustic modeling. Using this mixture model, each observation $\mathbf{o}_t$ at state $s_t = j$ is expressed by

$$p(\mathbf{o}_t|\Theta_j) \triangleq \sum_{k=1}^{K} \omega_{jk}\mathcal{N}(\mathbf{o}_t|\mathbf{\Phi}_{jk}\mathbf{w}_t, \mathbf{R}_{jk}^{-1}), \tag{5.47}$$

where ω_{jk} is the mixture weight of jth component with the constraint

$$\sum_{k=1}^{K} \omega_{jk} = 1. \tag{5.48}$$

Here, the reconstruction error of an observation vector $\mathbf{o}_t$ due to the jth component with basis vectors $\mathbf{\Phi}_{jk} = [\boldsymbol{\phi}_{jk1}, \cdots, \boldsymbol{\phi}_{jkN}]$ is assumed to be Gaussian distributed with zero mean and precision matrix $\mathbf{R}_{jk}$. In addition, BS-HMMs can be constructed by incorporating a non-zero mean vector $\boldsymbol{\mu}_j^w$ in the prior density of sensing weights, i.e.,

$$p(\mathbf{w}_t|\mathbf{0}, \mathcal{A}_j) \rightarrow p(\mathbf{w}_t|\boldsymbol{\mu}_j^w, \mathcal{A}_j). \tag{5.49}$$

Similarly to the Maximum Likelihood Linear Regression (MLLR) adaptation for HMMs, BS-HMMs are developed for speaker adaptation where the nth BS-HMM basis vector is transformed by

$$\hat{\boldsymbol{\phi}}_{jn} = \mathbf{M}\widetilde{\boldsymbol{\phi}}_{jn}, \tag{5.50}$$

where $\mathbf{M}$ is a $D \times (D+1)$ regression matrix and $\widetilde{\boldsymbol{\phi}}_{jn} = [\boldsymbol{\phi}_{jn}^{\mathsf{T}}\ 1]^{\mathsf{T}}$ is the extended basis vector. The type-2 ML estimation can be applied to calculate the optimal solutions to non-zero mean vector $\boldsymbol{\mu}_j^w$ and regression matrix $\mathbf{M}$.

5.2.8 Discriminative training

Finally, BS-HMMs are sophisticated, incorporating both model-space and feature-space discriminative training, which is crucial to improve classification of confusing patterns in pattern recognition systems. Developing discriminative training for BS-HMMs is important for LVCSR. Instead of the goodness-of-fit criterion using marginal likelihood

https://doi.org/10.1017/CBO9781107295360.006 Published online by Cambridge University Press

function, the objective function for discriminative training is established according to the mutual information between observation data $\mathbf{O}$ and the sequence of reference words W^r (Bahl *et al*. 1986, Povey & Woodland 2002, Povey, Kanevsky, Kingsbury *et al*. 2008):

$$\begin{aligned}\mathcal{F}(\Theta) &\triangleq I_\Theta(\mathbf{O}, W^r) = \log \frac{p_\Theta(\mathbf{O}, W^r)}{p_\Theta(\mathbf{O})p(W^r)} \\ &= \log p_\Theta(\mathbf{O}|W^r) - \log \sum_W p_\Theta(\mathbf{O}|W)p(W) \\ &\triangleq \mathcal{F}^{\mathsf{num}}(\Theta) - \mathcal{F}^{\mathsf{den}}(\Theta),\end{aligned} \tag{5.51}$$

which consists of a numerator term $\mathcal{F}^{\mathsf{num}}(\Theta)$ and a denominator term $\mathcal{F}^{\mathsf{den}}(\Theta)$. The Maximum Mutual Information (MMI) estimation of BS-HMMs Θ^{MMI} is performed for discriminative training. To solve the optimization problem, we calculate the *weak-sense auxiliary function* (Povey & Woodland 2002, Povey 2003), where the HMM state sequence S is incorporated as follows:

$$\begin{aligned}Q(\Theta'|\Theta) &= Q^{\mathsf{num}}(\Theta'|\Theta) - Q^{\mathsf{den}}(\Theta'|\Theta) + Q^{\mathsf{sm}}(\Theta'|\Theta) \\ &= \sum_S p(S|\mathbf{O}, W^r, \Theta) \log p(S, \mathbf{O}|\Theta') \\ &\quad - \sum_S \sum_W p(S, W|\mathbf{O}, \Theta) \log p(S, \mathbf{O}|\Theta') \\ &\quad + Q^{\mathsf{sm}}(\Theta'|\Theta).\end{aligned} \tag{5.52}$$

