社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

混合指数型分布の変分ベイズ学習における確率的複雑さ

渡辺　一帆†　　渡辺　澄夫††

† 東京工業大学大学院総合理工学研究科知能システム科学専攻　〒 226–8503 横浜市緑区長津田 4259
†† 東京工業大学精密工学研究所　〒 226–8503 横浜市緑区長津田 4259 MailBox:R2-5
E-mail: †{kazuho23,swatanab}@pi.titech.ac.jp

あらまし　　変分ベイズ法はベイズ学習の一つの近似手法として提案され、少ない計算量と、その有効性が多くの実問題を通じて検証されてきた。しかしながら、近似精度や汎化性能といった変分ベイズ法自体の理論的性質は未だに不明な部分が多い。　本研究では、指数型分布族の混合モデルの変分ベイズ学習について考察し、確率的複雑さの上界と下界を与える。混合分布モデルの真のベイズ法における確率的複雑さと比較することで、変分ベイズ法の近似法としての精度について議論する。

キーワード　混合分布、変分ベイズ学習、確率的複雑さ、特異モデル

Stochastic Complexities for Mixture of Exponential Family in Variational Bayes Approach

Kazuho WATANABE† and Sumio WATANABE††

† Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology　4259 Nagatsuta, Midori-ku, Yokohama, 226–8503 Japan
†† P&I Lab, Tokyo Institute of Technology　4259 Nagatsuta, Midori-ku, Yokohama, 226–8503 Japan
E-mail: †{kazuho23,swatanab}@pi.titech.ac.jp

Abstract　The Variational Bayes learning, proposed as an approximation of the Bayesian learning, has provided computational tractability and good generalization performance in many applications. However, little has been done to investigate its theoretical properties. In this paper, we discuss the Variational Bayes learning of the mixture of exponential families and derive the upper and lower bounds of the stochastic complexities or the marginal likelihoods. We show that the stochastic complexities become smaller than those of regular statistical models, which means the advantage of the Bayesian learning still remains in the Variational Bayes learning.

Key words　Mixture Model, Variational Bayes Learning, Stochastic Complexity, Singular Statistical Model

1. Introduction

The Variational Bayes (VB) framework was proposed as an approximation of the Bayesian learning in the models with hidden variables [3] [6]. This framework provides computationally tractable posterior distributions over the hidden variables and parameters with an iterative algorithm. The Variational Bayes learning has been applied to various learning machines and it has performed good generalization with only modest computational costs compared to Markov chain Monte Carlo (MCMC) methods that are the major schemes of the Bayesian learning.

In spite of its tractability and its wide range of applications, little has been done to investigate the theoretical properties of the Variational Bayes learning itself. Although the Variational Bayes framework is an approximation, questions like how accurately it approximates the true one remained unanswered until quite recently. To address these issues, the stochastic complexity in the Variational Bayes learning of gaussian mixture models was clarified and the accuracy of the Variational Bayes learning was discussed [15].

In this paper, we focus on the Variational Bayes learning of more general mixture models, namely the mixtures of exponential families which include mixtures of distributions such as gaussian, binomial and gamma. Mixture models are known to be non-regular statistical models since they have

the non-identifiability of parameters caused by their hidden variables. In some recent studies, it has been proven that the Bayesian stochastic complexities of non-regular models become smaller than those of regular models [11] [12] [13]. This indicates an advantage of the Bayesian learning when it is applied to non-regular models.

In this paper, we derive the upper and lower bounds of the stochastic complexity in the Variational Bayes learning of the mixture of exponential families and show that the stochastic complexity becomes smaller than those of regular models, which means the advantage of the Bayesian learning still remains in the Variational Bayes learning. In addition, the derived bounds give us an indication of how the hyperparameter in the prior distribution influences the process of the learning. We consider the case in which the true distribution is contained in the learner model. Analyzing the stochastic complexity in the case is most valuable for comparing the Variational Bayes learning with the true Bayesian learning. This is because the advantage of the Bayesian learning is typical in the case [11]. Furthermore, this analysis is necessary and essential for addressing the model selection problem and hypothesis testing.

