AN ALTERNATIVE VIEW OF VARIATIONAL BAYES AND MINIMUM VARIATIONAL STOCHASTIC COMPLEXITY

Kazuho Watanabe

Graduate School of Information Science, Nara Institute of Science and Technology,
8916-5, Takayama-cho, Ikoma, Nara, 630-0192, JAPAN, wkazuho@is.naist.jp

ABSTRACT

Bayesian learning is widely used in many applied data-modelling problems and is often accompanied with approximation schemes since it requires intractable computation of the posterior distributions. In this study, we focus on the two approximation methods, the variational Bayes and the local variational approximation. We show that the variational Bayes approach for statistical models with latent variables can be viewed as a special case of the local variational approximation, where the log-sum-exp function is used to form the lower bound of the log-likelihood. The minimum variational stochastic complexity, that is the objective function of the variational Bayes, is also examined and related to the asymptotic theory of Bayesian learning.

1. INTRODUCTION

Bayesian estimation provides a powerful framework for learning from data. Recently, its asymptotic theory has been established, which supports its effectiveness for latent variable models such as the Gaussian mixture model (GMM) and the hidden Markov model (HMM). More specifically, a formula for evaluating asymptotic forms of the stochastic complexity or the free energy was obtained and the generalization errors of statistical models have been intensively analyzed [1, 2, 3, 4].

Practically, however, Bayesian estimation requires some approximation method since computing the Bayesian posterior distribution is intractable in general. In this study, we focus on two approximation methods, the variational Bayes and the local variational approximation, for Bayesian estimation. The former has been successfully applied to latent variable models such as mixture models and HMMs [5, 6, 7]. Furthermore, its asymptotic analysis has progressed in several statistical models [8, 9, 10, 11]. The latter, also known as direct site bounding, has been applied to the logistic regression [12] and the sparse linear models [13] representatively. It is known that this approximation is generally characterized and described by using the Bregman divergence [14].

In this paper, by providing the general framework for the local variational approximation, we show that the variational Bayes for the latent variable models can be interpreted as an application of the local variational approximation. From this viewpoint, we also investigate the asymptotic behavior of the variational stochastic complexity also known as the variational free energy, which is the objective function to be minimized by the variational Bayes. More specifically, we present a formula for evaluating the asymptotic form of the minimum variational stochastic complexity relating it to the asymptotic theory of Bayesian estimation. This formula explains relationships between several previous works where asymptotic stochastic complexity and the minimum variational stochastic complexity have been analyzed respectively [2, 3, 4, 8, 9, 10, 11]. We apply it to the GMM as an example and demonstrate another proof of the upper bound of the minimum variational stochastic complexity previously obtained in [8].

The rest of this paper is organized as follows. Section 2 describes the Bayesian estimation and briefly introduces its asymptotic theory. Section 3 reviews the variational Bayes for the latent variable models and the general framework for the local variational approximation. Section 4 shows that a special case of the local variational approximation reduces to the variational Bayes for latent variable models. Section 5 presents the formula for the asymptotic analysis of the minimum variational stochastic complexity. Section 6 demonstrates its application to the GMM. Discussion and conclusion follow in Section 7 and 8.

2. BAYESIAN LEARNING

Assume we are given training examples or observations $\boldsymbol{x} = \{x_1, x_2, \cdots, x_n\}$ where each observation x_i is defined in some domain $\mathcal{X}$. Let $\boldsymbol{w} \in R^d$ be the parameter vector and consider Bayesian learning for a model $p(\boldsymbol{x}|\boldsymbol{w})$. By using the prior distribution $p_0(\boldsymbol{w})$, the Bayesian posterior distribution of the parameter $\boldsymbol{w}$ is defined by,

$$p(\boldsymbol{w}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{w})p_0(\boldsymbol{w})}{\int p(\boldsymbol{x}|\boldsymbol{w})p_0(\boldsymbol{w})d\boldsymbol{w}} = \frac{p(\boldsymbol{w},\boldsymbol{x})}{Z}. \quad (1)$$

As is often the case, the normalizing constant,

$$Z = p(\boldsymbol{x}) = \int p(\boldsymbol{x}|\boldsymbol{w})p_0(\boldsymbol{w})d\boldsymbol{w},$$

called the marginal likelihood or the evidence, is analytically intractable and so is the Bayesian posterior distribution (1). The stochastic complexity or the free energy F is defined by the negative logarithm of Z,

$$F = -\log Z.$$

Although it is an issue to compute or approximate the stochastic complexity practically, an asymptotic theory for analyzing the Bayesian stochastic complexity was established, over which we provide a brief overview.