The property of weak-sense auxiliary function turns out to meet the condition

$$\left.\frac{\partial Q(\Theta'|\Theta)}{\partial \Theta'}\right|_{\Theta'=\Theta} = \left.\frac{\partial \mathcal{F}(\Theta')}{\partial \Theta'}\right|_{\Theta'=\Theta}, \tag{5.53}$$

where the mode of MMI auxiliary function and its weak-sense auxiliary function have the same value. The smoothing function $Q^{\mathsf{sm}}(\Theta'|\Theta)$ in Eq. (5.52) is added to ensure that the objective function $Q(\Theta'|\Theta)$ is improved by this extended EM algorithm. For this, the smoothing function should satisfy

$$\left.\frac{\partial Q^{\mathsf{sm}}(\Theta'|\Theta)}{\partial \Theta'}\right|_{\Theta'=\Theta} = 0. \tag{5.54}$$

In what follows, we address the discriminative training of basis vectors $\mathbf{\Phi}_j$ of state j. The same procedure can be applied to estimate the discriminative precision matrix of reconstruction errors $\mathbf{R}_j$.

One possible choice of smoothing function meeting Eq. (5.54) is formed by the Kullback–Leibler divergence $\mathrm{KL}(\cdot\|\cdot)$ between marginal likelihoods of the current estimate $p(\mathbf{o}_t|\Theta)$ and the new estimate $p(\mathbf{o}_t|\Theta')$ given by

$$\begin{aligned}Q^{\mathsf{sm}}(\{\mathbf{\Phi}'_j\}|\{\mathbf{\Phi}_j\}) &\triangleq -\sum_{j=1}^{J} D_j \mathrm{KL}(p(\mathbf{o}|\mathbf{\Phi}_j)\|p(\mathbf{o}|\mathbf{\Phi}'_j)) \\ &\propto \sum_{j=1}^{J} D_j \int p(\mathbf{o}|\mathbf{\Phi}_j) \log p(\mathbf{o}|\mathbf{\Phi}'_j) d\mathbf{o}\end{aligned}$$

https://doi.org/10.1017/CBO9781107295360.006 Published online by Cambridge University Press

$$
\begin{aligned}
&\propto \sum_{j=1}^{J} D_j \Bigg[\log |\mathbf{R}'_j| - \int p(\mathbf{o}|\mathbf{\Phi}_j) \\
&\quad \times (\mathbf{o} - \mathbf{\Phi}_j \mathbf{w} + \mathbf{\Phi}_j \mathbf{w} - \mathbf{\Phi}'_j \mathbf{w})^{\mathsf{T}} \\
&\quad \times \mathbf{R}'_j (\mathbf{o} - \mathbf{\Phi}_j \mathbf{w} + \mathbf{\Phi}_j \mathbf{w} - \mathbf{\Phi}'_j \mathbf{w}) d\mathbf{o} \Bigg] \\
&= \sum_{j=1}^{J} D_j \Bigg[\log |\mathbf{R}'_j| - \int p(\mathbf{o}|\mathbf{\Phi}_j) \\
&\quad \times (\mathbf{o} - \mathbf{\Phi}_j \mathbf{w})^{\mathsf{T}} \mathbf{R}'_j (\mathbf{o} - \mathbf{\Phi}_j \mathbf{w}) d\mathbf{o} \\
&\quad - (\mathbf{\Phi}_j \mathbf{w} - \mathbf{\Phi}'_j \mathbf{w})^{\mathsf{T}} \mathbf{R}'_j (\mathbf{\Phi}_j \mathbf{w} - \mathbf{\Phi}'_j \mathbf{w}) \Bigg].
\end{aligned} \tag{5.55}
$$

Here, D_j is a state-dependent smoothing constant. Typically, the smoothing function is mathematically intractable when applying marginal likelihoods. Noting this, we can approximate the marginal likelihood by using an average plug-in MAP estimate:

$$
\mathbf{w} \approx \frac{1}{T} \sum_{t=1}^{T} \mathbf{w}_t^{\mathrm{MAP}}, \tag{5.56}
$$

obtained by taking an ensemble average of the MAP estimates in Eq. (5.35) using all observation frames $\{\mathbf{o}_1, \cdots, \mathbf{o}_T\}$. Ignoring the terms independent of $\mathbf{\Phi}_j$, the smoothing function is obtained by substituting the approximate MAP estimate of Eq. (5.56) into

$$
Q^{\mathsf{sm}}(\{\mathbf{\Phi}'_j\}|\{\mathbf{\Phi}_j\}) = -\sum_{j=1}^{J} D_j \mathbf{w}^{\mathsf{T}} (\mathbf{\Phi}_j - \mathbf{\Phi}'_j)^{\mathsf{T}} \mathbf{R}'_j (\mathbf{\Phi}_j - \mathbf{\Phi}'_j) \mathbf{w}. \tag{5.57}
$$

