Mach Learn (2012) 86:273–293
DOI 10.1007/s10994-011-5264-5

An alternative view of variational Bayes and asymptotic approximations of free energy

Kazuho Watanabe

Received: 11 February 2011 / Accepted: 21 September 2011 / Published online: 13 October 2011

Abstract Bayesian learning, widely used in many applied data-modeling problems, is often accomplished with approximation schemes because it requires intractable computation of the posterior distributions. In this study, we focus on two approximation methods, variational Bayes and local variational approximation. We show that the variational Bayes approach for statistical models with latent variables can be viewed as a special case of local variational approximation, where the log-sum-exp function is used to form the lower bound of the log-likelihood. The minimum variational free energy, the objective function of variational Bayes, is analyzed and related to the asymptotic theory of Bayesian learning. This analysis additionally implies a relationship between the generalization performance of the variational Bayes approach and the minimum variational free energy.

Keywords Variational Bayes · Local variational approximation · Variational free energy · Generalization error · Asymptotic analysis

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 hidden Markov model (HMM). More specifically, a formula for evaluating asymptotic forms of stochastic complexity or free energy was obtained and the generalization errors of statistical models have been intensively analyzed (Watanabe 2009; Yamazaki and Watanabe 2003a, 2003b, 2005; Rusakov and Geiger 2005; Aoyagi and Watanabe 2005; Yamazaki et al. 2010).

Editor: Kevin P. Murphy.

K. Watanabe (✉)
Graduate School of Information Science, Nara Institute of Science and Technology,
8916-5, Takayama-cho, Ikoma, Nara, 630-0192, Japan
e-mail: wkazuho@is.naist.jp

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, variational Bayes and local variational approximation, for Bayesian estimation. The former has been successfully applied to latent variable models such as mixture models and HMMs (Attias 1999; Beal 2003; Bishop 2006). Furthermore, its asymptotic analysis has progressed in several statistical models (Watanabe and Watanabe 2006, 2007; Hosino et al. 2005; Watanabe et al. 2009). The latter, also known as direct site bounding, has been applied to logistic regression (Jaakkola and Jordan 2000) and sparse linear models (Seeger 2008, 2009). This approximation is generally characterized and described by using the Bregman divergence (Watanabe et al. 2011).

In this paper, by providing a general framework for local variational approximation, we show that variational Bayes for the latent variable models can be interpreted as an application of local variational approximation. From this viewpoint, we investigate the asymptotic behavior of variational free energy, which is the objective function to be minimized by variational Bayes. More specifically, we present a formula for evaluating the asymptotic form of the minimum variational free energy relating it to the asymptotic theory of Bayesian estimation. This formula is applicable to general latent variable models and explains relationships between several previous works where asymptotic free energy and the minimum variational free energy have been analyzed respectively (Yamazaki and Watanabe 2003a, 2003b, 2005; Watanabe and Watanabe 2006, 2007; Hosino et al. 2005; Watanabe et al. 2009). We apply this formula to the GMM as an example and demonstrate another proof of the upper bound of the minimum variational free energy previously obtained in Watanabe and Watanabe (2006).

Furthermore, a byproduct of this analysis provides a quantity which is related to the generalization ability of the variational Bayesian approach. Analysis of generalization ability of a learning machine when it is used with the variational Bayesian approximation has been successful in quite limited cases (Nakajima and Watanabe 2007; Nakajima and Sugiyama 2010). We extend the asymptotic analysis of the minimum variational free energy (Watanabe 2010) and show an inequality which implies a relationship between the minimum variational free energy and the generalization error of the approximate predictive distribution. This relationship is also examined by a numerical experiment.

The rest of this paper is organized as follows. Section 2 describes Bayesian estimation and briefly introduces its asymptotic theory. Section 3 reviews variational Bayes for the latent variable models and the general framework for local variational approximation. Section 4 shows that a special case of the local variational approximation reduces to the variational Bayes approach for latent variable models. Section 5 presents the formula for the asymptotic analysis of the minimum variational free energy. Section 6 demonstrates its application to the GMM. Section 7 derives an inequality relating the minimum variational free energy and the generalization error of the variational Bayes approach. Discussion and conclusion follow in Sects. 8 and 9.

2 Bayesian learning

Assume we are given i.i.d. training examples or observations $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$ 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 $p(\boldsymbol{x}|\boldsymbol{w}) = \prod_{i=1}^n p(x_i|\boldsymbol{w})$. In this paper, we focus on the model $p(x|\boldsymbol{w})$ that is formulated by using a latent (unobserved) variable y. More

specifically, we consider the following model,

$$p(x|\boldsymbol{w}) = \sum_{y} p(x, y|\boldsymbol{w}), \tag{1}$$

which is obtained by marginalizing the joint distribution $p(x, y|\boldsymbol{w})$ of the model, that is, summing over all possible states of y. We assume a discrete latent variable y to include the examples such as the GMM and HMM.

Let $\boldsymbol{y} = \{y_1, y_2, \ldots, y_n\}$ be the latent (unobserved) variables corresponding to the observations $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$. Then, the likelihood function of the parameter $\boldsymbol{w}$ is expressed as,

$$p(\boldsymbol{x}|\boldsymbol{w}) = \sum_{\boldsymbol{y}} p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{w}), \tag{2}$$

where $p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{w}) \equiv \prod_{i=1}^{n} p(x_i, y_i|\boldsymbol{w})$ and $\sum_{\boldsymbol{y}}$ denotes the summation over all possible realizations of the latent variables.

