Variational Bayesian Approach for Stochastic Grammatical Inference

Department of Computational Intelligence and Systems Science
Interdisciplinary Graduate School of Science and Engineering
Tokyo Institute of Technology

Tikara Hosino

2007

Abstract

Stochastic grammar has been widely used in real world application such as speech recognition, natural language processing and bioinformatics. In many field, it was shown that the stochastic approach have being the competitor of the ordinary rule based approach. In spite of the successful application, the theoretical property, for example, the generalization error or the model selection have not been clarified.

The difficult part of the analysis is caused by the non-identifiability and the singularities in the stochastic grammar. Recently, in Bayesian Learning, asymptotic behavior of the free energy and the generalization error was clarified based on the algebraic geometry. The result show that Bayesian learning is more effective than the maximum likelihood method by the effect of the entropy of the singularities. However, in general, the exact performing of Bayesian learning is prohibited because of the multiple integral operation for the posterior distribution. Therefore, we need some approximation scheme. As the efficient approximation method, the variational Bayesian method was proposed. The computational costs of the variational Bayesian method is same order of the maximum likelihood method. In the numerical experiment and the real world application, it was reported that the variational Bayesian method had a better generalization performance. However, the theoretical aspects of the variational Bayesian methods have not been well clarified.

In this thesis, we derive the asymptotic free energy of variational Bayesian stochastic grammar. The variational free energy can be used to evaluate the approximation accuracy and can be used to compare the several optimization method. The result show that in variational Bayesian learning, the conditions of learning machines satisfies that the redundant components is effectively eliminated. This cause the suppression of the overfitting. Finally, we propose a novel approach for the model selection on the stochastic grammar and evaluate the proposed method by the the numerical experiments.

Preface

This work has been carried out at Watanabe laboratory, Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology. I wish to thank my supervisor Prof. Sumio Watanabe for his support and a lot of encouraging comments during my Ph.D. studies. I would also like to express my great respect and gratitude to him for giving me the opportunity to meet such fascinating studies.

I grate thank Nihon Unisys, Ltd. for giving the opportunity to apply the Ph.D. course at Watanabe laboratory and the financial support during the period. I also would like to express my gratitude to the member of HR Strategic Innovation & Training, especially to Motonori Saeki, Representative Director and Senior Corporate Officer, and Kumiko Shirai, General Manager of the HR Strategic Innovation & Training, for providing a constant support to this work.

I am grateful to Prof. Yoshiyuki Kabashima, Prof. Katumi Nitta , Prof. Misako Takayasu, and Prof. Manabu Okumura for reviewing this thesis and providing comments to improve it.

I also wish to thank present and former members of Watanabe laboratory: Dr. Keisuke Yamazaki, Dr. Shinichi Nakajima, Dr. Miki Aoyagi, Kazumi Imbe, Dr. Kazuho Watanabe, Dr. Motoki Shiga, Nobuhiro Nakano, Satoshi Tsuji, Taruhi Iwagaki, Michiaki Hamada, Takeshi Matsuda, Kaori Fujiwara, Kenji Nagata, Shingo Takamatsu, Yu Nishiyama, and Ryosuke Iriguchi.

Tikara Hosino

Contents

Chapter 1

Introduction

Today, very large data set is collected via such as World Wide Web or Point Sales system. This direction is accelerated by sensor networks or Radio Frequency Identification which are called ubiquitous computing. Moreover, Bioinformatics which deals with RNA or DNA sequence analysis emerges as a new field[4, 9]. In utilizing such huge data, for the uncertainty of the data or the difficulty of the description of the concrete rules, the learning from example is widely used and shows their effectiveness in many fields.

In this thesis, we focus on a stochastic grammatical inference which is inferring a grammar from the given symbols. Concretely, we treat two classes of stochastic grammar. The first is Hidden Markov Model (HMM) which is stochastic version of the regular grammar [15]. Hidden Markov model is established as the standard method for speech recognition[27]. Recently, HMM extends the application area to robot navigation, clustering of the time series and DNA sequence analysis. The second is Stochastic Context Free Grammar (SCFG). SCFG is used in natural language parser, RNA secondary structure analysis and Knowledge Discovery where the problem has the nested or the tree structure.

In spite of the wide range of successful applications, the theoretical property of the stochastic grammar is not well established. The difficult part of the analysis is caused by the non-identifiability of the models. In statistical

models, if the map from the parameter to the model is one-to-one, the model is defined as identifiable, or if otherwise the model is non-identifiable. Many statistical models that have hierarchical structure, such as neural networks and stochastic grammar is non-identifiable. In non-identifiable models, the Fisher information matrix is degenerate and the model has many singularities in the parameter space. Therefore, the conventional method which assumes the positive definiteness of the Fisher information matrix cannot be applied directory to the non-identifiable models.

Recently, many approach are proposed for investigating the theoretical property of the non-identifiable model. In particular, in Bayesian learning, asymptotic property of the free energy and the generalization error is derived by the method of the algebraic analysis[33]. The result shows that in Bayesian learning, the free energy and the generalization error are much smaller than identifiable models. On the contrary in the maximum likelihood method, the generalization behavior is worse than identifiable models because of the overfitting. This result theoretically shows that Bayesian learning is effective in the non-identifiable models.

Although the theoretical advantage of Bayesian learning, a direct implementation of Bayesian learning is difficult. In Bayesian Learning, all the inference problem are executing by the posterior distribution. For example, the prediction of the new data is done by the expectation to the model over the posterior distribution and this higher dimensional integration cannot be analytically executed and the numerical integration is highly intensive in non-identifiable models. Therefore we need some approximation method.

There are two lines of approximation schema. The first is sampling method. The most popular technique is Markov Chain Monte Carlo (MCMC) method. MCMC creates Markov Chain that converges to the posterior distribution[13]. In spite of the generality and ensure the exactness in the limit, the MCMC methods are computationally intensive and have difficulty to determine the convergence to the distribution. The second is the determin-

istic approach. In this line, Variational Bayesian approach was proposed[3, 6].

Variational Bayesian learning approximates the posterior distribution directly via minimizing Kullback information to the true posterior. Computational efficiency for the exponential family with hidden variables and the good generalization performance for the real world application were reported. Recently, for some statistical models, the theoretical property of variational Bayesian learning was revealed[32, 25].

In this thesis, we consider the theoretical aspects of variational Bayesian Stochastic grammar. Main contribution of the thesis is deriving the asymptotic property of the variational free energy in HMM and SCFG. The variational free energy is the objective function on the optimization method, has the information of the approximation accuracy and is used for the model selection problem.

Additionally, based on the analysis and the non-identifiable model selection method[34], we propose a new approach to the model selection for HMM and SCFG. The numerical experiment is executed for evaluating the proposed method by the synthetic data. In HMM and SCFG, the conventional method such as AIC[1], BIC[30], or MDL[28] cannot be theoretically justified as explained above, the proposed approach has a theoretical foundation for the non-identifiable problems.

This thesis is organized as follows.

- Chapter 2 explains the standard procedure of the statistical learning.
- Chapter 3 explains the variational Bayesian learning.
- Chapter 4 introduces the basics of the stochastic grammar and their variational Bayesian learning.
- Chapter 5 derives the asymptotic variational free energy in HMM and SCFG.

- Chapter 6 proposes a new model selection procedure based on the analysis of the previous chapter.
- Chapter 7 evaluates the proposed method by the numerical experiments.
- Chapter 8 gives the discussions of the result and the suggestion for the future works.
- Chapter 9 summarizes the results and concludes the thesis.

Chapter 2

Statistical Learning

2.1 Learning from Example

Learning from example is defined as the system that adapts itself from the given example $X^n = \{X_1, X_2, \cdots, X_n\}$ where n is the number of the example. In the case of statistical learning, we suppose that the given example X^n is an independent and identical realization from the true probability distribution $p_0(x)$. Then, we prepare two elements, the one is a learning machine and the other is a learning algorithm. The learning machine is defined as a probability function or probability distribution $p(x|\theta)$ over x and θ is a d-dimensional model parameter. After the learning machine is specified, we choose the learning algorithm which adapts the model parameter θ so that the learning machine $p(x|\theta)$ captures the some characteristics of the true distribution $p_0(x)$ by using example X^n.

2.2 Learning Machines

The learning machine can be classified into two types, identifiable and non-identifiable. We define that the learning machine $p(x|\theta)$ is identifiable when

the Fisher information matrix, whose ij element is given by

$$I_{ij}(\theta) = E_{p(x|\theta)}[\frac{\partial \log p(x|\theta)}{\partial \theta_i} \frac{\partial \log p(x|\theta)}{\partial \theta_j}]$$

is finite and positive definite, where $E_{p(x|\theta)}[.]$ is the expectation over $p(x|\theta)$. In an identifiable case, the family of the learning machines $S = \{p(x|\theta)|\theta\}$, $\theta_1 \neq \theta_2$ implies $p(x|\theta_1) \neq p(x|\theta_2)$, i.e., the mapping from the parameter to the probability distribution is one-to-one. If the learning machine is non-identifiable, then the Fisher information matrix degenerates on the singularities. It is noted that widely used learning machines, such as neural networks, gaussian mixture models, Bayesian networks and stochastic grammar are non-identifiable models.

2.3 Learning Algorithms

If the learning machine $p(x|\theta)$ is specified, then we select the learning algorithm. In this section, We treat three major learning algorithms, maximum likelihood (ML), maximum a posteriori (MAP), and Bayesian method.

In ML method, the parameter θ is treated as a specified value and is chosen to maximize the likelihood function which is defined as

$$\hat{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{n} p(X_i|\theta).$$

The prediction of the new sample X_{n+1} is given by the plug-in $\hat{\theta}_{ML}$ into the model and given by

$$p(X_{n+1}|X^n) = p(X_{n+1}|\hat{\theta}_{ML}).$$

In MAP method, the parameter θ is treated as a random variable and we give the prior distribution of the parameter $p(\theta)$. Then the MAP parameter is chosen at the maximum of the posterior distribution of the parameter θ,

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{n} p(X_i|\theta)p(\theta).$$

Again the prediction of the new sample X_{n+1} is given by

$$p(X_{n+1}|X^n) = p(X_{n+1}|\hat{\theta}_{MAP}).$$

In Bayesian method, same as MAP method, the parameter θ is treated as random variable and we give the prior distribution of the parameter $p(\theta)$. Then we make the posterior distribution of the parameter.

$$p(\theta|X^n) = \frac{\prod_{i=1}^n p(X_i|\theta)p(\theta)}{Z(X^n)},$$

where $Z(X^n)$ is the normalizing constant given by

$$Z(X^n) = \int \prod_{i=1}^n p(X_i|\theta)p(\theta)d\theta.$$

The prediction of the new sample X_{n+1} is given by the averaging the model over the posterior distribution.

$$p(X_{n+1}|X^n) = \int p(X_{n+1}|\theta)p(\theta|X^n)d\theta.$$

2.4 Learning Theory

In this section, we introduce the basic of learning theory, and the some result on the asymptotic case.

2.4.1 Generalization Error and Free Energy

In the learning theory, there are two fundamental quantities. The one is the generalization error and the other is the free energy.

The measure of information content of the random variable x with distribution $p(x)$ is defined by the entropy, which is given by

$$S = -\int p(x)\log p(x)dx.$$

For the measure of two distributions between $q(x)$ and $p(x)$, the relative entropy or the Kullback information from $p(x)$ to $q(x)$ is defined by

$$K(q(x)||p(x)) = \int q(x) \log \frac{q(x)}{p(x)} dx.$$

The Kullback information is non negative and equal to zero only when $q(x) = p(x)$ holds. It is noted that the Kullback information is not symmetric.

The generalization error is defined as the Kullback information from the true distribution to the predictive distribution.

$$G(X^n) = \int p_0(x) \log \frac{p_0(x)}{p(x|X^n)} dx.$$

The generalization error is a criterion of how well the learning machine and the learning algorithm can imitate the true distribution, so the purpose of the learning theory is to seek for the learning machine and the learning algorithm which minimizes the generalization error.