2. Mixture of Exponential Family

Denote by $c(x|b)$ a density function of the input $x \in R^N$ given an M-dimensional parameter vector $b = (b^{(1)}, b^{(2)}, \cdots, b^{(M)})^T \in B$ where B is a subset of R^M. The general mixture model $p(x|\theta)$ with a parameter vector θ is defined by

$$p(x|\theta) = \sum_{k=1}^{K} a_k c(x|b_k),$$

where integer K is the number of components and $\{a_k | a_k \geqq 0, \sum_{k=1}^{K} a_k = 1\}$ is the set of mixing proportions. The parameter θ of the model is $\theta = \{a_k, b_k\}_{k=1}^{K}$.

A mixture model is called a mixture of exponential family (MEF) model or exponential family mixture model if the probability distribution $c(x|b)$ for each component is given by the following form,

$$c(x|b) = \exp\{b \cdot f(x) + f_0(x) - g(b)\}, \tag{1}$$

where $b \in B$ is called the natural parameter, $b \cdot f(x)$ is the inner product with the vector $f(x)$ which is defined by using M functions $f_1, \cdots, f_M$ as $f(x) = (f_1(x), \cdots, f_M(x))^T$, $f_0(x)$ and $g(b)$ are real-valued functions of the input x and the parameter b, respectively.

The conjugate prior distribution $\varphi(\theta)$ for the mixture of exponential family model is given by the product of the following two distributions on $\mathbf{a} = \{a_k\}_{k=1}^{K}$ and $\mathbf{b} = \{b_k\}_{k=1}^{K}$

$$\varphi(\mathbf{a}) = \frac{\Gamma(K\phi_0)}{\Gamma(\phi_0)^k} \prod_{k=1}^{K} a_k^{\phi_0 - 1}, \tag{2}$$

$$\varphi(\mathbf{b}) = \prod_{k=1}^{K} \varphi(b_k) = \prod_{k=1}^{K} \frac{1}{C(\xi_0, \nu_0)} \exp\{\xi_0(b_k \cdot \nu_0 - g(b_k))\}, \tag{3}$$

where $\xi_0 > 0$, $\nu_0 \in R^M$ and $\phi_0 > 0$ are constants called hyperparameters and

$$C(\xi, \mu) = \int \exp\{\xi(\mu \cdot b - g(b))\} db \tag{4}$$

is a function of $\xi \in R$ and $\mu \in R^M$.

The mixture model can be rewritten as follows by using a hidden variable $y = (y^1, \cdots, y^K) \in \{(1, 0, \cdots, 0), (0, 1, \cdots, 0), \cdots, (0, 0, \cdots, 1)\}$,

$$p(x, y|\theta) = \prod_{k=1}^{K} \Big[a_k c(x|b_k)\Big]^{y^k}.$$

The hidden variable y is not observed and is representing a component from which the datum x is generated. If the datum x is from the kth component, then $y^k = 1$.

The mixture model is a non-regular statistical model, since it has non-identifiability of the parameter. More specifically, if the true distribution can be realized by a model with the smaller number of components, the true parameter is not a point but an analytic set with singularities. If a model parameter is non-identifiable, the usual asymptotic theory of regular statistical models cannot be applied. Some studies have revealed that the mixture model has quite different properties from those of regular statistical models [7] [12].

3. The Bayesian Learning

Suppose n training samples $X^n = \{x_1, \cdots, x_n\}$ are independently and identically taken from the true distribution $p_0(x)$. In the Bayesian learning of a model $p(x|\theta)$ whose parameter is θ, first, the prior distribution $\varphi(\theta)$ on the parameter θ is set. Then the posterior distribution $p(\theta|X^n)$ is computed from the given dataset and the prior by

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

where $Z(X^n)$ is the normalization constant that is also known as the marginal likelihood or the evidence of the dataset X^n [8]. The Bayesian predictive distribution $p(x|X^n)$ is given by averaging the model over the posterior distribution as follows,

$$p(x|X^n) = \int p(x|\theta) p(\theta|X^n) d\theta. \tag{5}$$

The stochastic complexity $F(X^n)$ is defined by

$$F(X^n) = -\log Z(X^n), \tag{6}$$

which is also called the free energy and is important in most data modeling problems. Practically, it is used as a criterion by which the model is selected and the hyperparameters in the prior are optimized [1] [10].