Let $p^*(x) = p(x|\boldsymbol{w}^*)$ be the true data generating distribution independently and identically and

$$S = -\left\langle \log p^*(x) \right\rangle_{p^*(x)}$$

be its entropy[1]. By the i.i.d. assumption we have that $p(\boldsymbol{x}|\boldsymbol{w}) = \prod_{i=1}^n p(x_i|\boldsymbol{w})$. We define the (average) normalized stochastic complexity by

$$\begin{aligned} F^* &= \left\langle -\log \int \prod_{i=1}^n \frac{p(x_i|\boldsymbol{w})}{p(x_i|\boldsymbol{w}^*)} p_0(\boldsymbol{w}) d\boldsymbol{w} \right\rangle_{\prod_{i=1}^n p^*(x_i)}, \\ &= \langle F \rangle_{\prod_{i=1}^n p^*(x_i)} - nS. \end{aligned}$$

Then, it was proved that the average normalized Bayesian stochastic complexity has the following asymptotic form,

$$F^* \simeq \lambda \log n - (m-1) \log \log n + O(1), \qquad (2)$$

where the $O(1)$ term is bounded by a constant independent of n. The constants λ and m are the rational number and the natural number respectively which are identified by the largest pole and its order of the zeta function,

$$J_H(z) = \int H(\boldsymbol{w})^z p_0(\boldsymbol{w}) d\boldsymbol{w}, \qquad (3)$$

where

$$H(\boldsymbol{w}) = \int p(x|\boldsymbol{w}^*) \log \frac{p(x|\boldsymbol{w}^*)}{p(x|\boldsymbol{w})} dx.$$

In regular statistical models such as exponential families, 2λ is equal to the number of parameters and $m = 1$, whereas in non-regular models such as GMMs, 2λ is not larger than the number of parameters and $m \geq 1$. For several statistical models, the coefficient λ or its upper bound was evaluated by analyzing the pole of the zeta function [1, 2, 3, 4].

3. APPROXIMATION METHODS

This section provides brief summaries of the two approximation methods of Bayesian estimation. The relationship between them is detailed in the next section.

3.1. Variational Bayes for Latent Variable Models

Let $\boldsymbol{y} = \{y_1, y_2, \cdots, y_n\}$ be the latent (unobserved) variables corresponding to the observations $\boldsymbol{x} = \{x_1, x_2, \cdots, x_n\}$ and consider the latent variable model,

$$p(\boldsymbol{x}|\boldsymbol{w}) = \sum_{\boldsymbol{y}} p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{w}),$$

where $\sum_{\boldsymbol{y}}$ denotes the summation over all possible realizations of the latent variables. Hereafter until Section 4, we do not assume $\boldsymbol{x}$ and $\boldsymbol{y}$ to be i.i.d. observations. In Section 5, we again assume i.i.d. observations and refer to the asymptotic theory presented in Section 2. We assume discrete latent variables $\boldsymbol{y}$ to include the examples such as the GMM and the HMM. The generalization to the continuous latent variables is straightforward.

The Bayesian posterior distribution of the latent variables and the parameter $\boldsymbol{w}$ is

$$p(\boldsymbol{y}, \boldsymbol{w}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{w}) p_0(\boldsymbol{w})}{\sum_{\boldsymbol{y}} \int p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{w}) p_0(\boldsymbol{w}) d\boldsymbol{w}},$$

which is intractable when $Z = \sum_{\boldsymbol{y}} \int p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{w}) p_0(\boldsymbol{w}) d\boldsymbol{w}$ requires the sum over exponentially many terms as in GMMs and HMMs and so is the posterior of the parameter

$$p(\boldsymbol{w}|\boldsymbol{x}) = \sum_{\boldsymbol{y}} p(\boldsymbol{y}, \boldsymbol{w}|\boldsymbol{x}). \qquad (4)$$

In the variational Bayesian framework, the Bayesian posterior distribution $p(\boldsymbol{y}, \boldsymbol{w}|\boldsymbol{x})$ of the hidden variables and the parameters is approximated by the variational posterior distribution $q(\boldsymbol{y}, \boldsymbol{w}|\boldsymbol{x})$, which factorizes as

$$q(\boldsymbol{y}, \boldsymbol{w}|\boldsymbol{x}) = q(\boldsymbol{y}|\boldsymbol{x}) q(\boldsymbol{w}|\boldsymbol{x}), \qquad (5)$$

where $q(\boldsymbol{y}|\boldsymbol{x})$ and $q(\boldsymbol{w}|\boldsymbol{x})$ are probability distributions on the hidden variables and the parameters respectively. The variational posterior $q(\boldsymbol{y}, \boldsymbol{w}|\boldsymbol{x})$ is chosen so that it minimizes the functional $\overline{F}[q]$, called the variational free energy or the variational stochastic complexity, defined by

$$\begin{aligned} \overline{F}[q] &= \sum_{\boldsymbol{y}} \int q(\boldsymbol{y}, \boldsymbol{w}|\boldsymbol{x}) \log \frac{q(\boldsymbol{y}, \boldsymbol{w}|\boldsymbol{x})}{p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{w}) p_0(\boldsymbol{w})} d\boldsymbol{w}, \\ &= F(\boldsymbol{x}) + K(q(\boldsymbol{y}, \boldsymbol{w}|\boldsymbol{x}) || p(\boldsymbol{y}, \boldsymbol{w}|\boldsymbol{x})), \qquad (6) \end{aligned}$$

where $K(q(\boldsymbol{y}, \boldsymbol{w}|\boldsymbol{x}) || p(\boldsymbol{y}, \boldsymbol{w}|\boldsymbol{x}))$ is the Kullback information from the true Bayesian posterior $p(\boldsymbol{y}, \boldsymbol{w}|\boldsymbol{x})$ to the variational posterior $q(\boldsymbol{y}, \boldsymbol{w}|\boldsymbol{x})$ [2]. This leads to the following alternate optimization over $q(\boldsymbol{y}|\boldsymbol{x})$ and $q(\boldsymbol{w}|\boldsymbol{x})$ [5, 6, 7]. For a fixed $q(\boldsymbol{y}|\boldsymbol{x})$, the functional $\overline{F}[q]$ as a function of $q(\boldsymbol{w}|\boldsymbol{x})$ is minimized by

$$q(\boldsymbol{w}|\boldsymbol{x}) = \frac{1}{C_w} p_0(\boldsymbol{w}) \exp \left\langle \log p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{w}) \right\rangle_{q(\boldsymbol{y}|\boldsymbol{x})}. \qquad (7)$$