By using the prior distribution $p_0(\boldsymbol{w})$, the Bayesian posterior distribution of the latent variables and parameter $\boldsymbol{w}$ is defined by

$$p(\boldsymbol{y}, \boldsymbol{w}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{w}) p_0(\boldsymbol{w})}{\sum_{\boldsymbol{y}} \int p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{w}) p_0(\boldsymbol{w})\, d\boldsymbol{w}}. \tag{3}$$

The normalizing constant,

$$Z(\boldsymbol{x}) \equiv \sum_{\boldsymbol{y}} \int p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{w}) p_0(\boldsymbol{w})\, d\boldsymbol{w} = \int p(\boldsymbol{x}|\boldsymbol{w}) p_0(\boldsymbol{w})\, d\boldsymbol{w}, \tag{4}$$

called the marginal likelihood or the evidence, is intractable since it requires the sum over exponentially many terms as in GMMs and HMMs and so is the posterior of the parameter,

$$p(\boldsymbol{w}|\boldsymbol{x}) = \sum_{\boldsymbol{y}} p(\boldsymbol{y}, \boldsymbol{w}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{w}) p_0(\boldsymbol{w})}{Z(\boldsymbol{x})}. \tag{5}$$

The negative logarithm of $Z(\boldsymbol{x})$,

$$F(\boldsymbol{x}) \equiv -\log Z(\boldsymbol{x}) \tag{6}$$

is termed the free energy or the stochastic complexity. This is a key quantity for model selection and is directly related to the average generalization error of the Bayesian predictive distribution as will be detailed in Sect. 7.1. Although it is an issue to compute or approximate the free energy practically, an asymptotic theory for analyzing the Bayesian free energy was established, of which we provide a brief overview.

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

$$S \equiv -\left\langle \log p(x|\boldsymbol{w}^*) \right\rangle_{p(x|\boldsymbol{w}^*)} \tag{7}$$

be its entropy.[1] For $p(\boldsymbol{x}|\boldsymbol{w}^*) = \prod_{i=1}^{n} p(x_i|\boldsymbol{w}^*)$, we define the (average) normalized free energy by

$$F^*(n) \equiv \left\langle F(\boldsymbol{x}) + \log p(\boldsymbol{x}|\boldsymbol{w}^*) \right\rangle_{p(\boldsymbol{x}|\boldsymbol{w}^*)} = \left\langle F(\boldsymbol{x}) \right\rangle_{p(\boldsymbol{x}|\boldsymbol{w}^*)} - nS. \tag{8}$$

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

Note that $F^*(n)$ is defined by the expectation of $F(\boldsymbol{x})$ over datasets generated by the true distribution $p(\boldsymbol{x}|\boldsymbol{w}^*)$ and hence is no longer a random variable. As a random variable, $F(\boldsymbol{x})$ has the leading term $-\log p(\boldsymbol{x}|\boldsymbol{w}^*)$, the expectation of which is nS. The above expression of $F^*(n)$ means that the expectation of $F(\boldsymbol{x})$ is $F^*(n) + nS$.

Then, it was proved that the average normalized Bayesian free energy has the following asymptotic form,

$$F^*(n) \simeq \lambda \log n - (m-1)\log\log n + O(1), \tag{9}$$

where the $O(1)$ term is bounded by a constant independent of n. The constants $-\lambda$ 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,

$$J_H(z) \equiv \int H(\boldsymbol{w})^z p_0(\boldsymbol{w})\, d\boldsymbol{w}, \tag{10}$$

where z is a complex number and

$$H(\boldsymbol{w}) \equiv \int p(x|\boldsymbol{w}^*) \log \frac{p(x|\boldsymbol{w}^*)}{p(x|\boldsymbol{w})}\, dx. \tag{11}$$

The free energy and the zeta function are related to the state density function of $H(\boldsymbol{w})$ by the Laplace and Mellin transforms respectively (Watanabe 2009). The asymptotic form (9) is then derived by the asymptotic expansion of the state density function.

In statistical models such as exponential families, 2λ is equal to the number of parameters and $m = 1$ (Schwarz 1978), whereas in latent variable models such as GMMs, 2λ is not larger than the number of parameters and $m \geq 1$. This means that the free energy of latent variable models deviates from the standard Bayesian Information Criterion (BIC) (Schwarz 1978). The asymptotic form (9) also plays an important role in assessing the approximation accuracy of the variational Bayes approach, which will be discussed in Sect. 8.2.2.

For several statistical models, the coefficient λ or its upper bound was evaluated by analyzing the pole of the zeta function (Yamazaki and Watanabe 2003a, 2003b, 2005; Rusakov and Geiger 2005; Aoyagi and Watanabe 2005; Watanabe 2009; Yamazaki et al. 2010). The condition that the true distribution is contained in the model is natural and essential for dealing with model selection problems, which is in fact assumed in these analyses (Schwarz 1978; Yamazaki and Watanabe 2003a, 2003b, 2005; Rusakov and Geiger 2005; Aoyagi and Watanabe 2005; Yamazaki et al. 2010).

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

In the variational Bayesian framework, the Bayesian posterior distribution (3) of the latent variables and the parameters is approximated by the variational posterior distribution $q(\boldsymbol{y}, \boldsymbol{w}|\boldsymbol{x})$, which factorizes as

$$q(\boldsymbol{y}, \boldsymbol{w}|\boldsymbol{x}) = q(\boldsymbol{y}|\boldsymbol{x})q(\boldsymbol{w}|\boldsymbol{x}), \tag{12}$$

where $q(\boldsymbol{y}|\boldsymbol{x})$ and $q(\boldsymbol{w}|\boldsymbol{x})$ are probability distributions on the latent 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 variational free energy. The variational free energy is defined in (13). We can express this as the sum of the Bayesian free energy and the Kullback information from the variational posterior $q(\boldsymbol{y}, \boldsymbol{w}|\boldsymbol{x})$ to the Bayesian posterior $p(\boldsymbol{y}, \boldsymbol{w}|\boldsymbol{x})$[2],

$$\overline{F}[q] \equiv \sum_{\boldsymbol{y}} \int q(\boldsymbol{y}, \boldsymbol{w}|\boldsymbol{x}) \log \frac{q(\boldsymbol{y}, \boldsymbol{w}|\boldsymbol{x})}{p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{w}) p_0(\boldsymbol{w})} d\boldsymbol{w} \tag{13}$$

$$= F(\boldsymbol{x}) + K(q(\boldsymbol{y}, \boldsymbol{w}|\boldsymbol{x})||p(\boldsymbol{y}, \boldsymbol{w}|\boldsymbol{x})). \tag{14}$$

This expression follows from the definitions of the free energy (6) and the posterior (3).