The Bayesian posterior can be rewritten as

$$p(\theta|X^n) = \frac{1}{Z_0(X^n)} \exp(-nH_n(\theta))\varphi(\theta), \tag{7}$$

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

$$H_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{p_0(x_i)}{p(x_i|\theta)}, \tag{8}$$

and $Z_0(X^n)$ is the normalization constant. Putting $S(X^n) = -\sum_{i=1}^{n} \log p_0(x)$, we define the normalized stochastic complexity $F_0(X^n)$ by

$$F_0(X^n) = F(X^n) - S(X^n). \tag{9}$$

It is noted that the empirical entropy $S(X^n)$ does not depend on the model $p(x|\theta)$ and its expectation value over all sets of training samples is equal to nS where $S = -\int p_0(x)\log p_0(x)dx$ is the entropy. Therefore minimization of $F(X^n)$ is equivalent to that of $F_0(X^n)$. We define the average normalized stochastic complexity $F_0(n)$ by

$$F_0(n) = E_{X^n}\left[F_0(X^n)\right], \tag{10}$$

where $E_{X^n}[\cdot]$ denotes the expectation value over all sets of training samples. And recently, it was proved that $F_0(n)$ has the following asymptotic form [11],

$$F_0(n) \simeq \lambda \log n - (m-1)\log\log n + O(1), \tag{11}$$

where λ and m are the rational number and the natural number respectively which are determined by the singularities of the set of true parameters. In regular statistical models, 2λ is equal to the number of parameters and $m = 1$, whereas in non-regular models such as the mixture model, 2λ is not larger than the number of parameters and $m \geqq 1$. This means an advantage of non-regular models in the case when the Bayesian learning is applied to them.

However, in the Bayesian learning, one computes the stochastic complexity or the predictive distribution by integrating over the posterior distribution, which typically cannot be performed analytically. As an approximation, the Variational Bayes framework was proposed [3] [4] [6]. In the following, we explain the outline of it.

4. The Variational Bayes Framework

Using the likelihood on the complete data $\{X^n, Y^n\}$ added the corresponding hidden variables $Y^n = \{y_1, \cdots, y_n\}$, we can rewrite the stochastic complexity eq.(6) as

$$F(X^n) = -\log \int \sum_{Y^n} p(X^n, Y^n, \theta)d\theta, \tag{12}$$

where the sum over Y^n ranges over all possible values of all hidden variables. In the Variational Bayes framework, the Bayesian posterior $p(Y^n, \theta|X^n)$ of the hidden variables and the parameters is approximated by the variational posterior $q(Y^n, \theta|X^n)$, which factorizes as

$$q(Y^n, \theta|X^n) = Q(Y^n|X^n)r(\theta|X^n), \tag{13}$$

where $Q(Y^n|X^n)$ and $r(\theta|X^n)$ are posteriors on the hidden variables and the parameters respectively. And the variational posterior $q(Y^n, \theta|X^n)$ is chosen to minimize the functional $\overline{F}[q]$ defined by

$$\overline{F}[q] = \sum_{Y^n}\int q(Y^n, \theta|X^n)\log\frac{q(Y^n, \theta|X^n)}{p(X^n, Y^n, \theta)}d\theta, \tag{14}$$

$$= F(X^n) + K(q(Y^n, \theta|X^n)||p(Y^n, \theta|X^n)), \tag{15}$$

where $K(q(Y^n, \theta|X^n)||p(Y^n, \theta|X^n))$ is the Kullback information between the true Bayesian posterior $p(Y^n, \theta|X^n)$ and the variational posterior $q(Y^n, \theta|X^n)$ (1). This leads to the following theorem. The proof is omitted [4] [9].

[Theorem 1] If the functional $\overline{F}[q]$ is minimized under the constraint eq.(13) then the variational posteriors, $r(\theta|X^n)$ and $Q(Y^n|X^n)$, satisfy

$$r(\theta|X^n) = \frac{1}{C_r}\varphi(\theta)\exp\left\langle \log p(X^n, Y^n|\theta)\right\rangle_{Q(Y^n|X^n)}, \tag{16}$$

$$Q(Y^n|X^n) = \frac{1}{C_Q}\exp\left\langle \log p(X^n, Y^n|\theta)\right\rangle_{r(\theta|X^n)}, \tag{17}$$

where C_r and C_Q are the normalization constants (2).

Note that eq.(16) and eq.(17) give only a necessary condition for the functional $\overline{F}[q]$ to be minimized. The variational posteriors that satisfy eq.(16) and eq.(17) are computed by an iterative algorithm whose convergence is guaranteed.