For a fixed $q(\boldsymbol{w}|\boldsymbol{x})$, it as a function of $q(\boldsymbol{y}|\boldsymbol{x})$ is minimized by

$$q(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{C_y} \exp \left\langle \log p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{w}) \right\rangle_{q(\boldsymbol{w}|\boldsymbol{x})}. \qquad (8)$$

Here C_w and C_y are the normalization constants. In Section 4, we show that this algorithm can be interpreted as an application of another approximation scheme, local variational approximation.

[1] For an arbitrary distribution $p(x)$, $\langle \cdot \rangle_{p(x)}$ denotes the expectation over $p(x)$.

[2] Throughout this paper, we use the notation $K(q(x)||p(x))$ for the Kullback information from a distribution $q(x)$ to a distribution $p(x)$, that is,

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

3.2. Local Variational Approximation

This section describes several facts regarding the local variational approximation [7, 14].

Local variational approximation forms a lower bound of $p(\boldsymbol{w}, \boldsymbol{x})$, denoted by $p_{\underline{\boldsymbol{\xi}}}(\boldsymbol{w}, \boldsymbol{x})$,

$$\underline{p_{\boldsymbol{\xi}}}(\boldsymbol{w}, \boldsymbol{x}) \quad \leq \quad p(\boldsymbol{w}, \boldsymbol{x}), \tag{9}$$

and approximates the posterior distribution (1) by

$$p_{\boldsymbol{\xi}}(\boldsymbol{w}|\boldsymbol{x}) \quad = \quad \frac{\underline{p_{\boldsymbol{\xi}}}(\boldsymbol{w}, \boldsymbol{x})}{\underline{Z}(\boldsymbol{\xi})}, \tag{10}$$

where $\underline{Z}(\boldsymbol{\xi}) = \int \underline{p_{\boldsymbol{\xi}}}(\boldsymbol{w}, \boldsymbol{x}) d\boldsymbol{w}$, and $\boldsymbol{\xi}$ is called the variational parameter. The above approximation is optimized by estimating the variational parameter $\boldsymbol{\xi}$ so that $\underline{Z}(\boldsymbol{\xi})$ is maximized since the inequality

$$\underline{Z}(\boldsymbol{\xi}) \leq Z \tag{11}$$

holds by definition. This is equivalent to the minimization of $\overline{F}(\boldsymbol{\xi}) = -\log \underline{Z}(\boldsymbol{\xi})$ which is an upper bound of the stochastic complexity, $F = -\log Z$.

Most existing local variational approximation techniques are based on the convexity of the log-likelihood function or the log-prior [7]. Let ϕ be a continuously-differentiable real-valued convex function and $\boldsymbol{h}$ be a vector-valued function. The lower bound of the joint distribution is formed as follows,

$$\begin{aligned} p(\boldsymbol{w}, \boldsymbol{x}) &= p(\boldsymbol{x}|\boldsymbol{w}) p_0(\boldsymbol{w}) \\ &\geq p(\boldsymbol{x}|\boldsymbol{w}) p_0(\boldsymbol{w}) \exp\{-d_\phi(\boldsymbol{h}(\boldsymbol{w}), \boldsymbol{h}(\boldsymbol{\xi}))\} \\ &\equiv \underline{p_{\boldsymbol{\xi}}}(\boldsymbol{w}, \boldsymbol{x}), \end{aligned} \tag{12}$$

where

$$d_\phi(\boldsymbol{u}, \boldsymbol{v}) = \phi(\boldsymbol{u}) - \phi(\boldsymbol{v}) - (\boldsymbol{u} - \boldsymbol{v}) \cdot \nabla\phi(\boldsymbol{v}) \geq 0, \tag{13}$$

is the Bregman divergence associated with the convex function ϕ [15].

Then we obtain the following expression,

$$\overline{F}(\boldsymbol{\xi}) - F = \langle d_\phi(\boldsymbol{h}(\boldsymbol{w}), \boldsymbol{h}(\boldsymbol{\xi})) \rangle_{p_{\boldsymbol{\xi}}} + K(p_{\boldsymbol{\xi}} || p). \tag{14}$$

From eq.(12), the approximating posterior is given by,

$$\begin{aligned} p_{\boldsymbol{\xi}}(\boldsymbol{w}|\boldsymbol{x}) \quad \propto \quad & \exp\{\boldsymbol{h}(\boldsymbol{w}) \cdot \nabla\phi(\boldsymbol{h}(\boldsymbol{\xi})) \\ & + \log p(\boldsymbol{x}, \boldsymbol{w}) - \phi(\boldsymbol{h}(\boldsymbol{w}))\}, \end{aligned} \tag{15}$$

which is a member of the exponential family. The expectation maximization (EM) algorithm for minimizing the upper bound $\overline{F}(\boldsymbol{\xi})$ updates the old estimate $\tilde{\boldsymbol{\xi}}$ to $\boldsymbol{\xi}$ so that

$$\boldsymbol{h}(\boldsymbol{\xi}) = \langle \boldsymbol{h}(\boldsymbol{w}) \rangle_{p_{\tilde{\boldsymbol{\xi}}}} \tag{16}$$

is satisfied [14].