This formulation leads to the following alternate optimization over $q(\boldsymbol{y}|\boldsymbol{x})$ and $q(\boldsymbol{w}|\boldsymbol{x})$ (Attias 1999; Beal 2003; Bishop 2006). 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

$$q(\boldsymbol{w}|\boldsymbol{x}) = \frac{1}{C_w} p_0(\boldsymbol{w}) \exp\left\langle \log p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{w}) \right\rangle_{q(\boldsymbol{y}|\boldsymbol{x})}. \tag{15}$$

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

$$q(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{C_y} \exp\left\langle \log p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{w}) \right\rangle_{q(\boldsymbol{w}|\boldsymbol{x})}. \tag{16}$$

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

3.2 Local variational approximation

This section describes several facts regarding the local variational approximation (Bishop 2006; Watanabe et al. 2011).

Local variational approximation forms a lower bound of $p(\boldsymbol{w}, \boldsymbol{x}) = p(\boldsymbol{x}|\boldsymbol{w}) p_0(\boldsymbol{w})$, denoted by $\underline{p}_{\boldsymbol{\xi}}(\boldsymbol{w}, \boldsymbol{x})$,

$$\underline{p}_{\boldsymbol{\xi}}(\boldsymbol{w}, \boldsymbol{x}) \leq p(\boldsymbol{w}, \boldsymbol{x}), \tag{17}$$

and approximates the posterior distribution (5) by

$$p_{\boldsymbol{\xi}}(\boldsymbol{w}|\boldsymbol{x}) \equiv \frac{\underline{p}_{\boldsymbol{\xi}}(\boldsymbol{w}, \boldsymbol{x})}{\underline{Z}(\boldsymbol{\xi})}, \tag{18}$$

where $\underline{Z}(\boldsymbol{\xi}) \equiv \int \underline{p}_{\boldsymbol{\xi}}(\boldsymbol{w}, \boldsymbol{x}) d\boldsymbol{w}$, 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

[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,

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

since the inequality

$$\underline{Z}(\boldsymbol{\xi}) \leq Z(\boldsymbol{x}) \tag{19}$$

holds by definition. This is equivalent to the minimization of

$$\overline{F}(\boldsymbol{\xi}) \equiv -\log \underline{Z}(\boldsymbol{\xi}), \tag{20}$$

which is an upper bound of the free energy, $F(\boldsymbol{x}) = -\log Z(\boldsymbol{x})$.

Most existing local variational approximation techniques are based on the convexity of the log-likelihood function or the log-prior (Bishop 2006; Jaakkola and Jordan 2000; Seeger 2008, 2009). We characterize these cases by using a general convex function ϕ and the Bregman divergence associated with ϕ. Let ϕ be a twice differentiable real-valued convex function and $\boldsymbol{h}$ be a vector-valued function. Let us consider the case where the lower bound of the joint distribution is formed as follows,

$$p(\boldsymbol{w}, \boldsymbol{x}) = p(\boldsymbol{x}|\boldsymbol{w})p_0(\boldsymbol{w}) \geq p(\boldsymbol{x}|\boldsymbol{w})p_0(\boldsymbol{w})\exp\{-d_\phi(\boldsymbol{h}(\boldsymbol{w}), \boldsymbol{h}(\boldsymbol{\xi}))\} \equiv \underline{p}_{\boldsymbol{\xi}}(\boldsymbol{w}, \boldsymbol{x}), \tag{21}$$

where

$$d_\phi(\boldsymbol{u}, \boldsymbol{v}) \equiv \phi(\boldsymbol{u}) - \phi(\boldsymbol{v}) - (\boldsymbol{u} - \boldsymbol{v}) \cdot \nabla\phi(\boldsymbol{v}) \geq 0, \tag{22}$$

is the Bregman divergence associated with the convex function ϕ (Banerjee et al. 2005). The interpretation of the bound (21) is summarized as follows (Watanabe et al. 2011). The normalization of $p(\boldsymbol{w}, \boldsymbol{x})$ with respect to $\boldsymbol{w}$ is intractable. We can multiply by the exponential of the Bregman divergence to obtain a lower bound of $p(\boldsymbol{w}, \boldsymbol{x})$. We choose the convex function ϕ of some function $\boldsymbol{h}$ transforming $\boldsymbol{w}$ such that the intractable terms in $p(\boldsymbol{w}, \boldsymbol{x})$ are canceled, giving us a tractable lower bound $\underline{p}_{\boldsymbol{\xi}}(\boldsymbol{w}, \boldsymbol{x})$. For example, in the latent variable model, $\log p(\boldsymbol{w}, \boldsymbol{x})$ has the intractable term, $\log \sum_{\boldsymbol{y}} p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{w})$ originated from the log-likelihood, $\log p(\boldsymbol{x}|\boldsymbol{w})$. We choose ϕ to be such that this term is canceled as will be detailed in the next section. Watanabe et al. (2011) demonstrates an example of the lower bound (21) for the logistic regression model together with its upper bound variant.

Then, as for the free energy bound of the local variational approximation using the general bound (21) with the convex function ϕ, we obtain the following expression,

$$\overline{F}(\boldsymbol{\xi}) - F = \left\langle d_\phi(\boldsymbol{h}(\boldsymbol{w}), \boldsymbol{h}(\boldsymbol{\xi}))\right\rangle_{p_{\boldsymbol{\xi}}(\boldsymbol{w}|\boldsymbol{x})} + K(p_{\boldsymbol{\xi}}(\boldsymbol{w}|\boldsymbol{x})||p(\boldsymbol{w}|\boldsymbol{x})), \tag{23}$$

the derivation of which is in Appendix A.

From (21), the approximating posterior is given by,

$$p_{\boldsymbol{\xi}}(\boldsymbol{w}|\boldsymbol{x}) \propto \exp\{\boldsymbol{h}(\boldsymbol{w}) \cdot \nabla\phi(\boldsymbol{h}(\boldsymbol{\xi})) + \log p(\boldsymbol{x}, \boldsymbol{w}) - \phi(\boldsymbol{h}(\boldsymbol{w}))\}, \tag{24}$$

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

$$\boldsymbol{h}(\boldsymbol{\xi}) = \langle \boldsymbol{h}(\boldsymbol{w})\rangle_{p_{\tilde{\boldsymbol{\xi}}}(\boldsymbol{w}|\boldsymbol{x})} \tag{25}$$

is satisfied (Watanabe et al. 2011).