We define the stochastic complexity in the Variational Bayes learning $\overline{F}(X^n)$ by the minimum value of the functional $\overline{F}[q]$, that is ,

$$\overline{F}(X^n) = \min_{r,Q} \overline{F}[q].$$

From eq.(15), the difference between $\overline{F}(X^n)$ and the original stochastic complexity $F(X^n)$ shows the accuracy of the Variational Bayes approach as an approximation of the Bayesian learning. Define the normalized stochastic complexity $\overline{F}_0(X^n)$

$$\overline{F}_0(X^n) = \overline{F}(X^n) - S(X^n). \tag{18}$$

Putting eq.(17) into eq.(14) gives the following lemma. (3).

(1) : $K(q(x)||p(x))$ denotes the Kullback information from a distribution $q(x)$ to a distribution $p(x)$, that is,

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

(2) : $\langle\cdot\rangle_{p(x)}$ denotes the expectation over $p(x)$.

(3) : Hereafter, we omit the condition X^n of the variational posteriors, and abbreviate them to $q(Y^n, \theta)$, $Q(Y^n)$ and $r(\theta)$.

[Lemma 1]

$$\overline{F}_0(X^n) = \min_{r(\theta)}\{K(r(\theta)||\varphi(\theta)) - (\log C_Q + S(X^n)\}, \quad (19)$$

where $C_Q = \sum_{Y^n} \exp\left\langle \log p(X^n, Y^n|\theta)\right\rangle_{r(\theta)}$.

5. Variational Posterior for MEF Model

In this section, we derive the variational posterior $r(\theta|X^n)$ for the mixture of exponential family model based on eq.(16) and then define the variational parameter for this model.

Using the complete data $\{X^n, Y^n\} = \{(x_1, y_1), \cdots, (x_n, y_n)\}$, we put

$$\overline{y}_i^k = \langle y_i^k \rangle_{Q(Y^n)}, \quad n_k = \sum_{i=1}^n \overline{y}_i^k, \quad \text{and} \quad \nu_k = \frac{1}{n_k}\sum_{i=1}^n \overline{y}_i^k f(x_i),$$

where $y_i^k = 1$ if the ith datum x_i is from the kth component, if otherwise, $y_i^k = 0$. The variable n_k is the expected number of the data that are estimated to be from the kth component. From eq.(16) and the respective prior eq.(2) and eq.(3), the variational posterior $r(\theta)$ is obtained as the product of the following two distributions,

$$r(\mathbf{a}) = \frac{\Gamma(n + K\phi_0)}{\prod_{k=1}^K \Gamma(n_k + \phi_0)} \prod_{k=1}^K a_k^{n_k + \phi_0 - 1}, \quad (20)$$

$$r(\mathbf{b}) = \prod_{k=1}^K r(b_k) = \prod_{k=1}^K \frac{1}{C(\gamma_k, \overline{\mu}_k)} \exp\{\gamma_k(\overline{\mu}_k \cdot b_k - g(b_k))\}, \quad (21)$$

where $\overline{\mu}_k = \frac{n_k\nu_k + \xi_0\nu_0}{n_k + \xi_0}$ and $\gamma_k = n_k + \xi_0$. Put

$$\overline{a}_k = \langle a_k \rangle_{r(\mathbf{a})} = \frac{n_k + \phi_0}{n + K\phi_0}, \quad (22)$$

$$\overline{b}_k = \langle b_k \rangle_{r(b_k)} = \frac{1}{\gamma_k}\frac{\partial \log C(\gamma_k, \overline{\mu}_k)}{\partial \overline{\mu}_k}, \quad (23)$$

and define the variational parameter $\overline{\theta}$ by

$$\overline{\theta} = \langle \theta \rangle_{r(\theta)} = \{\overline{a}_k, \overline{b}_k\}_{k=1}^K. \quad (24)$$

Then it is noted that the variational posterior $r(\theta)$ and C_Q in eq.(17) are parameterized by the variational parameter $\overline{\theta}$. Therefore, we denote them as $r(\theta|\overline{\theta})$ and $C_Q(\overline{\theta})$ henceforth. We define the variational estimator $\overline{\theta}_{vb}$ by the variational parameter $\overline{\theta}$ that attains the minimum value of the normalized stochastic complexity $\overline{F}_0(X^n)$. Then, Lemma 1 claims that

$$\overline{\theta}_{vb} = \operatorname*{argmin}_{\overline{\theta}}\{K(r(\theta|\overline{\theta})||\varphi(\theta)) - (\log C_Q(\overline{\theta}) + S(X^n))\}. \quad (25)$$

Therefore, our aim is to evaluate the minimum value of eq.(25) as a function of the variational parameter $\overline{\theta}$.