The other important quantity is the free energy, which is defined by

$$F(X^n) = -\log Z(X^n). \tag{2.1}$$

The free energy has a strong relation to the generalization error. To see the relation, we rewrite Bayesian posterior distribution as

$$p(\theta|X^n) = \frac{1}{Z_0(X^n)} \exp(-nH_n(\theta))p(\theta),$$

where $H_n(\theta)$ is the empirical Kullback information,

$$H_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log \frac{p_0(X_i)}{p(X_i|\theta)},$$

and $Z_0(X^n)$ is the normalizing constant. The normalized free energy is defined by $F_0(X^n) = -\log Z_0(X^n)$. Also the empirical entropy $S(X^n)$ is defined by

$$S(X^n) = -\frac{1}{n} \sum_{i=1}^{n} \log p_0(X_i).$$

Then the free energy $F(X^n)$ and the normalized free energy $F_0(X^n)$ satisfy the relation,

$$F(X^n) = F_0(X^n) + nS(X^n).$$

The empirical entropy $S(X^n)$ does not depend on the model. Hence minimization of $F_0(X^n)$ is equivalent to the minimization of $F(X^n)$. Finally, the generalization error and the normalized free energy have the following relation [22],

$$G(X^n) = E_{X_{n+1}}[F_0(X^{n+1}) - F_0(X^n)], \tag{2.2}$$

where the expectation is taken over X_{n+1}.

Moreover the free energy is used for the model selection and the hyperparameter optimization. First, consider the model selection in Bayesian framework. In Bayesian setting, the set of the candidate model $\{M\}$ has the prior probability $P(M)$ and each model M has a prior distribution of the parameter $p(\theta|M)$. Then the joint probability distribution is written as

$$p(x, \theta, M) = p(x|\theta)p(\theta|M)p(M).$$

Therefore, after the showing example X^n, the posterior distribution of the model is given by the Bayes rule

$$P(M|X^n) = \frac{\int p(X^n|\theta)p(\theta|M)p(M)d\theta}{\sum_M \int p(X^n|\theta)p(\theta|M)p(M)d\theta}.$$

Consider the case which p(M) is uniform over M, then

$$P(M|X^n) \propto \int p(X^n|\theta)p(\theta|M)d\theta$$

holds. The right hand side is the normalizing constant $Z(X^n)$ for each model M. Hence, from the relation between $Z(X^n)$ and $F(X^n)$ which is given by the form (2.1), the maximization of the model posterior is equivalent to the minimization of the free energy.

Next, in the case that the prior distribution has a hyperparameter ξ, the free energy is written as

$$F(X^n, \xi) = -\log \int \prod_{i=1}^{n} p(X_i|\theta)p(\theta|\xi)d\theta.$$

Then, by the same argument as the model selection, if we optimize ξ so that $F(X^n, \xi)$ takes the minimum value, the pair of the learning machine and the prior distribution is more probable for the given sample X^n. This procedure is called the empirical Bayes method [14, 10].

2.4.2 Asymptotic Theory

In identifiable models, using the positive definiteness of the Fisher information matrix, it is shown that the normalized free energy is asymptotically expanded as the number of samples n goes to infinity,

$$F_0(X^n) = \frac{d}{2}\log n + O_p(1),$$

where d is the dimension of the parameter θ.

Using this expansion, the following Bayesian information criterion (BIC) is defined [30],

$$(BIC) = -\sum_{i=1}^{n} \log p(X_i|\theta_{MAP}) + \frac{d}{2}\log n.$$

It is noted that this criterion accords with Minimum Description Length (MDL) [28] in Bayesian setting.

In similar fashion, for maximum likelihood setting, from the point of the generalization behavior, Akaike Information Criterion (AIC) is proposed [1].

$$(AIC) = -\sum_{i=1}^{n} \log p(X_i|\theta_{ML}) + d.$$

This criterion also assumes the identifiability. Obviously, in a non-identifiable machine includes stochastic grammar, these criteria do not have justification for the model selection.

Recently, for non-identifiable machines in Bayesian method, the asymptotic property of the free energy and the generalization error is revealed based on algebraic geometry and algebraic analysis [33]. The result shows that the averaged normalized free energy is asymptotically expanded as

$$E_{X^n}[F_0(X^n)] \simeq \lambda \log n - (m-1)\log\log n + O(1),$$

where the coefficients λ and m are the positive rational number and the natural number ($m \geq 1$) respectively and the expectation is taken over the sample set X^n. λ and m are called the learning coefficient and its multiplicity. The each value is determined by the geometric structure of the Kullback information in the parameter space. In identifiable machines, $\lambda = \frac{d}{2}, m = 1$ where d is the dimension of the parameter. However, many non-identifiable machines, it was shown that the value of λ is much smaller than $\frac{d}{2}$ [35, 34, 2]. In Bayesian learning, from the equation 2.2, the expectation of the generalization error has a following asymptotic form,

$$E_{X^n}[G(X^n)] \simeq \frac{\lambda}{n} + o(\frac{1}{n}).$$

Therefore, we see that the smaller λ indicates the smaller generalization error. These result show the effectiveness of non-identifiable machines in Bayesian learning.

Chapter 3

Variational Bayes

As described in the previous chapter, the efficiency of Bayesian learning for non-identifiable models is revealed. However, the practical implementation of Bayesian Learning is difficult at the integration of the posterior distribution and the predictive distribution. In this chapter, two major approximations of Bayesian learning are introduced. Then, variational Bayes learning is explained in detail.

3.1 Approximation schema

The approximation method for Bayesian learning can be divided into two categories. The one is a sampling method and the other is a deterministic method.

The sampling method makes random samples from the target distribution and the most major method is Markov chain Monte Carlo method (MCMC) [13]. MCMC method makes Markov chain sample sequence which converges to the target distribution. MCMC method has the generality over the Bayes learning and the convergence to the target distribution is guaranteed in the limit. However, MCMC still computationally burdens at the practical models that have many model parameters and it was reported that the convergence to the target distribution is slow especially in non-identifiable models [26].

The deterministic method approximates the integral by the optimization of the specific measure of each method and the computational cost is much smaller than sampling method. The famous method is Laplace approximation. The Laplace method approximates the posterior distribution by the gaussian distribution whose mean and variance are determined by the MAP estimator and the Fisher Information Matrix at the MAP estimator. However the Laplace method has two week points in non-identifiable machines. First, in the non-identifiable machines, the Fisher Information Matrix around the true parameter is not positive definite by definition. Second, it is often the case that the integration is not dominated by the maximum of the integrand but the entropy around the maximum. Then, MAP estimator which is only determined by the maximum of the posterior distribution does not well approximates the integral.

3.2 Variational Bayes

Recently, Variational Bayes learning was proposed for an another deterministic method [3]. Variational method itself has the long history for the approximation of the free energy in statistical physics [11, 24]. In Bayesian learning, this method is especially well suited to the machines that are the exponential families with hidden variables and conjugate prior. This class includes HMM and SCFG [6].

In exponential family with hidden variable, the complete likelihood which treats the joint distribution of the complete data

$$(X, Y)^n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\},$$

where the pair of the data X_i and the corresponding hidden variables Y_i are

considered. Then, the free energy is given by

$$\begin{aligned} F(X^n) &= -\log \int\int \prod_{i=1}^{n} p(x_i, y_i|\theta)\varphi(\theta) d\theta dy_i, \\ &= -\log \int\int p(X^n, Y^n, \theta) d\theta dY^n, \end{aligned}$$

which second integration is over hidden variables.

Variational Bayes learning approximates the free energy by the arbitrary conditional trial distribution $q(Y^n, \theta|X^n)$,

$$\begin{aligned} F(X^n) &= -\log \int\int q(Y^n, \theta|X^n) \frac{p(X^n, Y^n, \theta)}{q(Y^n, \theta|X^n)} d\theta dY^n \\ &\leq \int\int q(Y^n, \theta|X^n) \log \frac{q(Y^n, \theta|X^n)}{p(X^n, Y^n, \theta)} d\theta dY^n \\ &\equiv \bar{F}[q] \end{aligned}$$

which uses Jensen's inequality twice and $\bar{F}[q]$ is the functional of q. From this inequality, $\bar{F}[q]$ gives the upper bound of the free energy, that is, the minimization of the functional $\bar{F}[q]$ yields the better approximation of the free energy. The variational free energy is defined by the minimum value of this functional,

$$\bar{F}(X^n) \equiv \min_q \bar{F}[q].$$

By using the Bayes rule,

$$p(Y^n, \theta|X^n) = \frac{P(X^n, Y^n, \theta)}{\int\int P(X^n, Y^n, \theta) d\theta dy},$$

the relationship of the free energy and the variational free energy is given by

$$\bar{F}(X^n) - F(X^n) = \min_q KL(q(Y^n, \theta|X^n)||P(Y^n, \theta|X^n)).$$

From this equation, the difference of F and $\bar{F}$ is the Kullback information from the optimal trial posterior to the true posterior. Therefore if F and

$\bar{F}$ are clarified, the approximation accuracy of variational Bayes learning is obtained.

Next, in variational Bayesian method, for the computational efficiency, the trial distribution $q(Y^n, \theta|X^n)$ is constrained so that the parameter and the hidden variables are independent,

$$q(Y^n, \theta|X^n) = q(Y^n|X^n)r(\theta|X^n),$$

where q and r are the trial posterior distribution of the hidden variables and the parameter respectively.

Under the settings above, the efficient algorithm that minimize $\bar{F}[q]$ with respect to the distributions $q(Y^n|X^n)$ and $r(\theta|X^n)$ is derived by variational method [6]. The necessary condition which should be satisfied by the optimal trial distribution is obtained. By neglecting the constant term, the variational free energy is rewritten as follows,

$$\int\int q(Y^n|X^n)r(\theta|X^n)\log\frac{q(Y^n|X^n)r(\theta|X^n)}{p(X^n,Y^n,\theta)}d\theta dy \tag{3.1}$$

$$=\int r(\theta|X^n)\langle\log\frac{r(\theta|X^n)}{p(X^n,Y^n,\theta)}\rangle_{q(Y^n|X^n)}d\theta \tag{3.2}$$

$$=\int r(\theta|X^n)\log\frac{r(\theta|X^n)}{\frac{1}{C_r}\exp\langle\log p(X^n,Y^n,\theta)\rangle_{q(Y^n|X^n)}}d\theta, \tag{3.3}$$

where we use $\int r(\theta|X^n)d\theta = 1$ in the last equation.

$$\int\int q(Y^n|X^n)r(\theta|X^n)\log\frac{q(Y^n|X^n)r(Y^n|X^n)}{p(X^n,Y^n,\theta)}d\theta y \tag{3.4}$$

$$=\int q(Y^n|X^n)\langle\log\frac{q(Y^n|X^n)}{p(X^n,Y^n|\theta)}\rangle_{r(\theta|X^n)}dy \tag{3.5}$$

$$=\int q(Y^n|X^n)\log\frac{q(Y^n|X^n)}{\frac{1}{C_q}\exp\langle\log p(X^n,Y^n|\theta)\rangle_{r(\theta|X^n)}}dy, \tag{3.6}$$

where we use $\int q(Y^n|X^n)dy = 1$ in the last equation. Equations [3.3, 3.6] are Kullback informations from the trial distributions, then the necessary

condition satisfies the following,

$$q(Y^n|X^n) = \frac{1}{C_q} \exp\langle \log p(X^n, Y^n|\theta) \rangle_{r(\theta|X^n)} \tag{3.7}$$

$$r(\theta|X^n) = \frac{1}{C_r} \exp\langle \log p(X^n, Y^n, \theta) \rangle_{q(Y^n|X^n)}, \tag{3.8}$$

where C_q, C_r are the normalizing constants. Optimization of the variational free energy is achieved by iterating these equations alternately. When the prior distribution $\varphi(\theta)$ is taken from the conjugate family, these iteration are very efficient, and computational cost is the same order of the ordinary Expectation-Maximization algorithm[8]. Moreover, this algorithm is a kind of natural gradient method, each iteration minimizes the cost function monotonically and at least the convergence to the local minimum is guaranteed [29].