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 (5). By the convexity of the function $\log \sum_{\boldsymbol{y}} \exp(\cdot)$, the log-likelihood is bounded below as follows,

$$\begin{aligned}\log p(\boldsymbol{x}|\boldsymbol{w}) &= \log \sum_{\boldsymbol{y}} \exp\{\log p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{w})\} \\ &\geq \log p(\boldsymbol{x}|\boldsymbol{\xi}) + \sum_{\boldsymbol{y}} \left(\log \frac{p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{w})}{p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\xi})} \right) p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\xi}), \end{aligned} \tag{26}$$

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 $\boldsymbol{y}$. The inequality (26) is derived from lower bounding the log-sum-exp function ϕ by its tangent hyperplane at $\boldsymbol{h}(\boldsymbol{w}) = \boldsymbol{h}(\boldsymbol{\xi})$ and the fact that the gradient vector of ϕ consists of the elements $\frac{\partial \phi(\boldsymbol{h})}{\partial h_i} = \frac{\exp(h_i)}{\sum_j \exp(h_j)}$. We can see that the sum $\sum_{\boldsymbol{y}}$ is inside the logarithm in the left hand side of the inequality (26) while it is outside in the right hand side. Hence, replacing the left hand side, the intractable term in $p(\boldsymbol{w}, \boldsymbol{x})$, with the right hand side yields the tractable lower bound (21). These choices of ϕ and $\boldsymbol{h}$ yield the Bregman divergence,

$$d_{\phi}(\boldsymbol{h}(\boldsymbol{w}), \boldsymbol{h}(\boldsymbol{\xi})) = \sum_{\boldsymbol{y}} p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\xi}) \log \frac{p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\xi})}{p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w})} = K(p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\xi}) || p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w})), \tag{27}$$

which is also verified by subtracting the right hand side of the inequality (26) from the left hand side of it.

From (23), we have

$$\begin{aligned}\overline{F}(\boldsymbol{\xi}) &= F + K(p_{\boldsymbol{\xi}}(\boldsymbol{w}|\boldsymbol{x}) || p(\boldsymbol{w}|\boldsymbol{x})) + \langle K(p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\xi}) || p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w})) \rangle_{p_{\boldsymbol{\xi}}(\boldsymbol{w}|\boldsymbol{x})} \\ &= F + K(p_{\boldsymbol{\xi}}(\boldsymbol{w}|\boldsymbol{x}) p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\xi}) || p(\boldsymbol{w}, \boldsymbol{y}|\boldsymbol{x})), \end{aligned} \tag{28}$$

which is exactly the variational free energy (14) of the factorized distribution, $p_{\boldsymbol{\xi}}(\boldsymbol{w}|\boldsymbol{x}) \times p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\xi})$. In fact, from (24) and (26), the approximating posterior is given by

$$\begin{aligned} p_{\boldsymbol{\xi}}(\boldsymbol{w}|\boldsymbol{x}) &\propto \exp\left\{ \sum_{\boldsymbol{y}} \log p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{w}) p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\xi}) \right\} p_0(\boldsymbol{w}) \\ &= \exp \langle \log p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{w}) \rangle_{p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\xi})} \, p_0(\boldsymbol{w}). \end{aligned} \tag{29}$$

From (25), the EM update for $\boldsymbol{\xi}$ yields

$$\log p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\xi}) = \langle \log p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{w}) \rangle_{p_{\tilde{\boldsymbol{\xi}}}(\boldsymbol{w}|\boldsymbol{x})}, \tag{30}$$

which implies

$$p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\xi}) \propto \exp \langle \log p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{w}) \rangle_{p_{\tilde{\boldsymbol{\xi}}}(\boldsymbol{w}|\boldsymbol{x})}. \tag{31}$$

Equations (29) and (31) are exactly the same as the variational Bayes algorithm for minimizing the variational free energy over the factorized distributions, (15) and (16).

This view of the variational Bayes approach is partly mentioned in Jordan et al. (1999). It provides the interpretation of the variational free energy in the form, $\overline{F}(\boldsymbol{\xi}) = -\log \underline{Z}(\boldsymbol{\xi})$. This implies that the asymptotic analysis of the minimum variational free energy can be reduced to the formula developed for that of the Bayesian free energy, $F = -\log Z$. In the next section, based on this view, we relate the minimum variational free energy to the asymptotic analysis of the Bayesian free energy (Watanabe 2009).

5 Minimum variational free energy

Let

$$\overline{F}_{\min}(\boldsymbol{x}) \equiv \min_{\boldsymbol{h}(\boldsymbol{\xi})} \overline{F}(\boldsymbol{\xi}) \tag{32}$$

be the minimum variational free energy. Recall that the variational free energy (13) is minimized with respect to the distributions $q(\boldsymbol{w}|\boldsymbol{x})$ and $q(\boldsymbol{y}|\boldsymbol{x})$. Also note from (28) and (29) that the free energy bound $\overline{F}(\boldsymbol{\xi})$ depends on the variational parameter $\boldsymbol{\xi}$ only through the form of the posterior distribution of the latent variables, $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\xi})$ in the alternative view presented in the previous section. Hence, the minimization of the variational free energy is equivalent to that of the free energy bound $\overline{F}(\boldsymbol{\xi})$ with respect not only to $\boldsymbol{\xi}$ but directly to $\boldsymbol{h}(\boldsymbol{\xi})$ that has one-to-one mapping to $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\xi})$.

We assume that $p(x|\boldsymbol{w}^*)$ with the parameter $\boldsymbol{w}^*$ is the underlying distribution generating the data $\boldsymbol{x}$ independently and identically. Because of the non-identifiability of the latent variable model, the set of true parameters

$$W^* \equiv \left\{ \tilde{\boldsymbol{w}} | \sum_y p(x, y|\tilde{\boldsymbol{w}}) = p(x|\boldsymbol{w}^*) \right\}, \tag{33}$$

is not generally a point but can be a union of several manifolds with singularities (Watanabe 2009).