4. AN ALTERNATIVE VIEW OF VARIATIONAL BAYES

Let us consider an application of the local variational method for approximating the posterior distribution of the latent variable model, $p(\boldsymbol{w}|\boldsymbol{x})$ in eq.(4). By the convexity of the function $\log \sum_{\boldsymbol{y}} \exp(\cdot)$, the log-likelihood is bounded below as follows,

$$\begin{aligned} & \log p(\boldsymbol{x}|\boldsymbol{w}) = \log \sum_{\boldsymbol{y}} \exp\{\log p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{w})\} \\ \geq \quad & \log p(\boldsymbol{x}|\boldsymbol{\xi}) + \sum_{\boldsymbol{y}} \left(\log \frac{p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{w})}{p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\xi})} \right) p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\xi}), \end{aligned} \tag{17}$$

where $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\xi}) = \frac{p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\xi})}{\sum_{\boldsymbol{y}} p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\xi})}$. This corresponds to the case where $\phi(\boldsymbol{h}) = \log \sum_i \exp(h_i)$ and $\boldsymbol{h}(\boldsymbol{w})$ is the vector-valued function which consists of the elements $\log p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{w})$ for all possible $\boldsymbol{y}$. Since

$$d_\phi(\boldsymbol{h}(\boldsymbol{w}), \boldsymbol{h}(\boldsymbol{\xi})) = \sum_{\boldsymbol{y}} p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\xi}) \log \frac{p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\xi})}{p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w})},$$

from eq.(14), we have

$$\begin{aligned} \overline{F}(\boldsymbol{\xi}) &= F + K(p_{\boldsymbol{\xi}}(\boldsymbol{w}|\boldsymbol{x}) || p(\boldsymbol{w}|\boldsymbol{x})) \\ &\quad + \langle K(p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\xi}) || p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w})) \rangle_{p_{\boldsymbol{\xi}}(\boldsymbol{w}|\boldsymbol{x})} \\ &= F + K(p_{\boldsymbol{\xi}}(\boldsymbol{w}|\boldsymbol{x}) p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\xi}) || p(\boldsymbol{w}, \boldsymbol{y}|\boldsymbol{x})), \end{aligned}$$

which is exactly the variational stochastic complexity of the factorized distribution $p_{\boldsymbol{\xi}}(\boldsymbol{w}|\boldsymbol{x}) p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\xi})$. In fact, from eqs.(15) and (17) , the approximating posterior is given by

$$\begin{aligned} p_{\boldsymbol{\xi}}(\boldsymbol{w}|\boldsymbol{x}) \quad \propto \quad & \exp\left\{ \sum_{\boldsymbol{y}} \log p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{w}) p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\xi}) \right\} p_0(\boldsymbol{w}) \\ = \quad & \exp \langle \log p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{w}) \rangle_{p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\xi})} \, p_0(\boldsymbol{w}). \end{aligned} \tag{18}$$

From eq.(16), the EM update for $\boldsymbol{\xi}$ yields

$$\begin{aligned} & \log p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\xi}) = \langle \log p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{w}) \rangle_{p_{\tilde{\boldsymbol{\xi}}}(\boldsymbol{w}|\boldsymbol{x})} \\ \Rightarrow \quad & p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\xi}) \propto \exp \langle \log p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{w}) \rangle_{p_{\tilde{\boldsymbol{\xi}}}(\boldsymbol{w}|\boldsymbol{x})}. \end{aligned} \tag{19}$$

Eqs.(18) and (19) are exactly same as the variational Bayes algorithm for minimizing the variational stochastic complexity over the factorized distributions, eqs.(7) and (8).

5. MINIMUM VARIATIONAL STOCHASTIC COMPLEXITY

Let $\overline{F}_{\min} = \min_{\boldsymbol{h}(\boldsymbol{\xi})} \overline{F}(\boldsymbol{\xi})$ be the minimum variational stochastic complexity. From the inequality (17), we have

$$\begin{aligned} & \overline{F}_{\min} \\ = \quad & \min_{\boldsymbol{h}(\boldsymbol{\xi})} \left\{ -\log \int \underline{p_{\boldsymbol{\xi}}}(\boldsymbol{w}, \boldsymbol{x}) d\boldsymbol{w} \right\} \\ = \quad & \min_{\boldsymbol{h}(\boldsymbol{\xi})} \{ -\log p(\boldsymbol{x}|\boldsymbol{\xi}) \\ & - \log \int \exp\{ \sum_{\boldsymbol{y}} p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\xi}) \log \frac{p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{w})}{p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\xi})} \} p_0(\boldsymbol{w}) d\boldsymbol{w} \} \\ \leq \quad & -\log p(\boldsymbol{x}|\boldsymbol{w}^*) \\ & - \log \int \exp\{ \sum_{\boldsymbol{y}} p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w}^*) \log \frac{p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{w})}{p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{w}^*)} \} p_0(\boldsymbol{w}) d\boldsymbol{w}. \end{aligned} \tag{20}$$