Chapter 4

Stochastic Grammar

Grammatical inference, also known as grammatical induction is the system of learning grammars from the given symbols. Grammatical inference was mainly treated in a formal language. In the formal language, there is a famous classification of the grammar which is known as the Chomsky hierarchy. Indeed, HMM and SCFG are the stochastic version of regular grammar and context free grammar in Chomsky hierarchy.

4.1 Formal Language

The formal grammar G is defined as the set of the quad-tuple (N, Σ, P, S), a finite set N of nonterminal symbols, a finite set Σ of terminal symbols, a finite set of production rules P which has the form

$$(\Sigma \cup N)^* N (\Sigma \cup N)^* \rightarrow (\Sigma \cup N)^*,$$

where $*$ is Kleene star operator. The language of a formal grammar G, denoted as $L(G)$ is defined as all strings over Σ that can be generated by the starting symbol $S \in N$ and then applying the production rules in P until no nonterminal symbols are present.

Noam Chomsky classified the formal language to four types and each class is characterized by the constraints of the production rule [7]. Each grammar

has the recognition machine which can accept the language generated from the grammar. The grammar in order of generating the restricted language is as follows,

- Regular grammar, $A \to Aa$ or $A \to a \leftrightarrow$ Finite state automaton
- Context free grammar, $A \to AB$ or $A \to \gamma \leftrightarrow$ Pushdown automaton
- Context sensitive grammar, $\alpha A \beta \to \alpha B \beta \leftrightarrow$ Linear-bounded automaton
- Unrestricted grammar, unrestricted, $\leftrightarrow$ Turing machine

where $A, B \in N$, $a \in \Sigma$ and $\alpha, \beta \in (N \cup \Sigma)$.

In practice, regular grammar and context free grammar are important by the reason that the acceptance of the language is decided efficiently. In computer science, regular grammar is used as regular expression and context free grammar is used as the syntax of the programming language.

4.2 Grammatical Inference

In grammatical inference, under the given strings, the sufficient condition for the identification of the grammar consists of three parts which are nested from the top.

- inference of a grammar hierarchy
- inference of the number of non-terminals under the given grammar
- inference of production rules under the given grammar and the number of non-terminals.

In this thesis, we assume that the hierarchy of the grammar is given and inference of the number of non-terminals and identification of the production rules are considered.

4.3 Hidden Markov Model

Stochastic grammar is made by giving the probability for each production rule P to a formal grammar. By making the grammar stochastic, the several inference task on the grammar can be quantitatively evaluated.

HMM is the stochastic version of the regular grammar and widely used in many fields for modeling nonlinear time series data. HMM is a standard method in speech recognition, natural language processing, and bioinformatics [5, 27].

4.3.1 Definition

We define HMM and briefly review the inference procedures. HMM assumes that observed T length M-dimensional vector $X_{1:T} = \{X_1, \ldots, X_T\}$ whose element satisfies $X_{t,m} = \{0, 1\}$ and $\sum_{m=1}^{M} X_{t,m} = 1$. Moreover, we assume $X_{1:T}$ was produced by K-dimensional hidden states vector $Y_{1:T} = \{Y_1, \ldots, Y_T\}$ whose element satisfies $Y_{t,i} = \{0, 1\}$ and $\sum_{i=1}^{K} Y_{t,i} = 1$. The hidden sates $Y_{1:T}$ is assumed to obey a first-order Markov process. Then the joint distribution of the data and the hidden states is given by

$$p(X_{1:T}, Y_{1:T}) = p(Y_1)p(X_1|Y_1) \prod_{t=2}^{T} p(Y_t|Y_{t-1})p(X_t|Y_t),$$

where $p(Y_t)$ is the prior probability of the first hidden state and this matches the probability of the start symbol in a formal grammar, $p(Y_t|Y_{t-1})$ is the transition probability from state Y_{t-1} to Y_t, and $p(X_t|Y_t)$ are the emission probabilities for hidden state Y_t emit the vector X_t. The probability of the observation $X_{1:T}$ is given by summing over all possible hidden state sequences,

$$p(X_{1:T}|\theta) = \sum_{Y_{1:T}} p(X_{1:T}, Y_{1:T}|\theta). \tag{4.1}$$

The set of the model parameter θ is given by the triplet (A, B, π),

$$\begin{aligned}
\theta &= (A, B, \pi) \\
A &= \{a_{ij}\} : a_{ij} = p(Y_{t,i} = 1 | Y_{t-1,j} = 1), \sum_{i=1}^{K} a_{ij} = 1 \\
B &= \{b_{im}\} : b_{im} = p(X_{t,m} = 1 | Y_{t,i} = 1), \sum_{m=1}^{M} b_{im} = 1 \\
\pi &= \{\pi_i\} : \pi_i = p(Y_{1,i} = 1), \sum_{i=1}^{K} \pi_i = 1,
\end{aligned}$$

where the last equation represents constraints of each parameter and we assume that the parameter is stationary over t. Using these parameters, joint likelihood is given by

$$p(X_{1:T}, Y_{1:T} | \theta) = \prod_{i=1}^{K} \pi_i^{y_{1,i}} \prod_{t=2}^{T} \prod_{i=1}^{K} \prod_{j=1}^{K} a_{ij}^{y_{t-1,j} y_{t,j}} \prod_{t=1}^{L} \prod_{i=1}^{K} \prod_{m=1}^{M} b_{im}^{y_{t,i} x_{t,m}}.$$

4.3.2 Variational Bayesian Learning

Next, we derive the iterative algorithm for variational Bayesian HMM [23, 6]. First, the log complete likelihood of a sequence is given by,

$$\begin{aligned}
&\log p(X_{1:T}, Y_{1:T} | \theta) = \\
&\sum_{t=1}^{T} \sum_{i=i}^{K} \sum_{j=1}^{K} y_{t,i} y_{t-1,j} \log a_{ij} + \sum_{t=1}^{T} \sum_{i=1}^{K} \sum_{j=1}^{K} y_{t,i} x_{t,m} \log b_{im}.
\end{aligned} \tag{4.2}$$

The prior probabilities of each parameter are given by Dirichlet distribution with hyperparameter ϕ_0, ξ_0.

$$\begin{aligned}
\varphi(A | \phi_0) &= \frac{\Gamma(K\phi_0)^K}{\Gamma(\phi_0)^{K^2}} \prod_{i=1}^{K} \prod_{j=1}^{K} a_{ij}^{\phi_0 - 1}, \\
\varphi(B | \xi_0) &= \frac{\Gamma(M\xi_0)^K}{\Gamma(\xi_0)^{KM}} \prod_{i=1}^{K} \prod_{m=1}^{M} b_{im}^{\xi_0 - 1}.
\end{aligned} \tag{4.3}$$

Using the equation (3.8), the variational posterior of the parameters is given by

$$\begin{aligned} r(\theta|X_{1:T}) &= \frac{1}{C_r}\varphi(\theta)\exp\langle\log p(X_{1:T}, Y_{1:T}|\theta)\rangle_{q(Y_{1:T})} \\ &= \frac{1}{C_r}\varphi(\theta)\exp\{\sum_{t=2}^{T}\sum_{j=1}^{K}\sum_{j=1}^{K}\{\langle y_{t,i}y_{t-1,j}\rangle_{q(Y_{1:T}|X_{1:T})}\log a_{ij} \\ &+ \sum_{t=1}^{T}\sum_{i=1}^{K}\sum_{m=1}^{M}\langle y_{t,i}\rangle_{q(Y_{1:T}|X_{1:T})}x_{t,m}\log b_{im}\}\}. \end{aligned} \tag{4.4}$$

We define the expected sufficient statistics.

$$\begin{aligned} n_{i,j} &\equiv \sum_{t=2}^{T}\langle y_{t,i}y_{t-1,j}\rangle_{q(Y_{1:T}|X_{1:T})}, \\ n_{i,m} &\equiv \sum_{t=1}^{T}\langle y_{t,i}\rangle_{q(Y_{1:T}|X_{1:T})}x_{t,m}, \\ n_i &= \sum_{j=1}^{K} n_{i,j} = \sum_{i=1}^{M} n_{i,m}. \end{aligned}$$

Then, the variational posterior is written as

$$\begin{aligned} r(A|X_{1:T}) &\propto a_{ij}^{n_{ij}+\phi_0-1}, \\ r(B|X_{1:T}) &\propto b_{im}^{n_{im}+\xi_0-1}. \end{aligned} \tag{4.5}$$

which are again Dirichlet distributions.

For the posterior distribution of the hidden variables, using the equation (3.7),

$$\begin{aligned} q(Y_{1:T}|X_{1:T}) &= \frac{1}{C_q}\exp\langle\log p(X_{1:T}, Y_{1:T}|\theta)\rangle_{r(\theta)} \\ &= \frac{1}{C_q}\exp\{\sum_{t=2}^{T}\sum_{i=1}^{K}\sum_{j=1}^{K} y_{t,i}y_{t-1,i}\langle\log a_{ij}\rangle_{r(\theta|X_{1:T})} \\ &+ \sum_{t=1}^{T}\sum_{i=1}^{K}\sum_{m=1}^{M} y_{t,i}x_{t,m}\langle\log b_{im}\rangle_{r(\theta|X_{1:T})}\}. \end{aligned} \tag{4.6}$$

From the equation

$$\langle \log \pi_i \rangle_{Dirichlet(\pi|\mu)} = \psi(\mu_i) - \psi(\sum_{i=1}^{K} \mu_i),$$

where $\psi(x)$ is the psi function and is defined by

$$\psi(x) = \frac{d}{dx} \log \Gamma(x).$$

Using the expected sufficient statistics, the expectation of the log parameter is given by

$$\langle \log a_{ij} \rangle = \psi(n_{ij} + \phi_0) - \psi(\sum_{j=1}^{K} (n_{ij} + \phi_0))$$

$$\langle \log b_{im} \rangle = \psi(n_{im} + \phi_0) - \psi(\sum_{m=1}^{M} (n_{im} + \phi_0)).$$

Optimization of the variational free energy is computed by the equations (4.6) and (4.4) iteratively. In HMM, the method which efficiently computes the posterior of the hidden variable (4.6) is known as forward backward algorithm [5].

4.3.3 Example

We give an example of simple HMM which has two non-terminal symbols $\{A, B\}$ and two terminal symbols $\{0, 1\}$ (figure 4.1). Let the observation sequence is length 3 and (1 0 0) and the hidden non-terminal sequence is (A A B). The parameter of the initial distribution is given by π, the transition probability is given by matrix A and the emission probability is given by B.

$$\pi = \begin{pmatrix} \pi_A \\ \pi_B \end{pmatrix}$$

$$A = \begin{pmatrix} a_{AA} & a_{AB} \\ a_{BA} & a_{BB} \end{pmatrix}$$

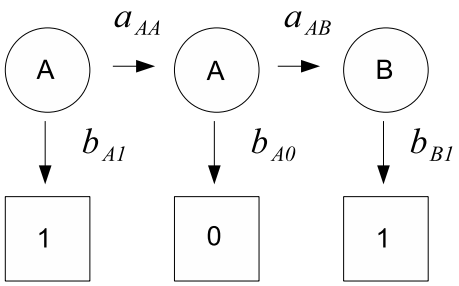

Figure 4.1: Example of HMM

$$B = \begin{pmatrix} b_{A0} & b_{A1} \\ b_{B0} & b_{B1} \end{pmatrix}.$$

Then, the joint probability of the data and hidden variables is given by

$$p(X, Y|\theta) = \pi_A b_{A0} a_{AA} b_{A0} a_{AB} b_{B1}.$$

In ordinary setting, we can only observe the data sequence. Then, the probability of the data is given by the all possible combinations of the hidden variables $\alpha_0\alpha_1\alpha_2$ where α_i is an element of $\{A, B\}$.

$$p(X|\theta) = \sum_{\alpha_0\alpha_1\alpha_2} \pi_{\alpha_0} b_{\alpha_0 0} a_{\alpha_0\alpha_1} b_{\alpha_1 0} a_{\alpha_1\alpha_2} b_{\alpha_2 1}.$$

4.4 Stochastic Context Free Grammar

SCFG is the stochastic version of the context free grammar and mainly used in natural language processing, RNA secondary structure analysis, and knowledge discovery.