As a result, the weak-sense auxiliary function is expressed in terms of state occupation posteriors as follows:

$$
\begin{aligned}
Q(\{\mathbf{\Phi}'_j\}|\{\mathbf{\Phi}_j\}) &= \sum_{t=1}^{T} \sum_{j=1}^{J} (\gamma_t^{\mathsf{num}}(j) - \gamma_t^{\mathsf{den}}(j)) \\
&\quad \times \Big[\log |\mathcal{A}'_j| + \log |\mathbf{R}'_j| + \log |\mathbf{\Sigma}'_j| \\
&\qquad - \mathbf{o}_t^{\mathsf{T}} \mathbf{R}'_j \mathbf{o}_t + (\mathbf{m}'_{tj})^{\mathsf{T}} (\mathbf{\Sigma}'_j)^{-1} \mathbf{m}'_{tj} \Big] \\
&\quad - \sum_{j=1}^{J} D_j \mathbf{w}^{\mathsf{T}} (\mathbf{\Phi}_j - \mathbf{\Phi}'_j)^{\mathsf{T}} \mathbf{R}'_j (\mathbf{\Phi}_j - \mathbf{\Phi}'_j) \mathbf{w},
\end{aligned} \tag{5.58}
$$

where the state occupation posteriors of staying in state $s_t = j$ at time t in the numerator and denominator terms are calculated by

$$
\gamma_t^{\mathsf{num}}(j) \triangleq p(s_t = j|\mathbf{O}, W^r, \Theta), \tag{5.59}
$$

$$
\gamma_t^{\mathsf{den}}(j) \triangleq \sum_{W} p(s_t = j|\mathbf{O}, W, \Theta), \tag{5.60}
$$

https://doi.org/10.1017/CBO9781107295360.006 Published online by Cambridge University Press

given the reference word sequence W^r and all possible word sequences $\{W\}$, respectively. The current estimates $\{\mathbf{\Phi}_j\}$ are used in this calculation. To find an MMI estimate for basis parameters, we differentiate Eq. (5.58) with respect to $\mathbf{\Phi}'_j$ and set it to zero:

$$\begin{aligned}\frac{\partial}{\partial \mathbf{\Phi}'_j} Q(\mathbf{\Phi}'_j|\mathbf{\Phi}_j) \propto \mathbf{R}'_j \sum_{t=1}^{T} (\gamma_t^{\text{num}}(j) - \gamma_t^{\text{den}}(j)) \\ \times (-\mathbf{\Phi}'_j \mathbf{\Sigma}'_j - \mathbf{\Phi}'_j \mathbf{m}'_{tj} (\mathbf{m}'_{tj})^{\mathsf{T}} + \mathbf{o}_t (\mathbf{m}'_{tj})^{\mathsf{T}}) \\ - \mathbf{R}'_j D_j (\mathbf{\Phi}'_j - \mathbf{\Phi}_j) \mathbf{w}\mathbf{w}^{\mathsf{T}} = 0, \end{aligned} \tag{5.61}$$

which is derived by considering the definition of variables $(\mathbf{\Sigma}'_j)^{-1}$ and $\mathbf{m}'_j$ in Eq. (5.29) and Eq. (5.30), respectively. Again, the implicit solution to an MMI estimate of $\mathbf{\Phi}'_j$ should satisfy

$$\begin{aligned}\mathbf{\Phi}_j^{\text{MMI2}} = \left[\sum_{t=1}^{T} (\gamma_t^{\text{num}}(j) - \gamma_t^{\text{den}}(j)) \mathbf{o}_t (\mathbf{m}'_{tj})^{\mathsf{T}} + D_j \mathbf{\Phi}_j \mathbf{w}\mathbf{w}^{\mathsf{T}} \right] \\ \times \left[\sum_{t=1}^{T} (\gamma_t^{\text{num}}(j) - \gamma_t^{\text{den}}(j)) \right. \\ \left. \times (\mathbf{\Sigma}'_j + \mathbf{m}'_{tj} (\mathbf{m}'_{tj})^{\mathsf{T}}) + D_j \mathbf{w}\mathbf{w}^{\mathsf{T}} \right]^{-1} \\ \triangleq \mathbf{F}^{\text{MMI2}}(\mathbf{\Phi}'_j). \end{aligned} \tag{5.62}$$

This solution is expressed as a recursive function $\mathbf{F}^{\text{MMI2}}$ of new basis parameter $\mathbf{\Phi}'_j$. Strictly speaking, an MMI estimation based on marginal likelihood is seen as a type 2 MMI (MMI2) estimation, which is different from conventional MMI training (Bahl *et al*. 1986, Povey & Woodland 2002, Povey *et al*. 2008) based on likelihood function without marginalization. In Eq. (5.62), the second terms in the numerator and the denominator come from smoothing function $Q^{\text{sm}}(\mathbf{\Phi}'_j|\mathbf{\Phi}_j)$ and serve to prevent instability in the MMI2 optimization procedure. The solution is highly affected by the difference between statistics for the reference hypothesis and statistics for the competing hypotheses $\gamma_t^{\text{num}}(j) - \gamma_t^{\text{den}}(j)$.