6. Main Result

The average normalized stochastic complexity $\overline{F}_0(n)$ in the Variational Bayes learning is defined by

$$\overline{F}_0(n) = E_{X^n}[\overline{F}_0(X^n)]. \quad (26)$$

We assume the following conditions.

(i) The true distribution $p_0(x)$ can be represented by an MEF model $p(x|\theta_0)$ which has K_0 components and the parameter $\theta_0 = \{a_k^*, b_k^*\}_{k=1}^{K_0}$,

$$p(x|\theta_0) = \sum_{k=1}^{K_0} a_k^* \exp\{b_k^* \cdot f(x) + f_0(x) - g(b_k^*)\},$$

where $b_k^* \in R^M$ and $b_k^* \neq b_j^* (k \neq j)$. And suppose that the true distribution can be realized by the model, that is, the model $p(x|\theta)$ has K components,

$$p(x|\theta) = \sum_{k=1}^{K} a_k \exp\{b_k \cdot f(x) + f_0(x) - g(b_k)\},$$

and $K \geqq K_0$ holds.

(ii) The prior distribution of the parameters is the conjugate prior $\varphi(\theta) = \varphi(\mathbf{a})\varphi(\mathbf{b})$ given by eq.(2) and eq.(3).

(iii) Regarding the distribution $c(x|b)$ of each component, the Fisher information matrix

$$I(b) = \frac{\partial^2 g(b)}{\partial b \partial b}$$

satisfies $0 < |I(b)| < +\infty$, for arbitrary $b \in B$ [4]. The function $\mu \cdot b - g(b)$ has a stationary point at $\hat{b} \in B$ for each $\mu \in R^M$.

Under these conditions, we prove the following theorem. The proof is done in the next section.

[Theorem 2] **(Main Result)** Assume the conditions (i),(ii) and (iii). Then the average normalized stochastic complexity $\overline{F}_0(n)$ eq.(26) satisfies

$$\overline{\lambda}\log n + E_{X^n}\left[nH_n(\overline{\theta}_{vb})\right] + C_1 \leqq \overline{F}_0(n) \leqq \overline{\lambda}\log n + C_2, \quad (27)$$

as n tends to infinity where C_1, C_2 are constants independent of n and

$$\overline{\lambda} = \begin{cases} (K - K_0)\phi_0 + \frac{MK_0 + K_0 - 1}{2} & (\phi_0 \leqq \frac{M+1}{2}), \\ \frac{MK + K - 1}{2} & (\phi_0 > \frac{M+1}{2}). \end{cases} \quad (28)$$

In this theorem, $H_n(\overline{\theta}_{vb})$ is a training error and is computable during the learning. If the term $E_{X^n}\left[nH_n(\overline{\theta}_{vb})\right]$ is a bounded function of n, then it immediately follows from this theorem that

$$\overline{F}_0(n) = \overline{\lambda}\log n + O(1),$$

where $O(1)$ is a bounded function of n. In certain cases, such as binomial mixtures and mixtures of von-Mises distributions, it is actually a bounded function of n. In the case of gaussian mixtures, it is conjectured that the minus likelihood ratio $\min_\theta nH_n(\theta)$, a lower bound of $nH_n(\overline{\theta}_{vb})$, is at most of the order of $\log\log n$ [7].

(4) : $\frac{\partial^2 g(b)}{\partial b \partial b}$ denotes the matrix whose ijth entry is $\frac{\partial^2 g(b)}{\partial b^{(i)} \partial b^{(j)}}$ and $|\cdot|$ denotes the determinant of a matrix.

Since the dimension of the parameter θ is $MK+K-1$, the average normalized stochastic complexity of regular statistical models, which coincides with the Bayesian information criterion (BIC) [10] is given by $\lambda_{\rm BIC}\log n$ where

$$\lambda_{\rm BIC}=\frac{MK+K-1}{2}. \tag{29}$$

Theorem 2 claims that the coefficient $\overline{\lambda}$ of $\log n$ is smaller than $\lambda_{\rm BIC}$ when $\phi_0 \leqq (M+1)/2$. This means the stochastic complexity $\overline{F}_0(n)$ becomes smaller than the BIC and the advantage of non-regular models in the Bayesian learning still remains in the Variational Bayes learning.