Since $\overline{F}_{\min}(\boldsymbol{x})$ is defined by the minimum over $\boldsymbol{h}(\boldsymbol{\xi})$, we obtain an upper bound for the minimum value by substituting a specific choice of $\boldsymbol{h}(\boldsymbol{\xi})$. For arbitrary $\tilde{\boldsymbol{w}}^* \in W^*$, by substituting $\boldsymbol{h}(\boldsymbol{\xi}) = \boldsymbol{h}(\tilde{\boldsymbol{w}}^*)$, we have

$$\begin{aligned}
\overline{F}_{\min}(\boldsymbol{x}) &= \min_{\boldsymbol{h}(\boldsymbol{\xi})} \left\{ -\log \int \underline{p}_{\xi}(\boldsymbol{w}, \boldsymbol{x})\, d\boldsymbol{w} \right\} \\
&= \min_{\boldsymbol{h}(\boldsymbol{\xi})} \left\{ -\log \sum_y p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\xi}) - \log \int \exp\left\{ \sum_y p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\xi}) \log \frac{p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{w})}{p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\xi})} \right\} p_0(\boldsymbol{w})\, d\boldsymbol{w} \right\} \qquad (34) \\
&\leq -\log p(\boldsymbol{x}|\boldsymbol{w}^*) - \log \int \exp\left\{ \sum_y p(\boldsymbol{y}|\boldsymbol{x}, \tilde{\boldsymbol{w}}^*) \log \frac{p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{w})}{p(\boldsymbol{x}, \boldsymbol{y}|\tilde{\boldsymbol{w}}^*)} \right\} p_0(\boldsymbol{w})\, d\boldsymbol{w} \\
&\equiv U(\boldsymbol{x}). \qquad (35)
\end{aligned}$$

Here, the second equality follows from the fact that the bound $\underline{p}_{\xi}(\boldsymbol{w}, \boldsymbol{x})$ is the exponential of the right hand side of the inequality (26) multiplied by the prior $p_0(\boldsymbol{w})$. In the last inequality, we have substituted $\boldsymbol{h}(\boldsymbol{\xi}) = \boldsymbol{h}(\tilde{\boldsymbol{w}}^*)$, that is, $p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\xi}) = p(\boldsymbol{x}, \boldsymbol{y}|\tilde{\boldsymbol{w}}^*)$. The expression (34) is also obtained by substituting the optimal form (15) of $q(\boldsymbol{w}|\boldsymbol{x})$ into (13) and identifying $q(\boldsymbol{y}|\boldsymbol{x})$ with $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\xi})$.

By subtracting the entropy of the true distribution, we define the (average) normalized minimum variational free energy and its upper bound by,

$$\overline{F}^*_{\min}(n) \equiv \left\langle \overline{F}_{\min}(\boldsymbol{x}) + \log p(\boldsymbol{x}|\boldsymbol{w}^*) \right\rangle_{p(\boldsymbol{x}|\boldsymbol{w}^*)} = \left\langle \overline{F}_{\min}(\boldsymbol{x}) \right\rangle_{p(\boldsymbol{x}|\boldsymbol{w}^*)} - nS, \tag{36}$$

and

$$U^*(n) \equiv \langle U(\boldsymbol{x}) \rangle_{p(\boldsymbol{x}|\boldsymbol{w}^*)} - nS. \tag{37}$$

Note here again that $\overline{F}_{\min}(\boldsymbol{x})$ and $U(\boldsymbol{x})$ are random variables with the leading term $-\log p(\boldsymbol{x}|\boldsymbol{w}^*)$, the expectations of which are $\overline{F}^*_{\min}(n) + nS$ and $U^*(n) + nS$ respectively.

Then, as is proved in Appendix B, the following inequality holds,

$$\overline{F}^*_{\min}(n) \leq -\log \int e^{-n\overline{H}(\boldsymbol{w})} p_0(\boldsymbol{w})\, d\boldsymbol{w} \equiv \overline{U}^*(n), \tag{38}$$

where

$$\overline{H}(\boldsymbol{w}) \equiv \int \sum_y p(x, y|\tilde{\boldsymbol{w}}^*) \log \frac{p(x, y|\tilde{\boldsymbol{w}}^*)}{p(x, y|\boldsymbol{w})}\, dx. \tag{39}$$

The asymptotic theory of the Bayesian estimation (Watanabe 2009) shows that the asymptotic form of the right hand side of (38), providing an upper bound of $\overline{F}^*_{\min}(n)$, is given by

$$\overline{U}^*(n) \simeq \overline{\lambda} \log n - (\overline{m} - 1) \log\log n + O(1), \tag{40}$$

where $-\overline{\lambda}$ and $\overline{m}$ are respectively the largest pole and its order of the zeta function defined for a complex number z by

$$J_{\overline{H}}(z) \equiv \int \overline{H}(\boldsymbol{w})^z p_0(\boldsymbol{w})\, d\boldsymbol{w}. \tag{41}$$

This means that the asymptotic behavior of the minimum variational free energy is characterized by $\overline{H}(\boldsymbol{w})$ while that of the free energy F is characterized by $H(\boldsymbol{w}) = K(p(x|\boldsymbol{w}^*)||p(x|\boldsymbol{w}))$ as in (9) and (10) (Watanabe 2009). These two functions are related by the log-sum inequality,

$$H(\boldsymbol{w}) \leq \overline{H}(\boldsymbol{w}). \tag{42}$$

Note that $\overline{H}(\boldsymbol{w})$ depends on the choice of $\tilde{\boldsymbol{w}}^* \in W^*$. For different $\tilde{\boldsymbol{w}}^*$, we have different values of $\overline{\lambda}$, which in (40) is determined by the minimum over different $\tilde{\boldsymbol{w}}^* \in W^*$. Then, $\overline{m}$ is determined by the maximum of the order of the pole for the minimum $\overline{\lambda}$.

6 Example: Gaussian mixture model

In this section, we derive an asymptotic upper bound of the minimum variational free energy of the GMM. Although this upper bound was obtained in a previous work (Watanabe and Watanabe 2006), the derivation was by direct evaluation and minimization of the variational free energy with respect to the expected sufficient statistics which corresponds to the variational parameter $\boldsymbol{\xi}$ in this paper. We present another derivation through (38) for an illustration of the asymptotic analysis described in Sect. 5.