It is clear that *discriminative training* and *Bayesian learning* are simultaneously performed in a type 2 MMI estimation. By doing this, the performance of LVCSR can be significantly improved (Saon & Chien 2012*b*). The robustness to uncertainty of sensing weights in a basis representation can be assured. Some experimental results are described below.

5.2.9 System performance

Evaluation of BS-HMMs was performed by using the LVCSR task in the domain of Arabic broadcast news transcription which was part of the DARPA GALE program. In total, 1800 hours of manually transcribed Arabic broadcast news and conversations were used in this evaluation (Saon & Chien 2011). The results on several test sets were

https://doi.org/10.1017/CBO9781107295360.006 Published online by Cambridge University Press

Table 5.1 Comparison of the number of free parameters and word error rates for baseline acoustic models after ML training and BS-HMMs after ML2 training.

System	Nb. parameters	WER		
		DEV'07	DEV'08	DEV'09
Baseline 800K	64.8M	13.8%	16.4%	19.6%
Baseline 2.8M	226.8M	14.1%	16.2%	19.3%
BS-HMM 417K	148.5M	13.6%	16.0%	18.9%

reported: DEV'07 (2.5 hours), DEV'08 (3 hours) and DEV'09 (3 hours). The front-end processing was performed as mentioned in Saon & Chien (2012*b*). The vocal-tract-length-warped PLP (Hermansky 1990) cepstrum features were extracted with a context window of nine frames. The features were mean and variance normalized on a per speaker basis. Linear discriminant analysis was used to reduce the feature dimension to 40. The maximum likelihood training of the acoustic model was interleaved with the estimation of a global semi-tied covariance transform (Gales 1998). All models in this evaluation were estimated based on pentaphones and speaker adaptively trained with feature-space MLLR (fMLLR). Each pentaphone was modeled by a 3-state left-to-right HMM without state skipping. At test time, speaker adaptation was performed with vocal-tract-length normalization (Wegmann, McAllaster, Orloff *et al*. 1996), fMLLR and multiple regression MLLR transforms. The vocabulary contained 795K words. The decoding was done with 4-gram language models which were estimated with modified Kneser–Ney smoothing. The acoustic models were discriminatively trained in both feature space and model space according to the boosted MMI criterion (Povey, Kingsbury, Mangu *et al*. 2005). The baseline acoustic models had 5000 context-dependent HMM states and 800K 40-dimensional diagonal covariance Gaussians.

In the implementation, BS-HMM parameters were initialized by training a large HMM model with 2.8M diagonal covariance Gaussians by maximum likelihood method. The means of GMM were clustered and then treated as the initial basis $\mathbf{\Phi}_{jk}$ for state j and mixture component k. The resulting number of mixture components in BS-HMMs after the clustering step was 417K. The precision matrices $\mathbf{R}_{jk}$ and $\mathcal{A}_{jk}$ were assumed to be diagonal and were initialized to the identity matrix. Table 5.1 compares the performance of the baseline 800K Gaussians model and the 2.8M Gaussians model used to train baseline acoustic models after ML training and to seed BS-HMMs after ML2 training. The number of free parameters is included in this comparison. As we can see, BS-HMMs outperform both baseline systems in terms of word error rates (%). The possible reasons are twofold. The first one is that the covariance modeling in BS-HMMs is more accurate than that in HMMs. The second one is due to the Bayesian parameter updates which provide an effective smoothing.

In contrast, we conducted a model comparison for BS-HMMs according to the estimated hyperparameters of sensing weights. Model compression was performed by discarding 50% of the basis vectors Φ_{jkn} corresponding to the largest hyperparameters α_{jkn}. This results in a compressed model with approximately 91M free parameters (about

https://doi.org/10.1017/CBO9781107295360.006 Published online by Cambridge University Press

Table 5.2 Comparison of word error rates for original and compressed BS-HMMs before and after model-space discriminative training with MMI2 training.

Model	Training	DEV07	DEV08	DEV09
Original	ML2	12.0%	13.9%	17.4%
Compressed	ML2	12.4%	14.2%	17.6%
Original	MMI2	10.7%	11.9%	15.0%
Compressed	MMI2	10.4%	11.7%	14.8%

30% larger than the 800K baseline HMM). For both original and compressed models, we performed model-space discriminative training using the MMI2 criterion. Table 5.2 reports the recognition performance before and after model-space discriminative training. We find that the compressed models outperform the originals after discriminative training even though they start from a higher word error rate after ML2 estimation. However, discriminative training is significantly more expensive than ML estimation which makes it difficult to find the optimal model size.

5.3 Hierarchical Dirichlet language model

In what follows, we revisit the interpolation smoothing methods presented in Section 3.6 and used to deal with the small sample size problem in ML estimation of n-gram parameters $\Theta^{\mathsf{ML}} = \{p_{\mathsf{ML}}(w_i|w_{i-n+1}^{i-1})\}$. Different from the heuristic solutions to language model smoothing in Section 3.6, a *full Bayesian* language model is proposed to realize interpolation smoothing for n-grams. The theory of evidence approximation is developed and applied to construct the hierarchical Dirichlet language model (MacKay & Peto 1995, Kawabata & Tamoto 1996).