Here, in the last inequality, we have substituted $\boldsymbol{h}(\boldsymbol{\xi}) = \boldsymbol{h}(\boldsymbol{w}^*)$ and assumed that $p(x|\boldsymbol{w}^*)$ with the parameter $\boldsymbol{w}^*$ is the underlying distribution generating the data $\boldsymbol{x}$ independently and identically. We also assume that $p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{w}) = \prod_{i=1}^n p(x_i, y_i|\boldsymbol{w})$ which implies $p(\boldsymbol{x}|\boldsymbol{w}) = \prod_{i=1}^n p(x_i|\boldsymbol{w})$ and $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w}) = \prod_{i=1}^n p(y_i|x_i, \boldsymbol{w})$. By subtracting the entropy of the true distribution, we define the (average) normalized minimum variational stochastic complexity,

$$\begin{aligned}\overline{F}^*_{\min} &= \left\langle \overline{F}_{\min} + \log p(\boldsymbol{x}|\boldsymbol{w}^*) \right\rangle_{p(\boldsymbol{x}|\boldsymbol{w}^*)}, \\ &= \left\langle \overline{F}_{\min} \right\rangle_{p(\boldsymbol{x}|\boldsymbol{w}^*)} - nS.\end{aligned}$$

Then it follows from the above inequality and the convexity of $\log \int \exp(\cdot) d\boldsymbol{w}$ that

$$\overline{F}^*_{\min} \leq -\log \int e^{-n\overline{H}(\boldsymbol{w})} p_0(\boldsymbol{w}) d\boldsymbol{w}, \qquad (21)$$

where

$$\overline{H}(\boldsymbol{w}) = \int \sum_y p(x, y|\boldsymbol{w}^*) \log \frac{p(x, y|\boldsymbol{w}^*)}{p(x, y|\boldsymbol{w})} dx.$$

The asymptotic theory of the Bayesian estimation [1] tells that the asymptotic form of the right hand side of eq.(21), providing an upper bound of $\overline{F}^*_{\min}$, is given by

$$\begin{aligned}\overline{F}^*_{\min} &\leq -\log \int e^{-n\overline{H}(\boldsymbol{w})} p_0(\boldsymbol{w}) d\boldsymbol{w} \\ &\simeq \overline{\lambda} \log n - (\overline{m} - 1) \log \log n + O(1), \qquad (22)\end{aligned}$$

where $\overline{\lambda}$ and $\overline{m}$ are respectively the largest pole and its order of the zeta function defined by

$$J_{\overline{H}}(z) = \int \overline{H}(\boldsymbol{w})^z p_0(\boldsymbol{w}) d\boldsymbol{w}. \qquad (23)$$

This means that the asymptotic behavior of the minimum variational stochastic complexity is characterized by $\overline{H}(\boldsymbol{w})$ while that of the stochastic complexity F is characterized by $H(\boldsymbol{w}) = K(p(x|\boldsymbol{w}^*)||p(x|\boldsymbol{w}))$ as in eqs.(2) and (3) [1]. These two functions are related by the log-sum inequality, $H(\boldsymbol{w}) \leq \overline{H}(\boldsymbol{w})$.

6. EXAMPLE: GAUSSIAN MIXTURE MODEL

In this section, we derive an asymptotic upper bound of the minimum variational stochastic complexity of the GMM. Although this upper bound was obtained in a previous work [8], the derivation was by direct evaluation and minimization of the variational stochastic complexity with respect to the expected sufficient statistics which corresponds to the variational parameter $\boldsymbol{\xi}$ in this paper. We present another derivation through eq.(21) for an illustration of the asymptotic analysis described in Section 5.

6.1. Variational Bayes for GMM

Let $g(x|\mu) = \frac{1}{\sqrt{2\pi}^M} \exp\{-\frac{||x-\mu||^2}{2}\}$ be the M-dimensional Gaussian density and consider the GMM with K components,

$$p(x|\boldsymbol{w}) = \sum_{k=1}^K a_k g(x|\mu_k)$$

where $x \in R^M$ and the parameter vector $\boldsymbol{w}$ consists of the mean vectors $\{\mu_k\}_{k=1}^K$ and the mixing proportions $\boldsymbol{a} = \{a_k\}_{k=1}^K$ that satisfy $0 \leq a_k \leq 1$ for $k = 1, \cdots, K$ and $\sum_{k=1}^K a_k = 1$. As a latent variable model, this model is expressed as, $p(x|\boldsymbol{w}) = \sum_y p(x, y|\boldsymbol{w})$, where

$$p(x, y|\boldsymbol{w}) = \prod_{k=1}^K \{a_k g(x|\mu_k)\}^{y^{(k)}}.$$

The latent variable $y = (y^{(1)}, y^{(2)}, \cdots, y^{(K)})$ indicates the component from which the datum x is generated, that is, $y^{(k)} = 1$ if x is from the kth component and $y^{(k)} = 0$ otherwise. The variational Bayes framework is successfully applied to this model [5, 6, 7] using the prior distribution,

$$p_0(\boldsymbol{w}) = p_0(\boldsymbol{a}) \prod_{k=1}^K p_0(\mu_k), \qquad (24)$$

where

$$p_0(\boldsymbol{a}) = \frac{\Gamma(K\phi_0)}{\Gamma(\phi_0)^K} \prod_{k=1}^K a_k^{\phi_0 - 1}$$

is the Dirichlet distribution with the hyperparameter $\phi_0 > 0$ and

$$p_0(\mu_k) = \sqrt{\frac{\xi_0}{2\pi}}^M \exp\{-\frac{\xi_0 ||\mu_k - \nu_0||^2}{2}\}$$

is the Gaussian distribution with the hyperparameters $\xi_0 > 0$ and $\nu_0 \in R^M$. They are the conjugate prior distributions for the mixing proportions and each mean vector respectively.