6.1 Variational Bayes for GMM

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

$$p(x|\boldsymbol{w}) = \sum_{k=1}^{K} a_k g(x|\mu_k) \tag{43}$$

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 $0 \le a_k \le 1$ for $k = 1, \ldots, K$ 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

$$p(x, y|\boldsymbol{w}) = \prod_{k=1}^{K} \{a_k g(x|\mu_k)\}^{y^{(k)}}. \tag{44}$$

The latent variable $y = (y^{(1)}, y^{(2)}, \ldots, y^{(K)})$ 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 (Attias 1999; Beal 2003; Bishop 2006) using the prior distribution,

$$p_0(\boldsymbol{w}) \equiv p_0(\boldsymbol{a}) \prod_{k=1}^{K} p_0(\mu_k), \tag{45}$$

where

$$p_0(\boldsymbol{a}) \equiv \frac{\Gamma(K\alpha_0)}{\Gamma(\alpha_0)^K} \prod_{k=1}^{K} a_k^{\alpha_0 - 1} \tag{46}$$

is the Dirichlet distribution with hyperparameter $\alpha_0 > 0$ and

$$p_0(\mu_k) \equiv \sqrt{\frac{\beta_0}{2\pi}}^M \exp\left\{-\frac{\beta_0 \|\mu_k - \nu_0\|^2}{2}\right\} \tag{47}$$

is the Gaussian distribution with hyperparameters $\beta_0 > 0$ 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 minimum variational free energy

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

$$p(x|\boldsymbol{w}^*) = \sum_{k=1}^{K_0} a_k^* g(x|\mu_k^*), \tag{48}$$

and $K_0 \le K$ holds, that is, the true distribution is realizable by the postulated model. Then, it was proved in Watanabe and Watanabe (2006) that the normalized minimum variational free energy satisfies

$$\overline{F}^*_{\min}(n) \le \overline{\lambda} \log n + O(1), \tag{49}$$

where

$$\overline{\lambda} = \begin{cases} (K - K_0)\alpha_0 + \frac{MK_0 + K_0 - 1}{2} & (\alpha_0 \leq \frac{M+1}{2}), \\ \frac{MK + K - 1}{2} & (\alpha_0 > \frac{M+1}{2}). \end{cases} \tag{50}$$

It was empirically demonstrated that this upper bound is tight in some cases (Watanabe and Watanabe 2006). As will be discussed in Sect. 8.2.1, the inequality (49) holds for more general mixture components than the Gaussian $g(x|\mu)$.

Note that when $\alpha_0 > \frac{M+1}{2}$ or $K = K_0$, $2\overline{\lambda}$ is equal to the number of parameters and $\overline{\lambda} \log n$ turns out to be the penalty term in the BIC (Schwarz 1978). When $K_0 < K$, the set of true parameters, $W^* = \{\tilde{\boldsymbol{w}} | \sum_y p(x, y|\tilde{\boldsymbol{w}}) = p(x|\boldsymbol{w}^*)\}$, is not a point but a union of manifolds. Reflecting this fact, the coefficient $\overline{\lambda}$ indicates where in the parameter space the (approximate) posterior distribution converges depending on the value of the hyperparameter α_0.

6.3 Derivation of the upper bound

In this section, we derive (49) from (38), which provides another proof than that presented in Watanabe and Watanabe (2006). Similar proofs can be found in Yamazaki and Watanabe (2003a, 2003b, 2005) although they are intended for evaluating the Bayesian free energy (8) asymptotically. Here, we intend to evaluate the minimum variational free energy and present the details of the proof for the specific choice of the prior distribution (45) for the sake of self-containedness.

First, in order to define $p(x, y|\tilde{\boldsymbol{w}}^*)$ for y with K elements, we extend and redefine the true parameter $\boldsymbol{w}^*$ denoting it as $\tilde{\boldsymbol{w}}^* = \{\{\tilde{a}_k^*\}_{k=1}^K, \{\tilde{\mu}_k^*\}_{k=1}^K\}$. Suppose that the true distribution with parameter $\tilde{\boldsymbol{w}}^*$ has $\hat{K}$ non-zero mixing proportions. For example, we can take

$$\tilde{a}_k^* = \begin{cases} a_k^* & (1 \leq k \leq K_0 - 1), \\ a_{K_0}^* / (K - K_0 + 1) & (K_0 \leq k \leq \hat{K}), \\ 0 & (\hat{K} + 1 \leq k \leq K), \end{cases} \tag{51}$$

and

$$\tilde{\mu}_k^* = \begin{cases} \mu_k^* & (1 \leq k \leq K_0), \\ \mu_{K_0}^* & (K_0 + 1 \leq k \leq K). \end{cases} \tag{52}$$

Note that the marginal distribution of $p(x, y|\tilde{\boldsymbol{w}}^*)$ reduces to (48). Then, we have

$$\begin{aligned} \overline{H}(\boldsymbol{w}) &= \int \sum_y p(x, y|\tilde{\boldsymbol{w}}^*) \log \frac{p(x, y|\tilde{\boldsymbol{w}}^*)}{p(x, y|\boldsymbol{w})} \, dx \\ &= \int \sum_{k=1}^K \tilde{a}_k^* g(x|\tilde{\mu}_k^*) \log \frac{\tilde{a}_k^* g(x|\tilde{\mu}_k^*)}{a_k g(x|\mu_k)} \, dx \\ &= \sum_{k=1}^K \tilde{a}_k^* \left\{ \log \frac{\tilde{a}_k^*}{a_k} + \int g(x|\tilde{\mu}_k^*) \log \frac{g(x|\tilde{\mu}_k^*)}{g(x|\mu_k)} \, dx \right\} \\ &= \sum_{k=1}^{\hat{K}} \tilde{a}_k^* \left\{ \log \frac{\tilde{a}_k^*}{a_k} + \frac{\|\mu_k - \tilde{\mu}_k^*\|^2}{2} \right\}. \end{aligned} \tag{53}$$

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

$$\boldsymbol{w}_1 \equiv (a_2, a_3, \ldots, a_{\hat{K}}), \tag{54}$$

$$\boldsymbol{w}_2 \equiv (a_{\hat{K}+1}, \ldots, a_K), \tag{55}$$

$$\boldsymbol{w}_3 \equiv (\mu_1, \mu_2, \ldots, \mu_{\hat{K}}), \tag{56}$$

and define

$$W_1 \equiv \{\boldsymbol{w}_1 | \, |a_k - \tilde{a}_k^*| \leq \epsilon, 2 \leq k \leq \hat{K}\}, \tag{57}$$

$$W_2 \equiv \{\boldsymbol{w}_2 | \, |a_k| \leq \epsilon, \hat{K} \leq k \leq K\}, \tag{58}$$

$$W_3 \equiv \{\boldsymbol{w}_3 | \, \|\mu_k - \tilde{\mu}_k^*\| \leq \epsilon, 1 \leq k \leq \hat{K}\}, \tag{59}$$