7. Proof of Theorem 2

In this section, we prove Theorem 2. Firstly, we evaluate the term $K(r(\theta|\overline{\theta})||\varphi(\theta))$ in Lemma 1.

From the condition (iii), calculating $C(\gamma_k,\overline{\mu}_k)$ by the saddle point approximation, we obtain the following lemma[(5)].

[Lemma 2]

$$K(r(b_k|\overline{b}_k)||\varphi(b_k))=\frac{M}{2}\log(n_k+\xi_0)-\log\varphi(\overline{b}_k)+O_p(1).$$

Inequalities on the di-gamma function $\Psi(x)$ and the log-gamma function $\log\Gamma(x)$, for $x>0$ and for a positive constant C [2]

$$\frac{1}{2x}<\log x-\Psi(x)<\frac{1}{x}, \tag{30}$$

$$0\leqq\log\Gamma(x)-(x-\frac{1}{2})\log x-x+\frac{1}{2}\log 2\pi\leqq\frac{C}{x}, \tag{31}$$

give the following lemma.

[Lemma 3]

$$K(r(\theta|\overline{\theta})||\varphi(\theta))=G(\overline{\mathbf{a}})-\sum_{k=1}^{K}\log\varphi(\overline{b}_k)+O_p(1) \tag{32}$$

holds where we defined the function $G(\overline{\mathbf{a}})$ of $\overline{\mathbf{a}}=\{\overline{a}_k\}_{k=1}^K$ by

$$G(\overline{\mathbf{a}})=\frac{MK+K-1}{2}\log n+\{\frac{M}{2}-(\phi_0-\frac{1}{2})\}\sum_{k=1}^{K}\log\overline{a}_k. \tag{33}$$

Let us turn now to the term $\log C_Q(\overline{\theta})$ in Lemma 1. It is evaluated by eq.(30) as follows.

[Lemma 4]

$$nH_n(\overline{\theta})+O_p(1)\leqq-(\log C_Q(\overline{\theta})+S(X^n))\leqq n\overline{H}_n(\overline{\theta})+O_p(1)$$

holds where the function $H_n(\theta)$ is defined in eq.(8) and

$$\overline{H}_n(\overline{\theta})=\frac{1}{n}\sum_{i=1}^{n}\log\frac{p(x_i|\theta_0)}{\sum_{k=1}^{K}\overline{a}_k c(x_i|\overline{b}_k)\exp\{-\frac{C'}{n_k+\min\{\phi_0,\xi_0\}}\}},$$

where C' is a constant.

(5) : In this proof, $O_p(1)$ denotes a random variable bounded in probability.

Now from above Lemmas, we show the following theorem on the upper bound in eq.(27).

[Theorem 3] The normalized stochastic complexity $\overline{F}_0(X^n)$ in eq.(18) satisfies the following inequality.

$$\overline{F}_0(X^n)\leqq\overline{\lambda}\log n+O_p(1).$$

(Proof of Theorem 3) From Lemma 1,Lemma 3 and Lemma 4, it follows that

$$\begin{aligned}\overline{F}_0(X^n)&\leqq\min_{\overline{\theta}}\Big[G(\overline{\mathbf{a}})-\sum_{k=1}^{K}\log\varphi(\overline{b}_k)+n\overline{H}_n(\overline{\theta})\Big]+O_p(1).\\&\equiv\min_{\overline{\theta}}[T_n(\overline{\theta})]+O_p(1).\end{aligned} \tag{34}$$

From eq.(34), it is noted that the function values of $T_n(\overline{\theta})$ at specific points of the variational parameter $\overline{\theta}$ give upper bounds of the normalized stochastic complexity $\overline{F}_0(X^n)$. Hence, let us consider following two cases. In both cases, $\overline{H}_n(\overline{\theta})=O_p(\frac{1}{n})$ holds.