6.2. Asymptotic Analysis of $\overline{F}_{\min}$

We assume that the true data generating distribution is $p(x|\boldsymbol{w}^*)$ with the parameter $\boldsymbol{w}^* = \{\{a_k^*\}, \{\mu_k^*\}\}$,

$$p(x|\boldsymbol{w}^*) = \sum_{k=1}^{K_0} a_k^* g(x|\mu_k^*), \qquad (25)$$

and $K_0 \leq K$ holds, that is, the true distribution is realizable by the postulated model. Then it was proved in [8] that the normalized minimum variational stochastic complexity satisfies

$$\overline{F}^*_{\min} \leq \overline{\lambda} \log n + O(1), \qquad (26)$$

where

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

6.3. Derivation of eq.(26)

In this section, we derive eq.(26) from eq.(21), which provides another proof than that presented in [8].

Firstly, in order to define $p(x, y|\boldsymbol{w}^*)$ for y with K elements, we extend and redefine the true parameter $\boldsymbol{w}^*$ denoting it as $\tilde{\boldsymbol{w}}^* = \{\{\tilde{a}_k^*\}_{k=1}^K, \{\tilde{\mu}_k^*\}_{k=1}^K\}$,

$$\tilde{a}_k^* = \begin{cases} a_k^* & (1 \leq k \leq K_0 - 1), \\ a_{K_0}^*/(K - K_0 + 1) & (K_0 \leq k \leq \hat{K}), \\ 0 & (\hat{K} + 1 \leq k \leq K), \end{cases}$$

and

$$\tilde{\mu}_k^* = \begin{cases} \mu_k^* & (1 \leq k \leq K_0), \\ \mu_{K_0}^* & (K_0+1 \leq k \leq K), \end{cases}$$

where $\hat{K}$ is the number of components whose mixing proportions are non-zero. Note that the marginal distribution of $p(x,y|\tilde{\boldsymbol{w}}^*)$ reduces to eq.(25). Then we have

$$\begin{aligned} \overline{H}(\boldsymbol{w}) &= \int \sum_y p(x,y|\tilde{\boldsymbol{w}}^*) \log \frac{p(x,y|\tilde{\boldsymbol{w}}^*)}{p(x,y|\boldsymbol{w})} dx \\ &= \int \sum_{k=1}^{K} \tilde{a}_k^* g(x|\tilde{\mu}_k^*) \log \frac{\tilde{a}_k^* g(x|\tilde{\mu}_k^*)}{a_k g(x|\mu_k)} dx \\ &= \sum_{k=1}^{K} \tilde{a}_k^* \left\{ \log \frac{\tilde{a}_k^*}{a_k} + \int g(x|\tilde{\mu}_k^*) \log \frac{g(x|\tilde{\mu}_k^*)}{g(x|\mu_k)} \right\} \\ &= \sum_{k=1}^{\hat{K}} \tilde{a}_k^* \left\{ \log \frac{\tilde{a}_k^*}{a_k} + \frac{||\mu_k - \tilde{\mu}_k^*||^2}{2} \right\}. \end{aligned}$$

Secondly, we divide the parameter $\boldsymbol{w}$ into three parts,

$$\begin{aligned} \boldsymbol{w}_1 &= (a_2, a_3, \cdots, a_{\hat{K}}), \\ \boldsymbol{w}_2 &= (a_{\hat{K}+1}, \cdots, a_K), \\ \boldsymbol{w}_3 &= (\mu_1, \mu_2, \cdots, \mu_{\hat{K}}). \end{aligned}$$

and define

$$\begin{aligned} W_1 &= \{\boldsymbol{w}_1 |\ |a_k - \tilde{a}_k^*| \leq \epsilon, 2 \leq k \leq \hat{K}\}, \\ W_2 &= \{\boldsymbol{w}_2 |\ |a_k| \leq \epsilon, \hat{K} \leq k \leq K\}, \\ W_3 &= \{\boldsymbol{w}_3 |\ ||\mu_k - \tilde{\mu}_k^*|| \leq \epsilon, 1 \leq k \leq \hat{K}\}, \end{aligned}$$

for a sufficiently small constant ϵ. For an arbitrary parameter $\boldsymbol{w} \in W_1 \times W_2 \times W_3 \equiv W(\epsilon)$, we can decompose $\overline{H}(\boldsymbol{w})$ as,

$$\overline{H}(\boldsymbol{w}) = \overline{H}_1(\boldsymbol{w}_1) + \overline{H}_2(\boldsymbol{w}_2) + \overline{H}_3(\boldsymbol{w}_3), \tag{27}$$