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,

$$\overline{H}(\boldsymbol{w}) = \overline{H}_1(\boldsymbol{w}_1) + \overline{H}_2(\boldsymbol{w}_2) + \overline{H}_3(\boldsymbol{w}_3), \tag{60}$$

where

$$\overline{H}_1(\boldsymbol{w}_1) \equiv \sum_{k=2}^{\hat{K}} \tilde{a}_k^* \log \frac{\tilde{a}_k^*}{a_k} + \left(1 - \sum_{k=2}^{\hat{K}} \tilde{a}_k^*\right) \log \frac{1 - \sum_{k=2}^{\hat{K}} \tilde{a}_k^*}{1 - \sum_{k=2}^{\hat{K}} a_k}, \tag{61}$$

$$\overline{H}_2(\boldsymbol{w}_2) \equiv \frac{1}{1-c} \frac{1 - \sum_{k=2}^{K_0} \tilde{a}_k^*}{1 - \sum_{k=2}^{\hat{K}} a_k} \sum_{k=\hat{K}+1}^{K} a_k, \tag{62}$$

and

$$\overline{H}_3(\boldsymbol{w}_3) \equiv \sum_{k=1}^{\hat{K}} \frac{\tilde{a}_k^*}{2} \|\mu_k - \tilde{\mu}_k^*\|^2. \tag{63}$$

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

$$C_1 \sum_{k=2}^{\hat{K}} (a_k - \tilde{a}_k^*)^2 \leq \overline{H}_1(\boldsymbol{w}_1) \leq C_2 \sum_{k=2}^{\hat{K}} (a_k - \tilde{a}_k^*)^2, \tag{64}$$

and

$$C_3 \sum_{k=\hat{K}+1}^{K} a_k \leq \overline{H}_2(\boldsymbol{w}_2) \leq C_4 \sum_{k=\hat{K}+1}^{K} a_k. \tag{65}$$

Third, we evaluate the partial variational free energies defined, for $i = 1, 2, 3$, by

$$\overline{F}_i \equiv -\log \int_{W_i} \exp(-n\overline{H}_i(\boldsymbol{w}_i)) p_0(\boldsymbol{w}_i) \, d\boldsymbol{w}_i, \tag{66}$$

where $p_0(\boldsymbol{w}_i)$ is the product of the factors involving the corresponding parameters in (45).

It follows from (38), (60) and (66) that

$$\overline{F}^*_{\min}(n) \leq \overline{F}_1 + \overline{F}_2 + \overline{F}_3 + O(1). \tag{67}$$

From (64) and (63), as for $\overline{F}_1$ and $\overline{F}_3$, the Gaussian integration yields,

$$\overline{F}_1 = \frac{\hat{K}-1}{2}\log n + O(1), \tag{68}$$

and

$$\overline{F}_3 = \frac{M\hat{K}}{2}\log n + O(1). \tag{69}$$

Since

$$n^{\alpha_0}\int_0^{\epsilon} e^{-na_k} a_k^{\alpha_0-1}\, da_k \to \Gamma(\alpha_0) \quad (n\to\infty), \tag{70}$$

for $k = \hat{K}+1, \ldots, K$, it follows from (65),

$$\overline{F}_2 = (K-\hat{K})\alpha_0 \log n + O(1). \tag{71}$$

Finally, combining (68), (71), (69) and (67), we obtain

$$\overline{F}^*_{\min}(n) \leq \left\{(K-\hat{K})\alpha_0 + \frac{M\hat{K}+\hat{K}-1}{2}\right\}\log n + O(1). \tag{72}$$

Minimizing the right hand side of the above expression over $\hat{K}$ ($K_0 \leq \hat{K} \leq K$) leads to (49).

Alternatively, the above evaluations of all the partial variational free energies are obtained by using the zeta function method as mentioned in Sect. 5. For example, as for $\overline{F}_2$, the zeta function

$$J_{\overline{H}_2}(z) \equiv \int \overline{H}_2(\boldsymbol{w}_2)^z p_0(\boldsymbol{w}_2)\, d\boldsymbol{w}_2 \tag{73}$$

has a pole $z = -(K-\hat{K})\alpha_0$. This can be observed by the change of variables, the so-called blow-up,

$$a_k = a'_k a'_K \quad (k = \hat{K}+1, \ldots, K-1), \tag{74}$$

$$a_K = a'_K, \tag{75}$$

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

$$\int a_K'^z a_K'^{(K-\hat{K})\alpha_0-1} \tilde{J}_{\overline{H}_2}(\tilde{\boldsymbol{w}}'_2)\, da'_K = \frac{\tilde{J}_{\overline{H}_2}(\tilde{\boldsymbol{w}}'_2)}{z+(K-\hat{K})\alpha_0}, \tag{76}$$

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

$$\int \left(\sum_{k=\hat{K}+1}^{K-1} a'_k + 1\right)^z \prod_{k=\hat{K}+1}^{K-1} a_k'^{\alpha_0-1} \prod_{k=\hat{K}+1}^{K-1} da'_k. \tag{77}$$

Hence, we can see that $J_{\overline{H}_2}$ has a pole $z = -(K-\hat{K})\alpha_0$.

7 Variational free energy and generalization error

In this section, we focus on the relationship between the variational free energy and the generalization ability of the variational Bayes approach. We denote the dataset by $\boldsymbol{x}^n = \{x_1, x_2, \ldots, x_n\}$ indicating the number of data explicitly only in this section.