5.3.1 n-gram smoothing revisited

In general, the frequency estimator in a higher-order language model has large variance, because there are so many possible word combinations in an n-gram event $\{w_{i-n+1}^{i-1}, w_i\}$ that only a small fraction of them have been observed in the data. A simple linear interpolation scheme for an n-gram language model is performed by interpolating an ML model of an n-gram $p_{\mathsf{ML}}(w_i|w_{i-n+1}^{i-1})$ with that of an $(n-1)$-gram $p_{\mathsf{ML}}(w_i|w_{i-n+2}^{i-1})$ using

$$\hat{p}(w_i|w_{i-n+1}^{i-1}) = \lambda p_{\mathsf{ML}}(w_i|w_{i-n+1}^{i-1}) + (1-\lambda)p_{\mathsf{ML}}(w_i|w_{i-n+2}^{i-1}), \tag{5.63}$$

where λ denotes the interpolation weight, which can be determined empirically from validation data. Basically, it is not possible to make language models without making a-priori assumptions. The smoothed n-gram $\hat{p}(w_i|w_{i-n+1}^{i-1})$ in Eq. (5.63) can be seen as an integrated n-gram model which is determined from a hierarchical model based on the smoothing parameters a-posteriori from the data. In what follows, we would like to reverse-engineer the underlying model, which gives a probabilistic meaning to language

https://doi.org/10.1017/CBO9781107295360.006 Published online by Cambridge University Press

model smoothing. We will explain the construction of a hierarchical prior model and illustrate how this model is used to justify the smoothed language model from a Bayesian perspective. A type-2 ML estimation is conducted to find optimal hyperparameters from training data. No cross validation procedure is required in a Bayesian language model.

5.3.2 Dirichlet prior and posterior

n-gram model parameters $\Theta = \{p(w_i|w_{i-n+1}^{i-1})\}$ are known as multinomial parameters which are used to predict n-gram events of word w_i appearing after observing history words w_{i-n+1}^{i-1}. The ML estimation of an n-gram $\theta_{w_i|w_{i-n+1}^{i-1}} = p(w_i|w_{i-n+1}^{i-1})$ from a text corpus $\mathcal{D}$ has been shown in Eq. (3.192). Here, we are interested in a Bayesian language model where the prior density of multinomial parameters is introduced. A parameter vector consists of N multinomial parameters:

$$\boldsymbol{\theta} = \text{vec}(\Theta) = [\theta_1, \cdots, \theta_N]^\intercal \text{ subject to} \tag{5.64}$$
$$0 \le \theta_i \le 1 \text{ and } \sum_{i=1}^{N} \theta_i = 1.$$

The *conjugate prior* over multinomial parameters $\boldsymbol{\theta}$ is specified by a Dirichlet prior with hyperparameters $\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_N]^\intercal$, which is a multivariate distribution in the form of

$$p(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \text{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{1}{Z(\boldsymbol{\alpha})} \prod_{i=1}^{N} \theta_i^{\alpha_i - 1}. \tag{5.65}$$

We express the normalization constant of the Dirichlet distribution by

$$Z(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{N} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{N} \alpha_i)}. \tag{5.66}$$

The mean vector of Dirichlet distribution is given by

$$\int \boldsymbol{\theta} \text{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha}) d\boldsymbol{\theta} = \frac{\boldsymbol{\alpha}}{\sum_{i=1}^{N} \alpha_i}. \tag{5.67}$$

When we observe the training samples $\mathcal{D}$, the posterior distribution is derived as another Dirichlet distribution:

$$\begin{aligned} p(\boldsymbol{\theta}|\mathcal{D}, \boldsymbol{\alpha}) &= \frac{p(\mathcal{D}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{\alpha})}{p(\mathcal{D}|\boldsymbol{\alpha})} \\ &= \frac{\prod_{i=1}^{N} \theta_i^{c(\theta_i)} \prod_{i=1}^{N} \theta_i^{\alpha_i - 1}}{p(\mathcal{D}|\boldsymbol{\alpha}) Z(\boldsymbol{\alpha})} \\ &= \frac{\prod_{i=1}^{N} \theta_i^{c(\theta_i) + \alpha_i - 1}}{Z(\mathbf{c} + \boldsymbol{\alpha})} \\ &= \text{Dir}(\boldsymbol{\theta}|\mathbf{c} + \boldsymbol{\alpha}), \end{aligned} \tag{5.68}$$

with the updated hyperparameters $\mathbf{c} + \boldsymbol{\alpha}$. In Eq. (5.68), each entry $c(\theta_i)$ of

$$\mathbf{c} = [c(\theta_1), \cdots, c(\theta_N)]^\intercal \tag{5.69}$$

https://doi.org/10.1017/CBO9781107295360.006 Published online by Cambridge University Press

denotes the number of occurrences of the ith n-gram event in θ_i in training data $\mathcal{D}$. This shows the property of a conjugate prior by using the Dirichlet distribution.