4.4.1 Definition

We define SCFG. In this thesis, without loss of generality, we assume that SCFG is written by Chomsky normal form (CNF). In CNF, rewriting rules

$\alpha \to \beta$ are constrained to the two types $A \to BC$ (we call this rule the non-terminal emitting rule) and $A \to c$ (we call this rule the terminal emitting rule). Where A, B are non-terminal symbols and c is a terminal symbol. SCFG assumes that we observe L-length M-dimensional vector $X_{1:L} = \{X_1, \ldots, X_L\}$ whose elements satisfies $X_{l,m} = \{0, 1\}$ and $\sum_{m=1}^{M} X_{l,m} = 1$. Let the grammar has the K non-terminal symbols. Then, the complete likelihood is given by the pair of the data and the tree t which is the element of the set T of L leaves tree. Using the model parameter θ, we can write the joint likelihood as

$$\begin{aligned} p(x|\theta) &= \sum_{t \in T} p(x, t|\theta) \\ \theta &= \{A, B\} \\ A &= \{a^i_{jk} \ (1 \leq i, j, k \leq K)\}, \sum_{j,k=1}^{K} a^i_{jk} = 1 \\ B &= \{b_{im} \ (1 \leq i \leq K, 1 \leq m \leq M)\}, \sum_{m=1}^{M} b_{im} = 1, \end{aligned} \tag{4.7}$$

where a^i_{jk} is a probability that non-terminal symbol i rewrites to the non-terminal symbol (j, k) and b_{im} is a probability that non-terminal symbol rewrites to the terminal symbol m.

In CNF, the tree t which has L leaves is completely characterized by the left most derivation, $(\alpha_1 \to \beta_1, \alpha_2 \to \beta_2, \ldots, \alpha_{2L} \to \beta_{2L})$ where $(\alpha_i \to \beta_i)$ are rewriting rules that obeys CNF. Moreover, we introduce two hidden variables. The first is $y^i_{t,jk}$ which is $2L$ length K^3-dimensional vector whose element is 1 if the tth rule in the most left derivation is non-terminal emitting rule $i \to (j, k)$ otherwise 0. The second is $\hat{y}_{t,i}$ which is $2L$ length K-dimensional vector whose element is 1 if the tth rule in the left most derivation is terminal emitting rule $i \to m$ otherwise 0. It is noted that $\hat{y}$ is completely determined

by y. Then, the complete likelihood of given sequence is written by

$$p(x,t) = p(x, y_{1:2L}|\theta) = \prod_{l=1}^{2L}\prod_{i=1}^{K}\prod_{j=1}^{K}\prod_{k=1}^{K} {a_{jk}^{i}}^{y_{l,jk}^{i}} \prod_{i=1}^{K}\prod_{m=1}^{M} {b_{im}}^{\hat{y}_{l,i}\hat{x}_{l,m}},$$

where $\hat{x}_{l,m}$ has one-to-one mapping to $x_{h,m}$ and the index h is determined by the number of terminal emitting rules in the subset of the most left derivation.

4.4.2 Variational Bayesian Learning

We derive the variational Bayesian algorithm for SCFG [20]. First, the log complete likelihood of a sequence is given by,

$$\log p(X_{1:L}, Y_{1:2L}|\theta) = \sum_{l=1}^{2L}\sum_{i=i}^{K}\sum_{j=1}^{K}\sum_{k=1}^{K} y_{l,j,k}^{i} \log a_{jk}^{i} \sum_{i=1}^{K}\sum_{m=1}^{M} \hat{y}_{l,i}\hat{x}_{l,m} \log b_{im}.$$

Prior probabilities of each parameter are given by Dirichlet distribution with hyperparameter ϕ_0, ξ_0.

$$\begin{aligned}\varphi(A|\phi_0) &= \frac{\Gamma(K^2\phi_0)^K}{\Gamma(\phi_0)^{K^3}} \prod_{i=1}^{K}\prod_{j=1}^{K}\prod_{k=1}^{K} {a_{jk}^{i}}^{\phi_0-1}, \\ \varphi(B|\xi_0) &= \frac{\Gamma(M\xi_0)^K}{\Gamma(\xi_0)^{KM}} \prod_{i=1}^{K}\prod_{m=1}^{M} b_{im}^{\xi_0-1}. \end{aligned} \tag{4.8}$$

Using the equation (3.8), the variational posterior of the parameters is given by

$$\begin{aligned} r(\theta|X_{1:T}) &= \frac{1}{C_r}\varphi(\theta)\exp\langle \log p(X_{1:L}, Y_{1:2L}|\theta)\rangle_{q(Y_{1:2L}|X_{1:L})} \\ &= \frac{1}{C_r}\varphi(\theta)\exp\{\sum_{l=1}^{2L}\sum_{i=1}^{K}\sum_{j=1}^{K}\sum_{k=1}^{K}\{\langle y_{l,jk}^{i}\rangle_{q(Y_{1:2L}|X_{1:L})}\log a_{ij} \\ &+ \sum_{l=1}^{2L}\sum_{i=1}^{K}\sum_{m=1}^{M}\langle \hat{y}_{l,i}\rangle_{q(Y_{1:2L},|X_{1:L})}\hat{x}_{l,m}\log b_{im}\}\}. \end{aligned} \tag{4.9}$$

We define the expected sufficient statistics.

$$n^i_{jk} \equiv \sum_{l=1}^{2L} \langle y^i_{l,jk} \rangle_{q(Y_{1:2L}|X_{1:L})},$$
$$n_{im} \equiv \sum_{L=1}^{2L} \langle \hat{y}_{l,i} \rangle_{q(Y_{1:2L}|X_{1:L})} \hat{x}_{l,m},$$
$$n_i = \sum_{j=1}^{K} \sum_{k=1}^{K} n^i_{jk} = \sum_{m=1}^{M} n_{i,m}.$$

Then, the variational posterior is written as

$$r(A|X_{1:L}) \propto {a^i_{jk}}^{n^i_{jk}+\phi_0-1},$$
$$r(B|X_{1:L}) \propto b^{n_{im}+\xi_0-1}_{im},$$

which are again Dirichlet distributions.

For the posterior of the hidden variables, using the equation (3.7),

$$\begin{aligned} q(Y_{1:2L}|X_{1:L}) &= \frac{1}{C_q} \exp\langle \log p(X_{1:L}, Y_{1:2L}|\theta) \rangle_{r(\theta)} \\ &= \frac{1}{C_q} \exp\{ \sum_{l=1}^{2T} \sum_{i=1}^{K} \sum_{j=1}^{K} \sum_{k=1}^{K} y^i_{t,jk} \langle \log a^i_{jk} \rangle_{r(\theta|X_{1:L})} \\ &+ \sum_{l=1}^{2L} \sum_{i=1}^{K} \sum_{m=1}^{M} \hat{y}_{l,i} \hat{x}_{l,m} \langle \log b_{im} \rangle_{r(\theta|X_{1:L})} \}. \end{aligned} \tag{4.10}$$

Just same as HMM case, using the expected sufficient statistics, the expectation of logarithm parameter is given by

$$\langle \log a^i_{jk} \rangle = \psi(n^i_{jk} + \phi_0) - \psi(\sum_{j=1}^{K} \sum_{k=1}^{K} (n^i_{jk} + \phi_0))$$
$$\langle \log b_{im} \rangle = \psi(n_{im} + \phi_0) - \psi(\sum_{m=1}^{M} (n_{im} + \phi_0)).$$

Optimization of the variational free energy is computed by the equations (4.10) and (4.9) iteratively. In SCFG, the method which efficiently computes the posterior of the hidden variable (4.10) is known as inside-outside algorithm [21].

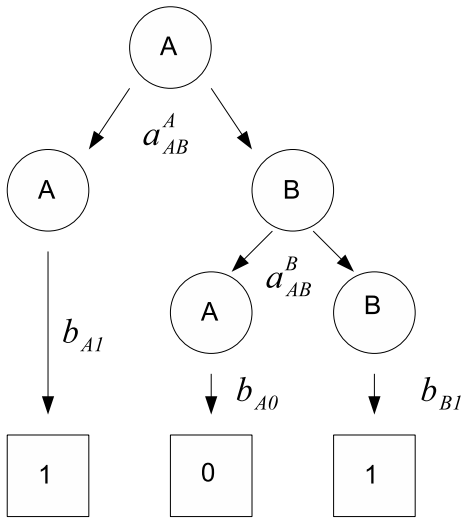

Figure 4.2: Example of SCFG

4.4.3 Example

We give an example of simple SCFG which has two non-terminal symbols $\{A, B\}$ and two terminal symbols $\{0, 1\}$ (figure 4.2). Let the observation sequence is length 3 and (1 0 0) and the hidden non-terminal tree is (A(A B(A B))) which is written by S-expression. The parameter of the initial distribution is given by π, the rewriting probability is given by matrix A and the emission probability is given by B.

$$\pi = \begin{pmatrix} \pi_A \\ \pi_B \end{pmatrix}$$

$$A = \begin{pmatrix} a^A_{AA} & a^A_{AB} & a^A_{BA} & a^A_{BB} \\ a^B_{AA} & a^B_{AB} & a^B_{BA} & a^B_{BB} \end{pmatrix}$$

$$B = \begin{pmatrix} b_{A0} & b_{A1} \\ b_{B0} & b_{B1} \end{pmatrix}.$$

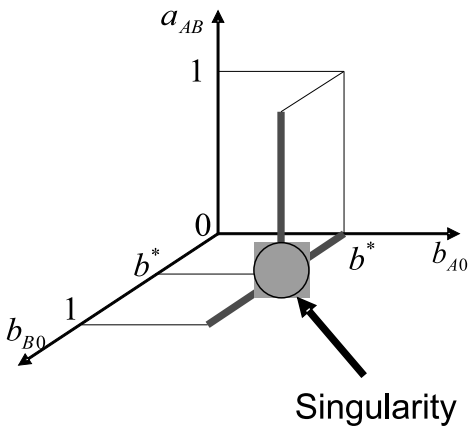

Figure 4.3: Singularity of the Stochastic Grammar. The thick line shows the set of $KL(p_0||p) = 0$ and the circle indicates the singular point.

Then, the joint probability of the data and the hidden variables is given by

$$p(X, Y|\theta) = \pi_A a^A_{AB} b_{A1} a^B_{AB} b_{A0} b_{b1}.$$

In ordinary setting, we can only observe the data sequence. Then, the probability of the data is given by the all possible combinations of the tree and the corresponding hidden variables $(\alpha_0(\alpha_1(\alpha_2 \quad \alpha_3)\alpha_4))$ and $(\alpha_0(\alpha_1\alpha_2(\alpha_3 \quad \alpha_4)))$ where α_i is $\{A, B\}$.

$$\begin{aligned} p(X|\theta) = &\sum_{\alpha_0\alpha_1\alpha_2\alpha_3\alpha_4} \pi_{\alpha_0} a^{\alpha_0}_{\alpha_1\alpha_2} a^{\alpha_1}_{\alpha_3\alpha_4} b_{\alpha_3 1} b_{\alpha_4 0} b_{\alpha_2 1} \\ &+ \sum_{\alpha_0\alpha_1\alpha_2\alpha_3\alpha_4} \pi_{\alpha_0} a^{\alpha_0}_{\alpha_1\alpha_2} b_{\alpha_1 1} a^{\alpha_2}_{\alpha_3\alpha_4} b_{\alpha_3 0} b_{\alpha_4 1}. \end{aligned}$$

4.5 Singularity of the Stochastic Grammar

In case that the learning machine has redundant non-terminal symbols against the true distribution, the set of the zero of the Kullback information from the true distribution to the learning machine is not the point but the algebraic

variety that has positive dimensions. Moreover, the variety has the singular points. We show this condition on the simplest example. Let the true distribution $p_0(x)$ has one non-terminal symbol and two terminal symbols whose parameter is constant A^*, B^*. The learning machine $p(x|\theta)$ has two non-terminal symbols whose parameter is A, B.

$$A^* = \begin{pmatrix} 1 \end{pmatrix}, \quad B^* = \begin{pmatrix} b^* & 1-b^* \end{pmatrix}$$

$$A = \begin{pmatrix} 1-a_{AB} & a_{AB} \\ 0 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} b_{A0} & 1-b_{A0} \\ b_{B0} & 1-b_{B0} \end{pmatrix}.$$

The Kullback information from the true distribution to the learning machine is given by,

$$KL(p_0||p) = \int p_0(x) \log \frac{p_0(x)}{p(x|\theta)} dx. \tag{4.11}$$

Then, the set of $KL(p_0||p) = 0$ in the parameter space is plotted by figure (4.3). This set is composed by two lines which are crossing each other and has singular point at $(a_{AB}, b_{A0}, b_{B0}) = (0, b^*, b^*)$. On this set, the Fisher information matrix is degenerate and we cannot use conventional statistical approach such as BIC, MDL.