where

$$\begin{aligned} &\overline{H}_1(\boldsymbol{w}_1) \\ &= \sum_{k=2}^{\hat{K}} \tilde{a}_k^* \log \frac{\tilde{a}_k^*}{a_k} + \left(1 - \sum_{k=2}^{\hat{K}} \tilde{a}_k^*\right) \log \frac{1 - \sum_{k=2}^{\hat{K}} \tilde{a}_k^*}{1 - \sum_{k=2}^{\hat{K}} a_k}, \end{aligned}$$

$$\overline{H}_2(\boldsymbol{w}_2) = \frac{1}{1-c} \frac{1 - \sum_{k=2}^{K_0} \tilde{a}_k^*}{1 - \sum_{k=2}^{\hat{K}} a_k} \sum_{k=\hat{K}+1}^{K} a_k,$$

and

$$\overline{H}_3(\boldsymbol{w}_3) = \sum_{k=1}^{\hat{K}} \frac{\tilde{a}_k^*}{2} ||\mu_k - \tilde{\mu}_k^*||^2. \tag{28}$$

Here we have used the mean value theorem $-\log(1-t) = \frac{1}{1-c}t$ for some c, $0 \leq c \leq t$ with $t = \frac{\sum_{k=\hat{K}+1}^{K} a_k}{1 - \sum_{k=2}^{\hat{K}} a_k}$. Furthermore, for $\boldsymbol{w} \in W(\epsilon)$, there exist positive constants C_1, C_2 C_3 and C_4 such that

$$C_1 \sum_{k=2}^{\hat{K}} (a_k - \tilde{a}_k^*)^2 \leq \overline{H}_1(\boldsymbol{w}_1) \leq C_2 \sum_{k=2}^{\hat{K}} (a_k - \tilde{a}_k^*)^2, \tag{29}$$

and

$$C_3 \sum_{k=\hat{K}+1}^{K} a_k \leq \overline{H}_2(\boldsymbol{w}_2) \leq C_4 \sum_{k=\hat{K}+1}^{K} a_k. \tag{30}$$

Thirdly, we evaluate the partial variational stochastic complexities defined, for $i = 1, 2, 3$, by

$$\overline{F}_i = -\log \int_{W_i} \exp(-n\overline{H}_i(\boldsymbol{w}_i)) p_0(\boldsymbol{w}_i) d\boldsymbol{w}_i, \tag{31}$$

where $p_0(\boldsymbol{w}_i)$ is the product of the factors involving the corresponding parameters in eq.(24).

It follows from eqs.(21), (27) and (31) that

$$\overline{F}^*_{\min} \leq \overline{F}_1 + \overline{F}_2 + \overline{F}_3 + O(1). \tag{32}$$

From eq.(29) and eq.(28), as for $\overline{F}_1$ and $\overline{F}_3$, the Gaussian integration yields,

$$\overline{F}_1 = \frac{\hat{K}-1}{2} \log n + O(1), \tag{33}$$

and

$$\overline{F}_3 = \frac{M\hat{K}}{2} \log n + O(1). \tag{34}$$

Since

$$n^{\phi_0} \int_0^{\epsilon} e^{-na_k} a_k^{\phi_0 - 1} da_k \to \Gamma(\phi_0) \quad (n \to \infty),$$

for $k = \hat{K}+1, \cdots, K$, it follows from eq.(30),

$$\overline{F}_2 = (K - \hat{K})\phi_0 \log n + O(1). \tag{35}$$

Finally, combining eqs.(33), (35),(34) and (32), we obtain

$$\overline{F}^*_{\min} \leq \left\{ (K - \hat{K})\phi_0 + \frac{M\hat{K} + \hat{K} - 1}{2} \right\} \log n + O(1).$$

Minimizing the right hand side of the above expression over $\hat{K}$ ($K_0 \leq \hat{K} \leq K$) leads to eq.(26).

Additionally, the above evaluations of all the partial variational stochastic complexities are obtained alternatively by using the zeta function method as mentioned in Section 5. For example, as for $\overline{F}_2$, the zeta function

$$J_{\overline{H}_2}(z) = \int \overline{H}_2(\boldsymbol{w}_2)^z p_0(\boldsymbol{w}_2) d\boldsymbol{w}_2$$

has a pole $z = (K - \hat{K})\phi_0$. This can be observed by the change of variables, the so-called blow-up,

$$\begin{aligned} a_k &= a'_k a'_K \quad (k = \hat{K}+1, \cdots, K-1), \\ a_K &= a'_K, \end{aligned}$$

which yields that $J_{\overline{H}_2}$ has a term

$$\int a_K'^z a_K'^{(K-\hat{K})\phi_0 - 1} \tilde{J}_{\overline{H}_2}(\tilde{\boldsymbol{w}}'_2) da'_K = \frac{\tilde{J}_{\overline{H}_2}(\tilde{\boldsymbol{w}}'_2)}{z - (K - \hat{K})\phi_0},$$

where $\tilde{J}_{\overline{H}_2}(\tilde{\boldsymbol{w}}'_2)$ is a function proportional to

$$\int \left(\sum_{k=\hat{K}+1}^{K-1} a'_k + 1 \right)^z \prod_{k=\hat{K}+1}^{K-1} a_k'^{\phi_0 - 1} \prod_{k=\hat{K}+1}^{K-1} da'_k.$$

Hence, we can see that $J_{\overline{H}_2}$ has a pole $z = (K - \hat{K})\phi_0$.