5.3.3 Evidence function

To obtain the predictive probability of a word w_i given history words w_{i-n+1}^{i-1} and training data $\mathcal{D}$, we apply the sum rule to calculate the evidence function or the marginal likelihood:

$$\begin{aligned} p(w_i|w_{i-n+1}^{i-1},\mathcal{D},\boldsymbol{\alpha}) &= \int p(w_i|w_{i-n+1}^{i-1},\mathcal{D},\boldsymbol{\alpha})p(\boldsymbol{\theta}|\mathcal{D},\boldsymbol{\alpha})d\boldsymbol{\theta} \\ &= \int \theta_{w_i|w_{i-n+1}^{i-1}} \mathrm{Dir}(\boldsymbol{\theta}|\mathbf{c}+\boldsymbol{\alpha})d\boldsymbol{\theta} \\ &= \frac{c(\theta_{i|j})+\alpha_i}{\sum_{k=1}^{N}[c(\theta_{k|j})+\alpha_k]}, \end{aligned} \tag{5.70}$$

which is marginalized over all values of parameter $\boldsymbol{\theta}$. This predictive distribution is equivalent to calculating the *mean* of an n-gram parameter:

$$p(w_i|w_{i-n+1}^{i-1}) \triangleq \theta_{i|j}, \tag{5.71}$$

based on the *posterior distribution* $p(\boldsymbol{\theta}|\mathcal{D},\boldsymbol{\alpha})$, which is a Dirichlet distribution with hyperparameters $\mathbf{c}+\boldsymbol{\alpha}$. Here, the n-gram probability $p(w_i|w_{i-n+1}^{i-1})$ is simply expressed by $\theta_{i|j}$ where j denotes the back-off smoothing information from the lower-order model $p(w_i|w_{i-n+1}^{i-2})$, which is addressed in Section 5.3.4.

We may further conduct the next level of inference by inferring the hyperparameters given the data. The posterior distribution of $\boldsymbol{\alpha}$ is expressed by

$$p(\boldsymbol{\alpha}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\alpha})p(\boldsymbol{\alpha})}{p(\mathcal{D})}. \tag{5.72}$$

The hierarchical prior/posterior model of parameters $\boldsymbol{\theta}$ and hyperparameters $\boldsymbol{\alpha}$ is constructed accordingly. The marginal likelihood over hyperparameters $\boldsymbol{\alpha}$ is yielded by

$$p(w_i|w_{i-n+1}^{i-1},\mathcal{D}) = \int p(w_i|w_{i-n+1}^{i-1},\mathcal{D},\boldsymbol{\alpha})p(\boldsymbol{\alpha}|\mathcal{D})d\boldsymbol{\alpha}. \tag{5.73}$$

We may find the most probable MAP estimate $\boldsymbol{\alpha}^{\mathsf{MAP}}$ by

$$\boldsymbol{\alpha}^{\mathsf{MAP}} = \arg\max_{\boldsymbol{\alpha}} \; p(\boldsymbol{\alpha}|\mathcal{D}). \tag{5.74}$$

Then the marginal distribution is approximated as

$$p(w_i|w_{i-n+1}^{i-1},\mathcal{D}) \approx p(w_i|w_{i-n+1}^{i-1},\mathcal{D},\boldsymbol{\alpha}^{\mathsf{MAP}}). \tag{5.75}$$

In addition, we would like to calculate the optimal hyperparameters via ML2 estimation from training data $\mathcal{D}$ based on the evidence framework. To do so, we need to determine the *evidence function* given hyperparameters $p(\mathcal{D}|\boldsymbol{\alpha})$. By referring to Eq. (5.68), this function is derived as

$$p(\mathcal{D}|\boldsymbol{\alpha}) = \frac{Z(\mathbf{c}+\boldsymbol{\alpha})}{Z(\boldsymbol{\alpha})} = \prod_j \left(\frac{\prod_{i=1}^{N}\Gamma(c(\theta_{i|j})+\alpha_i)}{\Gamma(\sum_{i=1}^{N}c(\theta_{i|j})+\alpha_i)} \cdot \frac{\Gamma(\sum_{i=1}^{N}\alpha_i)}{\prod_{i=1}^{N}\Gamma(\alpha_i)} \right), \tag{5.76}$$

https://doi.org/10.1017/CBO9781107295360.006 Published online by Cambridge University Press

which is viewed as a ratio of normalization constants of posterior probability $p(\boldsymbol{\theta}|\mathcal{D},\boldsymbol{\alpha})$ over prior probability $p(\boldsymbol{\theta}|\boldsymbol{\alpha})$.