4.6 Example of Variational Approximation

We show the example of the variational approximation for Bayes posterior distributions. Consider the example of the previous section and we set the true model parameter b^* to 0.7. We sample 200 data from the true distribution and estimate the posterior distribution of each method (figure 4.4, 4.5). For Bayes posterior, we use the MCMC method. For variational Bayes, we estimate the variational posterior by iterative algorithm. Then we sample the parameters from the variational posterior. In this example, we show that the variational posterior comprises the sub-variety of the Bayes posterior.

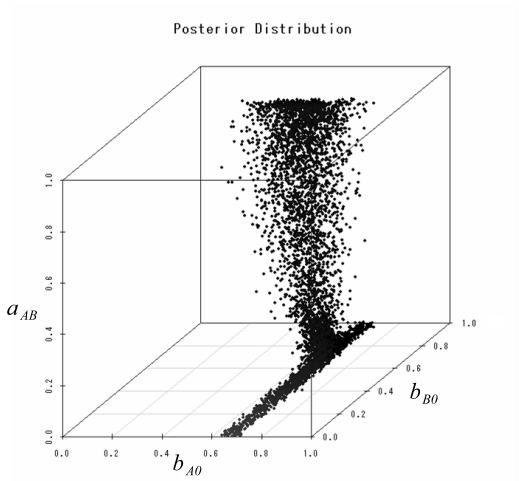

Figure 4.4: Bayes posterior distribution

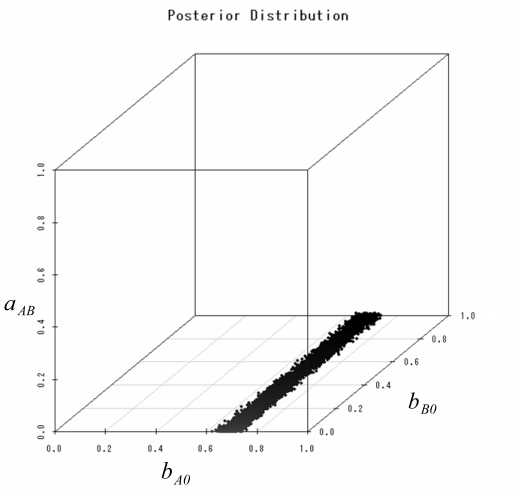

Figure 4.5: Variational Bayes posterior distribution

Chapter 5

Main Theorem

In this chapter, as main contribution of this thesis, we derive the asymptotic variational free energy in cases of HMM and SCFG.

5.1 Assumption

We assume following conditions.

(A1) The true distribution of HMM and SCFG has K_0 non terminal symbols and M terminal symbols and K_0 is minimum degree in the parametrization [19].

(A2) The learning machine is given by equations (4.1), (4.7) and includes the true distribution, namely, the number of non terminals K satisfies the inequality $K_0 \leq K$.

(A3) The prior distribution of the learning machine is Dirichlet distribution which is given by (4.3), (4.8).

(A4) In finite dimensional HMM and SCFG, variational Bayes estimator is consistent.

5.2 Normalized Variational Free Energy

We define the normalized variational free energy.

$$\begin{aligned}\bar{F}_0(X^n) &\equiv \bar{F}(X^n) - nS(X^n) \\ &= \min_{q,r}\{\log p_0(X^n) - \langle \log \frac{p(X^n, Y^n|\theta)}{q(Y^n|X^n)} \rangle_{q(Y^n|X^n)r(\theta|X^n)} \\ &+ \int r(\theta|X^n) \log \frac{r(\theta|X^n)}{\varphi(\theta)} d\theta\}. \end{aligned} \tag{5.1}$$

Using equation (3.7), the second term of $\bar{F}_0(X^n)$ is given by

$$-\langle \log \frac{p(X^n, Y^n|\theta)}{q(Y^n|X^n)} \rangle_{q(Y^n|X^n)r(\theta|X^n)} = -\log C_q.$$

First, we consider the first and the second terms of $\bar{F}_0(X^n)$.

Lemma 1. *In HMM and SCFG, the first and the second term of $\bar{F}_0(X^n)$ are lower bounded by $O_p(1)$.*

$$\log p_0(X^n) - \log C_q \geq \log \frac{p_0(X^n)}{p(X^n|\theta_{ML})} \to O_p(1).$$

Proof. Using Jensen's inequality, the second term is lower bounded by the likelihood of the Maximum Likelihood parameter.

$$-\log C_q \geq -\log \langle p(X^n|\theta) \rangle_{r(\theta|X^n)} \geq -\log p(X^n|\theta_{ML}).$$

Therefore, inequality

$$\log p_0(X^n) - \log C_q \geq \log \frac{p_0(X^n)}{p(X^n|\theta_{ML})}$$

holds. Additionally, HMM and SCFG which have L length and M-dimensional output, are sub-models of the regular model which has L^M-dimensional parameter. In this class of models, it is well known that likelihood ratio is converged $O_p(1)$ as the number of samples n goes to infinity, which concludes the lemma. □

5.3 HMM case

The true HMM has K_0 hidden states and the model parameter is given by the constant $\{A^*, B^*\}$ as follows.

$$
\begin{aligned}
p(X_{1:T}|\theta^*) &= \sum_{Y_{1:T}} p(X_{1:T}, Y_{1:T}|\theta^*). \\
\theta^* &= \{A^*, B^*\} \\
A^* &= \{a_{ij}^*\} \quad (1 \leq i, j \leq K_0), \sum_{i=1}^{K_0} a_{ij}^* = 1 \\
B^* &= \{b_{im}^*\} \quad (1 \leq i \leq K_0, 1 \leq m \leq M), \sum_{m=1}^{M} b_{im}^* = 1
\end{aligned}
$$

Theorem 1. *[16], [18] Under the conditions (A1) to (A4), the normalized variational free energy of HMM satisfies*

$$\bar{F}(X^n) = \bar{\lambda} \log n + O_p(1) \quad (n \to \infty),$$

as the number of samples n goes to infinity, where the learning coefficient of HMM $\bar{\lambda}$ *is given by*

$$
\bar{\lambda} = \begin{cases} \frac{1}{2}\{K_0(K_0-1) + K_0(M-1)\} + K_0(K-K_0)\phi_0 \\ \quad (\phi_0 \leq \frac{K+K_0+M-2}{2K_0}) \\ \frac{1}{2}\{K(K-1) + K(M-1)\} \\ \quad (\phi_0 > \frac{K+K_0+M-2}{2K_0}). \end{cases} \tag{5.2}
$$

Proof. First, we consider the lower bound. The second term of equation (5.1) is Kullback information from the variational posterior to the prior. In HMM, this term is Kullback information of Dirichlet distribution which is written

by the expected sufficient statistics n_i, n_{ij}, n_{im}.

$$\begin{aligned}
&\int r(\theta|X^n) \log \frac{r(\theta|X^n)}{\varphi(\theta)} d\theta \\
&= KL(r(\theta|X^n)||\varphi(\theta)) = KL(r(A|X^n)||\varphi(A)) + KL(r(B|X^n||\varphi(B)) \\
&= \sum_{i=1}^{K} \{\log \Gamma(n_i + K\phi_0) - n_i \Psi(n_i + K\phi_0) - \log \Gamma(K\phi_0) + K \log \Gamma(\phi_0) \\
&- \sum_{j=1}^{K} \{\log \Gamma(n_{ij} + \phi_0) - n_{ij} \Psi(n_{ij} + \phi_0)\}\} \\
&+ \sum_{i=1}^{K} \{\log \Gamma(n_i + M\xi_0) - n_i \Psi(n_i + M\xi_0) - \log \Gamma(M\xi_0) + M \log \Gamma(\xi_0) \\
&- \sum_{m=1}^{M} \{\log \Gamma(n_{im} + \xi_0) - n_{im} \Psi(n_{im} + \xi_0)\}\}. \qquad (5.3)
\end{aligned}$$

By using the Stirling formula for gamma function and Ψ function, we have

$$\begin{aligned}
\log \Gamma(x) &= (x - \frac{1}{2}) \log x - x + \frac{1}{2} \log 2\pi + \frac{1}{12x} + o(\frac{1}{x}), \\
\Psi(x) &= \log x - \frac{1}{2x} + o(\frac{1}{x}). \qquad (5.4)
\end{aligned}$$

The Kullback information is asymptotically expanded as,

$$\begin{aligned}
&KL(r(\theta|X^n)) \\
&= \sum_{i=1}^{K} \{(K\phi_0 - \frac{1}{2}) \log(n_i + K\phi_0) - \sum_{j=1}^{K} \{(\phi_0 - \frac{1}{2}) \log(n_{ij} + \phi_0)\} \\
&+ (M\xi_0 - \frac{1}{2}) \log(n_i + M\xi_0) - \sum_{m=1}^{M} \{(\xi_0 - \frac{1}{2}) \log(n_{im} + \xi_0)\}\} + O_p(1). \qquad (5.5)
\end{aligned}$$

Consider the minimization of above equation with respect to n_i, n_{ij}, n_{im}. We divide the problem into two cases $\phi_0 \leq \frac{1}{2}$ and $\phi_0 > \frac{1}{2}$.

In case $\phi_0 \leq \frac{1}{2}$, it it obvious that minimization (5.5) is achieved redundant n_i, n_{ij}, n_{im} $(K_0 < i, j \leq K)$ have order $O_p(1)$.

In case $\phi_0 > \frac{1}{2}$, from the equation (5.5) and constraint $n_i = \sum_{j=1}^{K} n_{ij}, n_i = \sum_{m=1}^{M} n_{im}$, minimization is achieved that n_{ij}, n_{im} have the same order as n_i. Therefore, let $n_i = p_i n^{\alpha_i}, n_{ij} = q_{ij} n^{\alpha_i}, n_{im} = r_{im} n^{\alpha_i}$ $(p_i, q_{ij}, r_{im} > 0, 0 \leq \alpha_i \leq 1)$. Then,

$$
\begin{aligned}
&KL(r(\theta|X^n)||\varphi(\theta)) \\
&= \sum_{i=1}^{K} \alpha_i \{(K\phi_0 + \frac{M}{2} - 1) - \sum_{j=1, \alpha_j \neq 0}^{K} (\phi_0 - \frac{1}{2})\} \log n + O_p(1)
\end{aligned}
$$

In minimizing above equation, by assumption (A4), $\alpha_i = 1$ $(1 \leq i \leq K_0)$ holds. Moreover, because of the α_i satisfies $\alpha_i \geq 0$ and linearity of the objective function, we divide the function to the true K_0 terms and additional l $(0 \leq l \leq K - K_0)$ terms that has $\alpha_i = 1$.

$$
\begin{aligned}
&\frac{KL(r(\theta|X^n)||\varphi(\theta))}{\log n} \\
&= (K\phi_0 + \frac{M}{2} - 1)K_0 - (\phi_0 - \frac{1}{2})K_0^2 \\
&+ (K\phi_0 + \frac{M}{2} - 1)l - (\phi_0 - \frac{1}{2})((K_0 + l)^2 - K_0^2) + O_p(1). \qquad (5.6)
\end{aligned}
$$

Minimization of the second term with respect to l has two cases which is determined by the value of ϕ_0 and given by

$$
\begin{cases}
l = 0 \ (\phi_0 \leq \frac{K + K_0 + M - 2}{2K_0}) \\
l = K - K_0 \ (\phi_0 > \frac{K + K_0 + M - 2}{2K_0}).
\end{cases}
$$

Finally, we substitute l into equation (5.6) and get the result.