7. DISCUSSION

We presented a formula for analyzing the asymptotic behavior of the minimum variational stochastic complexity in Section 5 and demonstrated its application to the GMM in Section 6. In this section, we discuss its relationships to the previous works where the stochastic complexity and the minimum variational stochastic complexity were respectively analyzed for specific latent variable models.

Asymptotic upper bounds of the stochastic complexity were obtained for some statistical models including the mixture model, HMM and the Bayesian network [2, 3, 4]. The upper bounds are given in such forms as

$$F^* \leq \nu \log n + O(1),$$

where the coefficient ν was identified for each model by analyzing the largest pole of the zeta function J_H in eq.(3). More specifically, however, these works analyzed the largest pole of $J_{\overline{H}}$ in eq.(23) instead of J_H by using the log-sum inequality [2, 3, 4]. Since the largest pole of $J_{\overline{H}}$ provides a lower bound for that of J_H, their analyses provided upper bounds of F^* for the above models.

On the other hand, the asymptotic forms of the minimum variational stochastic complexity were analyzed also for the same models [8, 9, 10, 11], each of which has the form

$$\overline{F}^*_{\min} \leq \overline{\lambda} \log n + O(1).$$

In most cases, the asymptotic upper bound of F^* and $\overline{F}^*_{\min}$ coincide, that is, $\nu = \overline{\lambda}$ holds. The assertion in Section 5 implies that this is generally true since it formally relates the asymptotic form of the minimum variational stochastic complexity $\overline{F}^*_{\min}$ to $\overline{H}(\boldsymbol{w})$ and the largest pole of $J_{\overline{H}}$.

Moreover, the previous analyses of the minimum variational stochastic complexity were based on the direct minimization of the variational stochastic complexity over the expected sufficient statistics which correspond to the variational parameter $\boldsymbol{\xi}$ in this paper. Hence the analyses were highly dependent on the concrete algorithm for the specific model and the choice of the prior distribution. Analyzing the right hand side of eq.(21) is more general and is independent of the concrete algorithm for the specific model. It does not even require that the prior distribution $p_0(\boldsymbol{w})$ to be conjugate prior although in this case the variational Bayes algorithm turns out not to be practical.

It is strongly implied from the inequality (20) that this inequality turns into an equality if the consistency in the sense that $p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\xi}^*) = p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{w}^*)$ holds for all $\boldsymbol{y}$ is guaranteed for the optimal variational parameter $\boldsymbol{\xi}^*$ as $n \to \infty$. It is an important undertaking to elucidate the condition under which this consistency holds and the upper bound (22) gives the exact asymptotic form of $\overline{F}^*_{\min}$. The approach presented in this paper is applicable for evaluating the asymptotic approximation accuracy of other models and other choices of the convex function ϕ. This will be pursued in the future.

8. CONCLUSION

In this paper, we provided an alternative view of the variational Bayes for latent variable models as an application of the local variational approximation. Combining this view with the asymptotic theory of Bayesian estimation, we derived a formula for asymptotic analysis of the minimum variational stochastic complexity.

9. REFERENCES

[1] S. Watanabe, *Algebraic Geometry and Statistical Learning Theory*, Cambridge University Press, 2009.

[2] K. Yamazaki and S. Watanabe, "Singularities in mixture models and upper bounds of stochastic complexity," *Neural Networks*, vol. 16, pp. 1029–1038, 2003.

[3] K. Yamazaki and S. Watanabe, "Stochastic complexity of Bayesian networks," in *Uncertainty in Artificial Intelligence*, 2003, pp. 592–599.

[4] K. Yamazaki and S. Watanabe, "Algebraic geometry and stochastic complexity of hidden Markov models," *Neurocomputing*, vol. 69, pp. 62–84, 2005.

[5] H. Attias, "Inferring parameters and structure of latent variable models by variational Bayes," in *Uncertainty in Artificial Intelligence*, 1999, pp. 21–30.

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

[7] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[8] K. Watanabe and S. Watanabe, "Stochastic complexities of Gaussian mixtures in variational Bayesian approximation," *Journal of Machine Learning Research*, vol. 7, pp. 625–644, 2006.

[9] K. Watanabe and S. Watanabe, "Stochastic complexities of general mixture models in variational Bayesian learning," *Neural Networks*, vol. 20, pp. 210–219, 2007.

[10] T. Hosino, K. Watanabe, and S. Watanabe, "Stochastic complexity of variational Bayesian hidden markov models," in *Proc. of IEEE International Joint Conference on Neural Networks*, 2005, vol. 2, pp. 1114–1119.

[11] K. Watanabe, M. Shiga, and S. Watanabe, "Upper bound for variational free energy of Bayesian networks," *Machine Learning*, vol. 75, pp. 199–215, 2009.

[12] T. Jaakkola and M. Jordan, "Bayesian parameter estimation via variational methods," *Statistics and Computing*, vol. 10, pp. 25–37, 2000.

[13] M. Seeger, "Bayesian inference and optimal design for the sparse linear model," *Journal of Machine Learning Research*, vol. 9, pp. 759–813, 2008.

[14] K.Watanabe, M.Okada, and K.Ikeda, "Information divergences in local variational approximation of Bayesian posterior distribution," in *Technical Report of IEICE*, 2010, vol. NC2009-138, pp. 297–302.

[15] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," *Journal of Machine Learning Research*, vol. 6, pp. 1705–1749, 2005.