(I): $\overline{a}_k=a_k^*,\ \overline{b}_k=b_k^*\ (1\leqq k\leqq K_0-1)$,

$\overline{a}_k=a_{K_0}^*/(K-K_0+1),\ \overline{b}_k=b_{K_0}^*\ (K_0\leqq k\leqq K)$,

then $T_n(\overline{\theta})=\dfrac{MK+K-1}{2}\log n+O_p(1)$.

(II): $\overline{a}_k=a_k^*\dfrac{n+K_0\phi_0}{n+K\phi_0},\ \overline{b}_k=b_k^*\ (1\leqq k\leqq K_0)$,

$\overline{a}_k=\dfrac{\phi_0}{n+K\phi_0},\ \overline{b}_k=\nu_0\quad(K_0+1\leqq k\leqq K)$,

then $T_n(\overline{\theta})=\Big\{(K-K_0)\phi_0+\dfrac{MK_0+K_0-1}{2}\Big\}\log n+O_p(1)$.

From eq.(34), we prove the theorem. **(Q.E.D)**

Next we show the lower bound in eq.(27).

[Theorem 4] The normalized stochastic complexity $\overline{F}_0(X^n)$ in eq.(18) satisfies the following inequality.

$$\overline{F}_0(X^n)\geqq\overline{\lambda}\log n+nH_n(\overline{\theta}_{vb})+O_p(1). \tag{35}$$

(Proof of Theorem 4) Since $\log\varphi(\overline{b}_k)$ is a bounded function of n, it follows from Lemma 3,

$$K(r(\theta|\overline{\theta})||\varphi(\theta))\geqq G(\overline{\mathbf{a}})+O_p(1). \tag{36}$$

Provided each n_k is α_k-th order of n , that is, $n_k=p_k n^{\alpha_k}+o(n^{\alpha_k}),(0<p_k,0\leqq\alpha_k\leqq 1,k=1,2,\cdots,K)$ as $n\to\infty$, since $\log(n_k+\phi_0)=\alpha_k\log n+O_p(1)$, we have

$$\frac{G(\overline{\mathbf{a}})}{\log n}\to\sum_{k=1}^{K}\{-\big(\phi_0-\frac{1}{2}\big)+\frac{M}{2}\}\alpha_k+\big(K\phi_0-\frac{1}{2}\big), \tag{37}$$

as $n\to\infty$. This means that the first term on the right-hand side of eq.(19) asymptotically only depends on $\{\alpha_k\}$ that are the orders of $\{n_k\}$ w.r.t. n as $n\to\infty$.

Also, from Lemma 1 and Lemma 4 and eq.(36), we have

$$\overline{F}_0(X^n)\geqq\min_{\overline{\theta}}\{G(\overline{\mathbf{a}})+nH_n(\overline{\theta})\}+O_p(1). \tag{38}$$

If the number of n_ks such that n_k/n converges to a positive constant is less than K_0, the number of components which the true distribution has, the distribution $p(x|\overline{\theta})$ has less components than the true one as $n \to \infty$. Then it doesn't approach the true distribution and

$$H_n(\overline{\theta}) \to K(p(x|\theta_0)||p(x|\overline{\theta})) > 0, \quad (n \to \infty)$$

holds in eq.(38). That means $nH_n(\overline{\theta})$ diverges in the order of n unless at least K_0 n_ks are in the order of n. Therefore, if $\{n_k\}$ is the minimum solution of the stochastic complexity $\overline{F}(X^n)$, it is necessary that at least K_0 n_ks are in the order of n. Finally, minimizing the limiting value of eq.(37) with respect to $\{\alpha_k\}$ under the constraint that $0 \leqq \alpha_k \leqq 1$ and at least K_0 α_ks satisfy $\alpha_k = 1$, which completes the proof.**(Q.E.D)**

Let us combine these theorems and finalize this section.

(Proof of Theorem 2) From Theorem 3 and Theorem 4, taking the expectation over all sets of training samples completes the proof. **(Q.E.D)**

8. Discussion and Conclusion

In this paper, we showed the upper and lower bounds of the stochastic complexity for the mixture of exponential family models in the Variational Bayes learning.