Next, we consider the upper bound. From the discussion of the lower bound, it is only necessary finding the concrete parameters which satisfies lower bound constraints and the first and the second term of $\bar{F}(X^n)$ (5.1) become $O_p(1)$ as the number of samples n goes to infinity. The first and the

second term are rewritten as follows.

$$\begin{aligned}
&\log p_0(X^n) - \log C_q \\
&= \log \sum_{Y^n} p_0(X^n, Y^n|\theta^*) - \log \sum_{Y^n} \exp\langle \log p(X^n, Y^n|\theta)\rangle_{r(\theta)} \\
&= \log \sum_{Y^n} \exp(\log p_0(X^n, Y^n|\theta^*) - \langle \log p(X^n, Y^n|\theta)\rangle_{r(\theta)}). \qquad (5.7)
\end{aligned}$$

Substituting the concrete form of HMM, this equation becomes

$$\begin{aligned}
&\log \sum_{Y^n} \exp(\sum_{t=2}^{T}\sum_{i=1}^{K}\sum_{j=1}^{K} y_{t-1,i}y_{t,j}\{\log a_{ij}^* - \Psi(n_{ij}+\phi_0) + \Psi(n_i + K\phi_0)\} \\
&+ \sum_{t=1}^{T}\sum_{i=1}^{K}\sum_{m=1}^{M} y_{t,i}x_{i,m}\{\log b_{im}^* - \Psi(n_{im}+\xi_0) + \Psi(n_i + M\xi_0)\}). \qquad (5.8)
\end{aligned}$$

We evaluate each parameter separately. For example

$$\log a_{ij}^* - \Psi(n_{ij}) + \Psi(n_i)$$

satisfy

$$\begin{aligned}
&\log a_{ij}^* - \log \frac{n_{ij}}{n_i} + \frac{1}{2n_{ij}} - \frac{1}{n_i} \\
&< \log a_{ij}^* - \Psi(n_{ij}) + \Psi(n_i) \\
&< \log a_{ij}^* - \log \frac{n_{ij}}{n_i} - \frac{1}{2n_{ij}} + \frac{1}{n_i}, \qquad (5.9)
\end{aligned}$$

where we use an inequality,

$$\frac{1}{2x} < \log x - \Psi(x) < \frac{1}{x}. \qquad (5.10)$$

We consider two conditions which attain the lower bound.

1) Eliminate redundant states

$$\begin{aligned}
&n_{ij} = n_i a_{ij}^* \ (1 \le i \le K_0, 1 \le j \le K_0) \\
&n_{im} = n_i b_{im}^* \ (1 \le i \le K_0, 1 \le m \le M) \\
&n_{ij} = n_i = 0 \ (K_0 + 1 \le i, j \le K) \\
&n_{im} = n_i = 0 \ (K_0 + 1 \le i \le K, 1 \le m \le M)
\end{aligned}$$

where each n_i $(1 \leq i \leq K_0)$ is $O_p(n)$.

2) Using all states

We set down the redundant states to the K_0th state. Let each n_i is $O_p(n)$ and is constrained by $\sum_{i=1}^{K} n_i = n$.

$$
\begin{aligned}
n_{ij} &= n_i a^*_{ij} \ (1 \leq i \leq K_0 - 1, 1 \leq j \leq K_0 - 1) \\
n_{ij} &= \frac{1}{1 + (K - K_0)} n_i a^*_{ik_0} \ (1 \leq i \leq K_0 - 1, k_0 \leq j \leq K) \\
n_{ij} &= n_i a^*_{k_0 j} \ (K_0 \leq i \leq K, 1 \leq j \leq K_0 - 1) \\
n_{ij} &= \frac{1}{1 + (K - K_0)} n_i a^*_{k_0 k_0} \ (k_0 \leq i \leq K, k_0 \leq j \leq K) \\
n_{im} &= n_i b^*_{im} \ (1 \leq i \leq K_0 - 1, 1 \leq m \leq M) \\
n_{im} &= \frac{1}{1 + (K - K_0)} n_i b^*_{im} \ (K_0 \leq i \leq K, 1 \leq m \leq M)
\end{aligned}
$$

When we set n_{ij}, n_{im} satisfy above conditions, from inequality (5.9), we can evaluate

$$\log a^*_{ij} - \Psi(n_{ij}) + \Psi(n_i) \to O_p(\frac{1}{n}) \ (n \to \infty).$$

Using these inequality for all parameters and equation (5.8), we conclude that the first and the second term of $\bar{F}(X^n)$ satisfy $O_p(1)$ $(n \to \infty)$. □

5.4 SCFG case

The true SCFG has K_0 non-terminal symbols and the model parameter is given by the constant $\{A^*, B^*\}$ as follows.

$$
\begin{aligned}
p(x|\theta^*) &= \sum_{t \in T} p(x, t|\theta^*) \\
\theta^* &= \{A^*, B^*\}, \\
A^* &= \{a^{i*}_{jk} \ (1 \leq i, j, k \leq K_0)\}, \sum_{j,k=1}^{K_0} a^{i*}_{jk} = 1 \\
B^* &= \{b^*_{im} \ (1 \leq i \leq K_0, 1 \leq m \leq M)\}, \sum_{m=1}^{M} b^*_{im} = 1.
\end{aligned}
$$

Theorem 2. *[17] Under the conditions (A1) to (A4), the normalized variational free energy of SCFG satisfies*

$$\bar{F}(X^n) = \bar{\lambda}\log n + O_p(1) \quad (n \to \infty)$$

as the number of sample n goes to infinity, where the learning coefficient of SCFG $\bar{\lambda}$ *is given by*

$$\bar{\lambda} = \begin{cases} \frac{1}{2}\{K_0(K_0^2-1) + K_0(M-1)\} + K_0(K^2-K_0^2)\phi_0 \\ (\phi_0 \leq \frac{K_0^2+KK_0+K^2+M-2}{2(K_0^2+KK_0)}) \\ \frac{1}{2}\{K(K^2-1) + K(M-1)\} \\ (\phi_0 > \frac{K_0^2+KK_0+K^2+M-2}{2(K_0^2+KK_0)}). \end{cases}$$

Proof. Similar to the HMM case, we first consider the lower bound. The second term of equation (5.1) is Kullback information from the variational posterior to the prior. In SCFG, this term is Kullback information of Dirichlet distribution which is written by the expected sufficient statistics n_i, n_{jk}^i, n_{im}.

$$\begin{aligned}
&\int r(\theta|X^n)\log\frac{r(\theta|X^n)}{\varphi(\theta)}d\theta \\
&= KL(r(\theta|X^n)||\varphi(\theta)) = KL(r(A|X^n)||\varphi(A)) + KL(r(B|X^n||\varphi(B)) \\
&= \sum_{i=1}^{K}\{\log\Gamma(n_i+K^2\phi_0) - n_i\Psi(n_i+K^2\phi_0) - \log\Gamma(K^2\phi_0) + K^2\log\Gamma(\phi_0) \\
&- \sum_{j=1}^{K}\sum_{k=1}^{K}\{\log\Gamma(n_{jk}^i+\phi_0) - n_{jk}^i\Psi(n_{jk}^i+\phi_0)\}\} \\
&+ \sum_{i=1}^{K}\{\log\Gamma(n_i+M\xi_0) - n_i\Psi(n_i+M\xi_0) - \log\Gamma(M\xi_0) + M\log\Gamma(\xi_0) \\
&- \sum_{i=1}^{M}\{\log\Gamma(n_{im}+\xi_0) - n_{im}\Psi(n_{im}+\xi_0)\}\}
\end{aligned}$$

Using equation (5.4), the Kullback information is asymptotically expanded

as,

$$
\begin{aligned}
&KL(r(\theta|X^n)||\varphi(\theta)) \\
&= \sum_{i=1}^{K}\{(K^2\phi_0 - \frac{1}{2})\log(n_i + K^2\phi_0) - \sum_{j=1}^{K}\sum_{k=1}^{K}\{(\phi_0 - \frac{1}{2})\log(n^i_{jk} + \phi_0)\} \\
&+ (M\xi_0 - \frac{1}{2})\log(n_j + M\xi_0) - \sum_{m=1}^{M}\{(\xi_0 - \frac{1}{2})\log(n_{jm} + \xi_0)\}\} + O_p(1).
\end{aligned} \tag{5.11}
$$

Consider the minimization of above equation with respect to n_i, n^i_{jk}, n_{im}. We divide the problem into two cases $\phi_0 \leq \frac{1}{2}$ and $\phi_0 > \frac{1}{2}$.

In case $\phi_0 \leq \frac{1}{2}$, it it obvious that the minimization (5.11) is achieved by the redundant expected sufficient statistics n_i, n^i_{jk}, n_{im} $(K_0 < i, j \leq K)$ having $O_p(1)$.

In case $\phi_0 > \frac{1}{2}$, from the equation (5.11) and the definition of n_i, minimization is achieved by the case that n^i_{jk}, n_{im} have the same order of n_i. Therefore, let $n_i = p_i n^{\alpha_i}, n^i_{jk} = q^i_{jk} n^{\alpha_i}, n_{im} = r_{im} n^{\alpha_i}$ $(p_i, q^i_{jk}, r_{im} > 0, 0 \leq \alpha_i \leq 1)$. Then,

$$
\begin{aligned}
&KL(r(\theta|X^n)) \\
&= \sum_{i=1}^{K}\alpha_i\{(K^2\phi_0 + \frac{M}{2} - 1) - \sum_{j,k=1,\alpha_i\neq 0}^{K}(\phi_0 - \frac{1}{2})\}\log n + O_p(1)
\end{aligned}
$$

In minimizing above equation, by assumption (A4), $\alpha_i = 1$ $(1 \leq i \leq K_0)$ holds. Moreover, because of the α_i satisfies $\alpha_i \geq 0$ and linearity to the objective function, we divide the function to true K_0 terms and additional l $(0 \leq l \leq K - K_0)$ terms that has $\alpha_i = 1$.

$$
\begin{aligned}
&\frac{KL(r(\theta|X^n)||\varphi(\theta))}{\log n} \\
&= (K_0^2\phi_0 + \frac{M}{2} - 1)K_0 - (\phi_0 - \frac{1}{2})K_0^3 \\
&+ (K^2\phi_0 + \frac{M}{2} - 1)l - (\phi_0 - \frac{1}{2})((K_0 + l)^3 - K_0^3) + O_p(1).
\end{aligned} \tag{5.12}
$$

Minimization of the second term with respect to l has two cases which is determined by the hyperparameter ϕ_0 and given by

$$\begin{cases} l = 0 \ (\phi_0 \leq \frac{K_0^2+KK_0+K^2+M-2}{2(K_0^2+KK_0)}) \\ l = K - K_0 \ (\phi_0 > \frac{K_0^2+KK_0+K^2+M-2}{2(K_0^2+KK_0)}). \end{cases}$$

Finally, substituting l into equation (5.12), we get the result.

Next, we consider the upper bound. From the discussion of the lower bound, it is only necessary finding the concrete parameters which satisfies lower bound constraints and the first and second term of $\bar{F}(X^n)$ (5.1) becomes $O_p(1)$. Substituting the concrete form of SCFG to equation (5.7).

$$\log \sum_{Y^n} \exp(\sum_{t=2}^{T}\sum_{i=1}^{K}\sum_{j=1}^{K}\sum_{k=1}^{K} y_{t,jk}^{i}\{\log(a_{jk}^{i})^{*} - \Psi(n_{jk}^{i}+\phi_0) + \Psi(n_i + K\phi_0)\}$$
$$+ \sum_{t=1}^{T}\sum_{i=1}^{K}\sum_{m=1}^{M} \hat{y}_{t,i}x_{i,m}\{\log b_{im}^{*} - \Psi(n_{im}+\xi_0) + \Psi(n_i + M\xi_0)\}). \tag{5.13}$$

We evaluate each parameter separately. For example

$$\log {a_{ij}^{i}}^{*} - \Psi(n_{jk}^{i}) + \Psi(n_i)$$

satisfy

$$\begin{aligned} &\log {a_{jk}^{i}}^{*} - \log \frac{n_{jk}^{i}}{n_i} + \frac{1}{2n_{jk}^{i}} - \frac{1}{n_i} \\ &< \log {a_{jk}^{i}}^{*} - \Psi(n_{jk}^{i}) + \Psi(n_i) \\ &< \log {a_{jk}^{i}}^{*} - \log \frac{n_{jk}^{i}}{n_i} - \frac{1}{2n_{ij}} + \frac{1}{n_i}, \end{aligned} \tag{5.14}$$

where we use equation (5.10).