5.3.4 Bayesian smoothed language model

It is important to illustrate the physical meaning of predictive distribution in Eq. (5.70). The hyperparameter α_i appears as an *effective initial count* for an n-gram event w_{i-n+1}^{i}. This marginal likelihood is integrated by the information sources from prior statistics $\boldsymbol{\alpha}$ as well as training data $\mathcal{D}$ or their counts of occurrences $\mathbf{c}$. On the other hand, the predictive distribution in Eq. (5.70) can be rewritten as

$$\begin{aligned} p(w_i|w_{i-n+1}^{i-1},\mathcal{D},\boldsymbol{\alpha}) &= \frac{c(w_{i-n+1}^{i})+\alpha_{w_i|w_{i-n+2}^{i-1}}}{\sum_{w_i}\left[c(w_{i-n+1}^{i})+\alpha_{w_i|w_{i-n+2}^{i-1}}\right]} \\ &= \lambda_{w_{i-n+1}^{i-1}} p_{\mathsf{ML}}(w_i|w_{i-n+1}^{i-1}) + (1-\lambda_{w_{i-n+1}^{i-1}})\frac{\alpha_{w_i|w_{i-n+2}^{i-1}}}{\sum_{w_i}\alpha_{w_i|w_{i-n+2}^{i-1}}}, \end{aligned} \tag{5.77}$$

where $p_{\mathsf{ML}}(w_i|w_{i-n+1}^{i-1})$ denotes the ML model introduced in Eq. (3.192) and $1-\lambda_{w_{i-n+1}^{i-1}}$ implies the interpolation weight for prior statistics and is herein obtained as

$$1-\lambda_{w_{i-n+1}^{i-1}} = \frac{\sum_{w_i}\alpha_{w_i|w_{i-n+2}^{i-1}}}{\sum_{w_i}\left[c(w_{i-n+1}^{i})+\alpha_{w_i|w_{i-n+2}^{i-1}}\right]}. \tag{5.78}$$

It is interesting to see that the predictive distribution in Eq. (5.77) is interpreted as the smoothed n-gram based on the *interpolation smoothing*. We build the tight connection between the Bayesian language model and a linearly smoothed language model as addressed in Section 3.6. The prior statistics or hyperparameters $\alpha_{w_i|w_{i-n+2}^{i-1}}$ should sufficiently reflect the backoff information from the low-order model $p(w_i|w_{i-n+2}^{i-1})$ when calculating the predictive n-gram probability $p(w_i|w_{i-n+1}^{i-1},\mathcal{D},\boldsymbol{\alpha})$. Comparing Eq. (5.78) and Eq. (3.211), the Bayesian language model is shown to be equivalent to the Witten–Bell smoothed language model in the case that the hyperparameters $\alpha_{w_i|w_{i-n+2}^{i-1}}$ are selected to meet the condition

$$\sum_{w_i}\alpha_{w_i|w_{i-n+2}^{i-1}} = N_{1+}(w_{i-n+1}^{i-1},\bullet). \tag{5.79}$$

(As before, the $\bullet$ represents any possible words at i that are summed over.)

Nevertheless, the advantage of the *Bayesian smoothed language model* is to automatically determine the optimal hyperparameters $\alpha_{w_i|w_{i-n+2}^{i-1}}$ from training data $\mathcal{D}$.

5.3.5 Optimal hyperparameters

According to the evidence framework, a type-2 ML estimation is carried out to find optimal hyperparameters $\boldsymbol{\alpha}=[\alpha_1,\cdots,\alpha_N]^{\intercal}$ by maximizing the evidence function

$$\boldsymbol{\alpha}^{\mathsf{ML2}} = \arg\max_{\boldsymbol{\alpha}}\ p(\mathcal{D}|\boldsymbol{\alpha}). \tag{5.80}$$

https://doi.org/10.1017/CBO9781107295360.006 Published online by Cambridge University Press

More specifically, we find individual parameter α_i^{ML2} through calculating the differentiation

$$\frac{\partial}{\partial \alpha_i} \log p(\mathcal{D}|\boldsymbol{\alpha}) = \sum_j \Bigg[\Psi(c(\theta_{i|j}) + \alpha_i) - \Psi\left(\sum_{i=1}^{N} c(\theta_{i|j}) + \alpha_i\right) + \Psi\left(\sum_{i=1}^{N} \alpha_i\right) - \Psi(\alpha_i) \Bigg], \tag{5.81}$$

where the di-gamma function

$$\Psi(x) \triangleq \frac{\partial}{\partial x} \log \Gamma(x) \tag{5.82}$$

is incorporated. We may use the conjugate gradient algorithm to find α_i^{ML2} or apply some approximation to derive an explicit optimization algorithm.