Firstly, we compare the stochastic complexity shown in Theorem 2 with the one in the true Bayesian learning. The stochastic complexities in the true Bayesian learning of several non-regular models have been clarified in some recent studies. On the mixture models with M parameters in each component, the following upper bound on the coefficient of $F_0(n)$ in eq.(11) is known [12] [13],

$$\lambda \leqq \begin{cases} (K + K_0 - 1)/2 & (M = 1), \\ (K - K_0) + (MK_0 + K_0 - 1)/2 & (M \geqq 2), \end{cases} \quad (39)$$

Under the same condition (i) about the true distribution and the model described in Section 5 and certain conditions about the prior distribution. Since these conditions about the prior are satisfied by putting $\phi_0 = 1$ in the condition (ii) of Theorem 2, we can compare the stochastic complexity in this case. Putting $\phi_0 = 1$ in eq.(28), we have

$$\overline{\lambda} = K - K_0 + (MK_0 + K_0 - 1)/2. \quad (40)$$

Let us compare this $\overline{\lambda}$ of the Variational Bayes learning to λ in eq.(39) of the true Bayesian learning. When $M = 1$, that is, each component has one parameter, $\overline{\lambda} \geqq \lambda$ holds since $K_0 \leqq K$. This means that the more redundant components the model has, the more the Variational Bayes learning differs from the true Bayesian learning. In this case, $2\overline{\lambda}$ is equal to $2K - 1$ that is the number of the parameters of the model. Hence the BIC [10] corresponds to $\overline{\lambda} \log n$ when $M = 1$. If $M \geqq 2$, the upper bound of λ is equal to $\overline{\lambda}$. This implies that the variational posterior is close to the true Bayesian posterior when $M \geqq 2$. More precise discussion about the accuracy of the approximation can be done for models on which more tighter bounds or exact values of the coefficient λ in eq.(11) are given [15].

Also, we point out that Theorem 2 shows how the hyperparameter ϕ_0 influence the process of the Variational Bayes learning. The coefficient $\overline{\lambda}$ in eq.(28) indicates that only when $\phi_0 \leqq (M + 1)/2$, the prior distribution works to eliminate the redundant components that the model has and otherwise it works to use all the components.

Moreover, the theoretical bounds would enable us to compare the accuracy of the Variational Bayes learning with that of the Laplace approximation or the MCMC method.

Acknowledgements.

This work was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for JSPS Fellows 4637 and for Scientific Research 15500130, 2005.

文　　献

[1] H. Akaike, "Likelihood and Bayes procedure," *Bayesian Statistics*, University Press, pp.143-166, 1980.

[2] H.Alzer, "On some inequalities for the Gamma and Psi functions," *Mathematics of computation*, Vol.66, No.217, pp.373-389, 1997.

[3] H.Attias, "Inferring parameters and structure of latent variable models by variational bayes," *Proc. of UAI'99*, 1999.

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

[5] L.D.Brown, "Fundamentals of statistical exponential families," IMS Lecture Notes-Monograph Series 9, 1986.

[6] Z.Ghahramani, M.J.Beal, "Graphical models and variational methods," *Advanced Mean Field Methods – Theory and Practice, eds. D. Saad and M. Opper, MIT Press*, 2000.

[7] J.A.Hartigan, "A Failure of likelihood asymptotics for normal mixtures," *Proc. of the Berkeley Conference in Honor of J.Neyman and J.Kiefer*, Vol.2, 807-810, 1985.

[8] D.J. Mackay, "Bayesian interpolation," *Neural Computation*, Vol.4, No.2, pp.415-447, 1992.

[9] M.Sato, "Online model selection based on the variational bayes," *Neural Computation*, Vol.13, No.7, pp.1649-1681, 2004.

[10] G.Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, Vol.6, No.2, pp.461-464, 1978.

[11] S.Watanabe,"Algebraic analysis for non-identifiable learning machines," *Neural Computation*, Vol.13, No.4, pp.899-933, 2001.

[12] K.Yamazaki, S.Watanabe, "Singularities in mixture models and upper bounds of stochastic complexity," *International Journal of Neural Networks*, 16, pp.1029-1038, 2003.

[13] K.Yamazaki, S.Watanabe "Stochastic complexity of bayesian networks," *Proc. of UAI'03*, 2003.

[14] K.Watanabe, S.Watanabe, "Stochastic complexities of normal mixture models in variational bayes learning," Proc. of IBIS2004, pp.259-265,2004.

[15] K.Watanabe, S.Watanabe, "Lower bounds of stochastic complexities in variational bayes learning of gaussian mixture models," *Proc. of IEEE conference on Cybernetics and Intelligent Systems (CIS04)*, pp.99-104, 2004.