We consider the two conditions which give the lower bound.

1) Eliminate redundant components

$$
\begin{aligned}
&n^i_{jk} = n_i a^*_{ij} \ (1 \le i, j, k \le K_0) \\
&n_{im} = n_i b^*_{im} \ (1 \le i \le K_0, 1 \le m \le M) \\
&n^i_{jk} = n_i = 0 \ (K_0 + 1 \le i, j, k \le K) \\
&n_{im} = n_i = 0 \ (K_0 + 1 \le i \le K, 1 \le m \le M)
\end{aligned}
$$

where each n_i $(1 \le i \le K_0)$ is $O_p(n)$.

2) Using all parameters

We set down the redundant states to the K_0th state. Let each n_i is $O_p(n)$ and has constraint $\sum_{i=1}^{K} n_i = n$.

$$
\begin{aligned}
&n^i_{jk} = n_i {a^i_{jk}}^* \ (1 \le i, j, k \le K_0 - 1) \\
&n^i_{jk} = \frac{1}{1 + (K^2 - K_0^2)} n_i a^{K_0 *}_{ij} \ (1 \le i \le K_0 - 1, K_0 \le j, k \le K) \\
&n^i_{jk} = n_i a^{K_0 *}_{jk} \ (K_0 \le i \le K, 1 \le j, k \le K_0 - 1) \\
&n^i_{jk} = \frac{1}{1 + (K^2 - K_0^2)} n_i {a^{K_0}_{K_0 K_0}}^* \ (K_0 \le i, j, k \le K) \\
&n_{im} = n_i b^*_{im} \ (1 \le i \le K_0 - 1, 1 \le m \le M) \\
&n_{im} = \frac{1}{1 + (K - K_0)} n_i b^*_{im} \ (K_0 \le i \le K, 1 \le m \le M)
\end{aligned}
$$

When we set n^i_{jk}, n_{im} satisfy above conditions, from inequality (5.14), we can evaluate

$$
\log {a^i_{jk}}^* - \Psi(n^i_{jk}) + \Psi(n_i) \to O_p(\frac{1}{n}) \ (n \to \infty).
$$

Using these inequalities for all parameters and equation (5.13), we conclude that the first and the second term of $\bar{F}(X^n)$ satisfies $O_p(1)$ $(n \to \infty)$. □

Chapter 6

Model Selection

In this chapter, we propose a new model selection method for stochastic grammar based on the main theorem.

6.1 Selection Criteria

In the theorem 1 and theorem 2, we can see that the asymptotic form of the normalized free energy is the function of the true number of non-terminals. Therefore, we can estimate the number of non-terminals by inverting the function with respect to K_0 [34]. It is noted that singularities of non-identifiable models produce this effect. On the contrary, in identifiable models, from the positive definiteness of the Fisher information matrix, the number of model parameters completely determines the entropy of the posterior distributions. Moreover, variational Bayesian approach has two advantages against the previous proposed method which uses true Bayes free energy [34]. The one is, as shown by the equation (5.1), the first term of the variational free energy is estimated entropy of the true distribution. Therefore, we directory estimate the coefficient $\bar{\lambda}$ by Kullback information from the variational posterior to the prior. The other is the computational demand of the optimization of the variational free energy is much smaller than estimation of the free energy by MCMC which is used in [34].

We consider the case of hyperparameter $\phi_0 = \frac{1}{2}$. This prior is known as Krichevsky-Trofimov mixture in information theory [12].

Corollary 1. *In HMM, we assume the condition (A1) to (A4). Then, the following criterion converges to the number of non-terminal K_0 as number of samples goes to infinity.*

$$\frac{KL(r(\theta||X^n)||\varphi(\theta))}{\frac{1}{2}(K+M-2)\log n} \to K_0 \quad (n \to \infty). \tag{6.1}$$

Corollary 2. *In SCFG, we assume the condition (A1) to (A4). Then, the following criterion converges to the number of non-terminal K_0 as number of samples goes to infinity.*

$$\frac{KL(r(\theta||X^n)||\varphi(\theta))}{\frac{1}{2}(K^2+M-2)\log n} \to K_0 \quad (n \to \infty).$$

6.2 Numerical Experiment

We evaluate the proposed method by numerical experiment in HMM case. For the index of the correctness, the judge of the number of the states is nearest integer of the equation (6.1).

In numerical experiment, we verify two effects, the one is the number of examples and the other is the redundancy of the learning machine.

We assume the true HMM has two states and two symbols and the parameter of distribution is set as follows,

$$A = \begin{pmatrix} 0.7 & 0.3 \\ 0.2 & 0.8 \end{pmatrix}, \quad B = \begin{pmatrix} 0.7 & 0.3 \\ 0.2 & 0.8 \end{pmatrix}.$$

The number of the model parameter is 4 in this model. The length of each sequence is 200 ($L = 200$) and we repeat the experiment over 200 sample sets. We set the hyperparameter of the transition probability $\xi_0 = \frac{1}{2}$.

6.2.1 Refinement criteria for HMM

The criteria (6.1) converges to the number of true non-terminals as the number of samples goes to infinity. Although, in finite sample case, we can refine this criteria by considering the constant terms. From the equation (5.5), we pick up the terms which the expectation sufficient statistics n_i, n_{ij} satisfy $O_p(n)$. Then, the Kullback information from the posterior to the prior of the transition probability A is written as,

$$\begin{aligned}
&K(r(A)||\varphi(A)) = \\
&\sum_{i=1}^{K_0} \log\Gamma(n_i + \frac{K}{2}) - K_0 \log\Gamma(\frac{K}{2}) \\
&- \sum_{i=1}^{K_0}\sum_{j=1}^{K_0} \log\Gamma(n_{ij} + \frac{1}{2}) + K_0^2 \log\Gamma(\frac{1}{2}) \\
&+ \sum_{i=1}^{K_0}\sum_{j=1}^{K_0} n_{ij}\Psi(n_{ij} + \frac{1}{2}) - \sum_{j=1}^{K_0}\sum_{i=1}^{K_0} n_{ij}\Psi(n_i + \frac{K}{2}) + O_p(1).
\end{aligned}$$

As the same argument of the main theorem, we have the asymptotic expansion of this equation using the Stirling formula (5.4).

$$\begin{aligned}
K(r(A)||\varphi(A)) &= K_0(\frac{K}{2} - \frac{1}{2})\log n - K_0\frac{K}{2} + \frac{K_0}{2}\log(2\pi) \\
&- K_0 \log\Gamma(\frac{K}{2}) + K_0^2\frac{1}{2} - \frac{K_0^2}{2}\log(2\pi) + K_0^2\log\Gamma(\frac{1}{2}) + O_p(1) \qquad (6.2)
\end{aligned}$$

We solve the quadratic equation (6.2) with respect to K_0, and use this criteria for the following experiments.

6.2.2 Experiment1

For testing the effect of the number of examples, we evaluate the case of learning the true HMM by four states learning machine and the numbers of examples are 200, 2000, 20000.

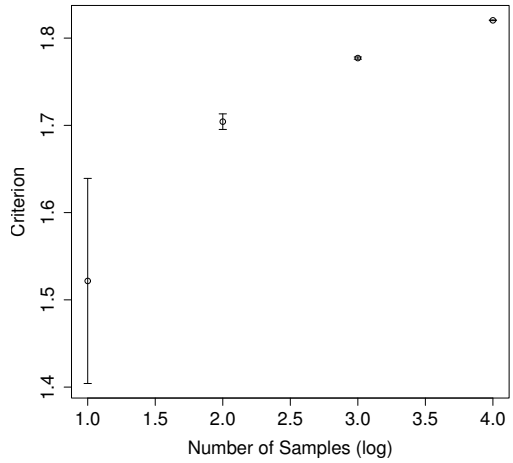

Figure 6.1: Result of experiment1, the horizontal axis is the number of samples ($200 \times \log_{10} x$) and the vertical axis is the estimated component. The true value is 2.0. The mean and the standard deviation are plotted.

The result of the experiment is displayed by figure (6.1). From the figure, we can see that, as increasing the number of the example, the mean of the estimator approaches to the true value from the beneath and the variance of the estimator is rapidly decreasing. In 200 samples case, the correctness of the estimation is 82 percent and all fails are occurred at the under estimation. More than 2000 samples case, the correctness of the estimation is 100 percent.

6.2.3 Experiment2

For testing the effect of redundancy of the learning machine, we evaluate the case of learning the true HMM by 2, 3, 4, 5 and 10 states learning machine and the number of examples is fixed at 2000.

The experimental result is displayed by figure (6.2). These results show that the increasing redundant states do not affect the estimator. The mean

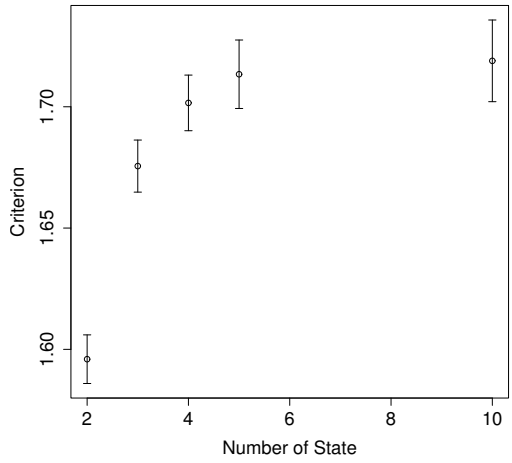

Figure 6.2: Result of experiment2, the horizontal axis is states of the learning machine and the vertical axis is the estimated components. The true value is 2.0. The mean and the standard deviation are plotted.

of the estimator is increasing and close to the true value and the variance of the estimator is slightly increased. The correctness of the estimation is 100 percent in the all learning machines.

Chapter 7

Discussion

In this chapter, we discuss the obtained result and give the suggestions for the future works.

7.1 Behavior of Variational Free Energy

7.1.1 Schematic Behavior of Variational Free Energy

From theorem 1 and theorem 2, the asymptotic variational free energy has two cases which is determined by the hyperparmeter of the non-terminals ϕ_0. In case of the large ϕ_0, the learning coefficient $\bar{\lambda}$ coincide with the half of the model parameters and equivalent to BIC [30]. In case of the small ϕ_0, the minimum of the free energy satisfies eliminates redundant non-terminals and much smaller than the model parameters (figure7.1). In case of eliminating the redundant non-terminals, the variational Bayes learning satisfies the small number of estimated parameters than the maximum likelihood method. This effect avoids the overfitting.