In general, it is reasonable that $\sum_{i=1}^{N} \alpha_i > 1$ and $\alpha_i < 1$. We can use the recursive formula of the di-gamma function (MacKay & Peto 1995),

$$\Psi(x+1) = \Psi(x) + \frac{1}{x}, \tag{5.83}$$

to combine the first and fourth terms in the brackets of Eq. (5.81) to obtain

$$\Psi(c(\theta_{i|j}) + \alpha_i) - \Psi(\alpha_i) = \frac{1}{c(\theta_{i|j}) - 1 + \alpha_i} + \frac{1}{c(\theta_{i|j}) - 2 + \alpha_i} + \cdots + \frac{1}{2 + \alpha_i} + \frac{1}{1 + \alpha_i} + \frac{1}{\alpha_i}. \tag{5.84}$$

The number of terms in the right-hand-side of Eq. (5.84) is $c(\theta_{i|j})$. Assuming α_i is smaller than 1, we can approximate Eq. (5.84), for $c(\theta_{i|j}) \geq 1$, by

$$\begin{aligned} \Psi(c(\theta_{i|j}) + \alpha_i) - \Psi(\alpha_i) &= \frac{1}{\alpha_i} + \sum_{c=2}^{c(\theta_{i|j})} \left[\frac{1}{c - 1 + \alpha_i} \right] \\ &\approx \frac{1}{\alpha_i} + \sum_{c=2}^{c(\theta_{i|j})} \left[\frac{1}{c-1} - \frac{\alpha_i}{(c-1)^2} + O(\alpha_i^2) \right] \\ &= \frac{1}{\alpha_i} + \sum_{c=2}^{c(\theta_{i|j})} \frac{1}{c-1} - \alpha_i \sum_{c=2}^{c(\theta_{i|j})} \frac{1}{(c-1)^2} + O(\alpha_i^2). \end{aligned} \tag{5.85}$$

Here, the function inside the brackets,

$$f(\alpha_i) = \frac{1}{c - 1 + \alpha_i}, \tag{5.86}$$

is approximated by a Taylor series at the point $\alpha_i = 0$. Further, we apply the approximation of a di-gamma function,

$$\Psi(x) \approx \log(x) - \frac{1}{2x} + O\left(\frac{1}{x^2}\right), \tag{5.87}$$

https://doi.org/10.1017/CBO9781107295360.006 Published online by Cambridge University Press

to approximate the second and third terms in Eq. (5.81) (MacKay & Peto 1995):

$$\begin{aligned} K(\boldsymbol{\alpha}) &= \sum_j \left\{ \Psi\left(\sum_{i=1}^N \alpha_i\right) - \Psi\left(\sum_{i=1}^N c(\theta_{i|j}) + \alpha_i\right) \right\} \\ &\approx \sum_j \log\left[\frac{\sum_{i=1}^N c(\theta_{i|j}) + \alpha_i}{\sum_{i=1}^N \alpha_i}\right] \\ &\quad + \frac{1}{2}\sum_j \left[\frac{\sum_{i=1}^N c(\theta_{i|j})}{\left(\sum_{i=1}^N \alpha_j\right)\left(\sum_{i=1}^N c(\theta_{i|j}) + \alpha_i\right)}\right]. \end{aligned} \tag{5.88}$$

For each count c and word i, let N_{ci} be the number of back-off contexts j such that $c(\theta_{i|j}) \geq c$, and let $c_i^{\max}$ denote the largest c such that $N_{ci} > 0$. Denote the number of entries in row i of $c(\theta_{i|j})$ that are non-zero N_{1i} by V_i. We compute the quantities:

$$G_i = \sum_{c=2}^{c_i^{\max}} \frac{N_{ci}}{c-1}, \tag{5.89}$$

$$H_i = \sum_{c=2}^{c_i^{\max}} \frac{N_{ci}}{(c-1)^2}. \tag{5.90}$$

Finally, the solution to optimal hyperparameters $\boldsymbol{\alpha}^{\mathsf{ML2}}$ is obtained. The hyperparameter α_i^{ML2} corresponding to each word i should satisfy

$$\begin{aligned} \alpha_i^{\mathsf{ML2}} &= \frac{2V_i}{K(\boldsymbol{\alpha}) - G_i + \sqrt{(K(\boldsymbol{\alpha}) - G_i)^2 + 4H_iV_i}} \\ &\triangleq f_i(\boldsymbol{\alpha}). \end{aligned} \tag{5.91}$$

Again, this solution to ML2 estimation is expressed as an *implicit solution* because the right-hand-side of Eq. (5.91) is a function of hyperparameters $\boldsymbol{\alpha}$. Starting from the initial hyperparameters $\boldsymbol{\alpha}$ and the resulting function $K(\boldsymbol{\alpha})$, the individual hyperparameter is then updated to $\alpha_j^{(1)}$. The function $K(\boldsymbol{\alpha}^{(1)})$ is updated as well. The estimation procedure $\boldsymbol{\alpha} \rightarrow \boldsymbol{\alpha}^{(1)} \rightarrow \cdots$ converges very rapidly.

https://doi.org/10.1017/CBO9781107295360.006 Published online by Cambridge University Press