7.1.2 Effect of HyperParameter

The rough explanation of the effect of the hyperparameter is given by the weight of the prior to the singularities. The figure (7.2, 7.3, 7.4) shows

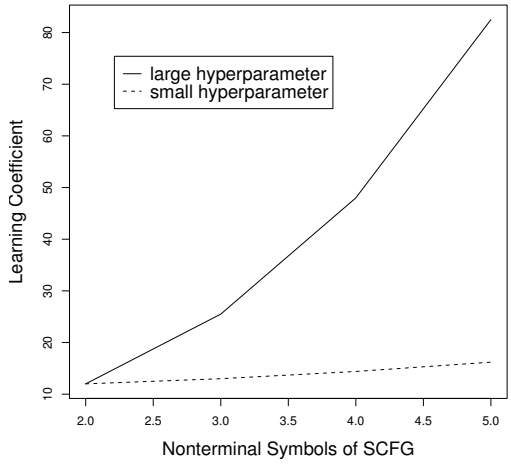

Figure 7.1: Schematic diagram of the variational free energy. In case of the true SCFG has 2 non-terminals and 10 terminals. The horizontal axis is the number of non-terminals of the learning machine and the vertical axis is the learning coefficient $\bar{\lambda}$. The solid line is large ϕ_0 and coincides with BIC. The dotted line is $\phi_0 = 0.1$.

the density of the prior distribution across the hyperparameter ϕ_0 when the model has 2 non-terminal symbols. These figure show that in case of the small hyperparameter ϕ_0, the density of the distribution at the singularity (corresponds at 0.0) diverges to infinity. This prior strongly enhances eliminating the redundant non-terminals. At the hyperparameter $\phi_0 = 1.0$, the density is the uniform and as increasing ϕ_0 the weight of the singularity is decreasing to 0. The figure (7.5) shows the learning coefficient $\bar{\lambda}$ by the hyperparameter ϕ_0. In case of limit of ϕ_0 to 0, the learning coefficient is the half of the number of parameters of the true distribution. The learning coefficient increases with the hyperparameter ϕ_0 and saturated at the point where the coefficient is the half of the number of the parameters of the learning machine.

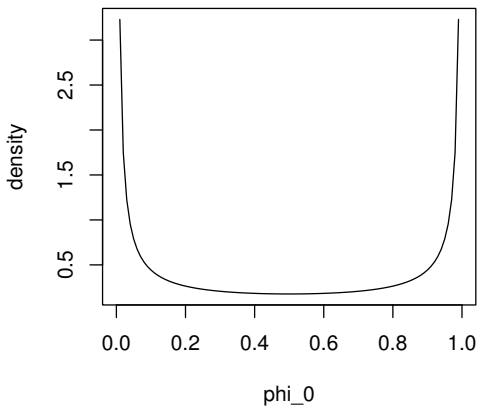

Figure 7.2: Density of prior for $\phi_0 = 0.1$

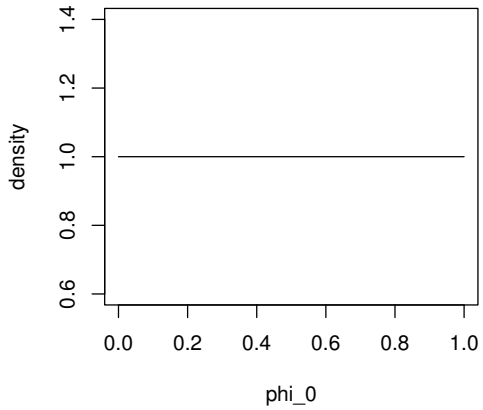

Figure 7.3: Density of prior for $\phi_0 = 1.0$

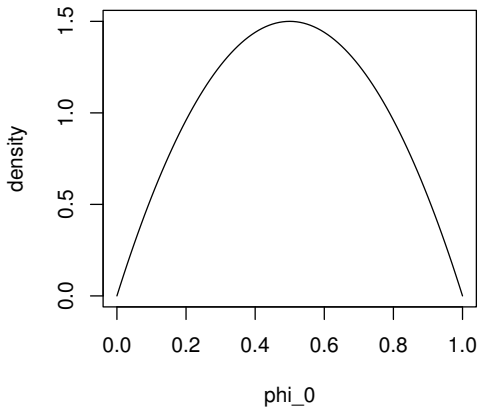

Figure 7.4: Density of prior for $\phi_0 = 2.0$

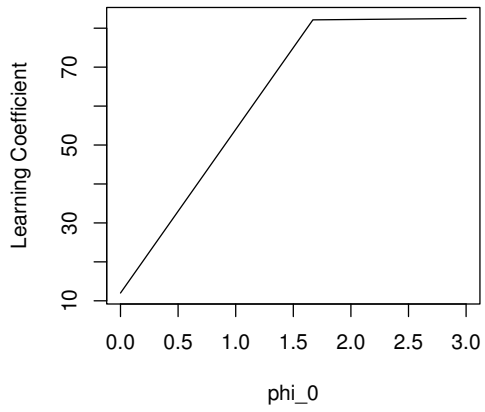

Figure 7.5: Effect for the Learning coefficient by hyperparameter

7.1.3 Approximation Accuracy

To evaluate approximation accuracy, we compare the variational free energy to the Bayes free energy. Compare the case of the model has uniform prior distribution ($\phi_0 = 1.0$). Then, we can see that the variational free energy is always greater than the Bayes upper bound. Hence, the variational Bayes posterior does not coincide with the Bayes one even asymptotically. However, in HMM and SCFG, only upper bound of the free energy is obtained [35],[34]. The elucidation of the Bayes free energy and characterizing the variational approximation by comparing among them are future works.

7.2 Model Selection on Variational Bayes

7.2.1 Advantage of the proposed method

Against the conventional method like BIC, MDL, AIC, our proposed model selection at the chapter 6 appreciates the non-identifiability theoretically.

Additionally, the proposed estimator is not a criterion for comparison but the direct indicator of the number of non-terminals. Therefore we can escape the procedure that calculates all candidate models and compares the results. The proposed method only needs that the learning machine is large enough for including the true model. However, if the redundancy has strong effect to the estimator variance, the choice of the learning machine becomes delicate. The numerical experiment2 suggests that the redundancy of the learning machine gives small impact on the estimator. The clarification of the variance of variational free energy is a future work.

7.2.2 Number of Samples

In experiment1, the needs of samples which ensures the correctness of the estimator is 500 times of a number of parameters. However, the result shows that the estimator approaches to the true value from beneath. If we assure this property, the judge is done by the ceil of the estimator and more accurate estimation will be obtained. At the same time, the estimator represents the actual number of the non-terminals for the resolution of the given samples. Therefore, the small estimation indicates that the variational Bayes estimator balances the bias and the variance. This effect avoids the overfitting at the case of small samples.

7.3 Evaluation of the Optimization Algorithm

It is conjectured that variational optimization algorithm in stochastic grammar has a large number of local minima. In this thesis, we obtained the theoretical value of the objective function. Therefore several initialize method or heuristics model search procedure such as split and merge algorithm [31] can be evaluated to the optimum variational free energy.

Chapter 8

Conclusion

In this thesis, we establish the theory of stochastic grammatical inference of variational Bayesian learning. The main contributions of this thesis are summarized as follows,

1. For HMM and SCFG, the asymptotic variational Free Energy which is objective function of variational Bayesian learning and fundamental quantity for the model selection was obtained.

2. Based on the analysis, a new model selection criteria for HMM and SCFG was proposed. This criteria is efficiently computable and does not suffer from the combinatorial problem.

3. Numerical experiments of the model selection of HMM were performed. The result indicated the effectiveness of the proposed method.

These result will be the step for developing the theory and the practical algorithm for stochastic grammatical inference.

Bibliography

[1] H. Akaike. A new look at statistical model identification. *IEEE Tans. Automatic Control*, 19(6):716–723, 1974.

[2] M. Aoyagi and S. Watanabe. Stochastic complexities of reduced rand regression in bayesian estimation. *Neural Networks*, 18(7), 2005.

[3] H. Attias. Inferring parameters and structure of latent variable models by variational bayes. In *Proceeding of 15th Conference on Uncertainty in Artificial Intelligence*, pages 21–30, 1999.

[4] P. Baldi and S. Brunak. *Bioinformatics. The machine Learning Approach.* The MIT Press, 1998.

[5] L. E. Baum and T. Petrie. Statistical inference for probabilistic function of finite state markov chains. *Annals of Mathematical Statistics*, 37(6):1554–1563, 1966.

[6] M. J. Beal. Variational algorithms for approximate bayesian inference. *Ph.D. Thesis*, 2003. University College London.

[7] N. Chomsky. Three models for the description of language. *IRE Transactions on Information Theory*, 2(2):113–123, 1956.

[8] A. P. Dempster, N. M. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39-B:1–38, 1977.

[9] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis*. Cambridge University Press, 1998.

[10] B. Efron and C. Morris. Stein's estimation rule and its competitors-an empirical bayes approach. *Journal of the American Statistical Association*, 68:117–130, 1973.

[11] R. P. Feynman. *Statistical Mechanics*. A Set of Lectures. W. A. Benjamin, Inc, 1972.

[12] E. Gassiat and S. Boucheron. Optimal error exponents in hidden markov models order estimation. *IEEE Transactions on Information Theory*, 49(4):964–980, 2003.

[13] W. R. Gilks, S. Richardson, and D. J. Speigelhalter, editors. *Markov Chain Monte Carlo in Practice*. Chapman and Hall, 1996.

[14] I. J. Good. *The Estimation of Probabilities*. MIT Press, 1965.

[15] J. E. Hopcroft, R. Motwani, and J. D. Ullman. *Introduction to Automata Theory, Languages, And Computation*. Addison-Wesley, 3rd edition, 2006.

[16] T. Hosino, K. Watanabe, and S. Watanabe. Stochastic complexity of variational bayesian hidden markov models. In *International Joint Conference of Neural Networks 2005*, pages 1114–1119, 2005.

[17] T. Hosino, K. Watanabe, and S. Watanabe. Free energy of stochastic context free grammar on variational bayes. In *International Conference on Neural Information Processing*, pages 407–416, 2006.

[18] T. Hosino, K. Watanabe, and S. Watanabe. Stochastic complexity of hidden markov models on the variational bayesian learning (in japanese. *IEICE Trans.*, J89-D(6):1279–1278, 2006.

[19] H. Ito, S. Amari, and K. Kobayashi. Identifiability of hidden markov information sources and their minimum degrees of freedom. *IEEE Transactions on Information Theory*, 38(2):324–333, 1992.

[20] K. Kurihara and T. Sato. An application of the variational bayesian approach to probablistic context-free grammars. In *IJCNLP-04 Workshop Beyond shallow analysis*, 2004.

[21] K. Lari and S. Yound. The estimation of stochastic context-free grammars using the inside-outside algorithm. *Computer Speech and Language*, 4:35–56, 1990.

[22] E. Levin, N. Tishby, and S. A. Solla. A statistical approaches to learning and generalization in layered neural networks. *Proc. of IEEE*, 78(10):1568–1674, 1990.

[23] D. J. C. Mackay. Ensemble learning for hidden markov models. *Technical report*, 1997. University of Cambridge.

[24] D. J. C. Mackay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[25] S. Nakajima. Asymptotic theory of empirical and variational bayes learning. *Ph.D. Thesis*, 2006. Tokyo Institute of Technology.

[26] N. Nakano, K. Takahashi, and S. Watanabe. On the evaluation criterion of the mcmc method in singular learning machines (in japanese). *IEICE Trans.*, J88-D2:2011–2020, 10 2005.

[27] L. Rabiner and B. H. Juang. *Fundamentals of Speech Recognition*. Pearson Education, 1993.

[28] J. Rissanen. Stochastic complexity and modeling. *Annals of Statistics*, 14:1080–1100, 1986.

[29] M. Sato. Online model selection based on the variational bayes. *Neural Computations*, 13(7):1649–1681, 2001.

[30] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.

[31] N. Ueda and Z. Ghahramani. Bayesian model search for mixture models based on optimizing variational bounds. *International Journal of Neural Networks*, 15:1223–1241, 2002.

[32] K. Watanabe and S. Watanabe. Lower bounds of stochastic complexities in variational bayes learning of gaussian mixture models. In *Proceedings of IEEE conference on Cybernetics and Intelligent Systems*, pages 99–104, 2004.

[33] S. Watanabe. Algebraic analysis for non-identifiable learning machines. *Neural Computation*, 13(4):899–933, 2001.

[34] K. Yamazaki, K. Nagata, and S. Watanabe. A new method of model selection based on learning coefficient. In *Proceedings of International Symposium on Nonlinear Theory and its Application*, pages 389–392, 2005.

[35] K. Yamazaki and S. Watanabe. Stochastic complexities of hidden markov models. In *IEEE International Workshop On Neural Networks For Signal Processing*, 2003.