7.1 Relationship between variational free energy and generalization error

Let $p(x, y|\tilde{\boldsymbol{w}}^*)$ be the true distribution of the observed variable x and the latent variable y which has the marginal distribution $p(x|\boldsymbol{w}^*)$. We define by

$$\overline{G}^*(\boldsymbol{x}^n) \equiv K(p(x, y|\tilde{\boldsymbol{w}}^*)||\tilde{p}^*(x, y|\boldsymbol{x}^n)), \tag{78}$$

the generalization error of the predictive distribution,

$$\tilde{p}^*(x, y|\boldsymbol{x}^n) \equiv \langle p(x, y|\boldsymbol{w})\rangle_{p_{\tilde{\boldsymbol{w}}^*}(\boldsymbol{w}|\boldsymbol{x}^n)} = \int p(x, y|\boldsymbol{w})\, p_{\tilde{\boldsymbol{w}}^*}(\boldsymbol{w}|\boldsymbol{x}^n)\, d\boldsymbol{w}, \tag{79}$$

where $p_{\tilde{\boldsymbol{w}}^*}(\boldsymbol{w}|\boldsymbol{x}^n)$ is the approximating posterior distribution (24) with $\boldsymbol{h}(\tilde{\boldsymbol{w}}^*)$ substituted for $\boldsymbol{h}(\boldsymbol{\xi})$. We denote its mean by

$$\overline{G}^*(n) \equiv \langle \overline{G}^*(\boldsymbol{x}^n)\rangle_{\prod_{i=1}^n p(x_i|\boldsymbol{w}^*)}. \tag{80}$$

Then, the following inequality holds,

$$U^*(n+1) - U^*(n) \geq \overline{G}^*(n), \tag{81}$$

where $U^*(n)$ is the upper bound (37) of the minimum variational free energy. The proof is put in Appendix C.

The inequality (81) is analogous to the equality,

$$F^*(n+1) - F^*(n) = G(n), \tag{82}$$

which holds for the average free energy (8) and the generalization error of the Bayesian predictive distribution,

$$G(n) \equiv \left\langle K(p(x|\boldsymbol{w}^*)||p(x|\boldsymbol{x}^n))\right\rangle_{\prod_{i=1}^n p(x_i|\boldsymbol{w}^*)}, \tag{83}$$

where

$$p(x|\boldsymbol{x}^n) \equiv \langle p(x|\boldsymbol{w})\rangle_{p(\boldsymbol{w}|\boldsymbol{x}^n)}. \tag{84}$$

If $U^*(n)$ has the asymptotic form $U^*(n) \simeq \overline{\lambda}\log n + O(1)$ as in (40) for $\overline{U}^*(n)$, the inequality (81) suggests that

$$\overline{G}^*(n) \leq \frac{\overline{\lambda}}{n} + o\left(\frac{1}{n}\right). \tag{85}$$

This means that the coefficient $\overline{\lambda}$ of the leading term of $U^*(n)$ is directly related to the generalization error of the variational Bayes approach measured by (78).

In the true sense, the generalization error of the variational Bayes approach should be evaluated for the predictive distribution $\langle p(x, y|\boldsymbol{w})\rangle_{p_{\hat{\boldsymbol{\xi}}_n}(\boldsymbol{w}|\boldsymbol{x}^n)}$ where $\boldsymbol{h}(\hat{\boldsymbol{\xi}}_n) =$

$\mathrm{argmin}_{\boldsymbol{h}(\boldsymbol{\xi})}\, \overline{F}(\boldsymbol{\xi})$.[3] Although the predictive distribution (79) consists of the true parameter $\tilde{\boldsymbol{w}}^*$ instead of $\hat{\boldsymbol{\xi}}_n$, it still depends on the samples $\boldsymbol{x}^n$ and inherits some property of the variational Bayesian approximation. In the next subsection, we will empirically examine the difference between $\overline{G}^*(n)$ and the generalization error in the true sense (as will be defined in (86)). At least, the inequality (81) implies the affinity of the minimum variational free energy and the generalization error measured by the Kullback information of the joint distributions.

7.2 Numerical experiment

We empirically examine the relationship between the generalization error and the asymptotic form of the minimum variational free energy and demonstrate that the asymptotic form (49) partly describes the generalization error of the variational Bayes approach for GMM.

We implemented the variational Bayesian learning of GMM with K components (43). For simplicity, we chose the true distribution to be the standard normal distribution in R^2, $g(x|(0,0)^T)$. According to the choice of $\tilde{\boldsymbol{w}}^*$ for evaluating $\overline{\lambda}$ in (49), we consider this distribution as the choice, $\tilde{a}_1^* = 1$, $\tilde{a}_k^* = 0$ for $k = 2, \ldots, K$, $\tilde{\mu}_k^* = (0,0)^T$ for $k = 1, 2, \ldots, K$ and focus on the case where $\alpha_0 < (M+1)/2 = 1.5$. Note that for this choice, the (Jensen's) inequality used to derive the inequality (81) holds with equality since $p(y|x_{n+1}, \tilde{\boldsymbol{w}}^*)$ is either 1 or 0.

Samples of the size $n = 100$ were generated by the true distribution. The variational Bayes algorithm was executed 21 times with 20 different random initializations and the one from the true parameter $\tilde{\boldsymbol{w}}^*$. We adopted the estimate $p(\boldsymbol{y}|\boldsymbol{x}^n, \hat{\boldsymbol{\xi}}_n)$ that attained the minimum of the variational free energy and evaluated the generalization error,

$$\overline{G}(\boldsymbol{x}^n) \equiv \sum_y \int p(x, y|\tilde{\boldsymbol{w}}^*) \log \frac{p(x, y|\tilde{\boldsymbol{w}}^*)}{\tilde{p}(x, y|\boldsymbol{x}^n)}\, dx, \tag{86}$$

where

$$\tilde{p}(x, y|\boldsymbol{x}^n) \equiv \langle p(x, y|\boldsymbol{w}) \rangle_{p_{\hat{\boldsymbol{\xi}}_n}(\boldsymbol{w}|\boldsymbol{x}^n)} \tag{87}$$

is the (approximate) predictive distribution.

In the above-mentioned case, the predictive distribution is given by the GMM,

$$\tilde{p}(x, y|\boldsymbol{x}^n) = \prod_{k=1}^{K} \left\{ \frac{\overline{a}_k}{\sqrt{2\pi(1+\overline{\sigma}_k^2)}^M} \exp\left(-\frac{\|x - \overline{\mu}_k\|^2}{2(1+\overline{\sigma}_k^2)} \right) \right\}^{y^{(k)}}. \tag{88}$$

Here $\overline{a}_k = \frac{n_k + \alpha_0}{n + K\alpha_0}$ and $\overline{\mu}_k = \frac{n_k \nu_k + \beta_0 \nu_0}{n_k + \beta_0}$ are (approximate) posterior means of a_k and μ_k where $n_k = \sum_{i=1}^n \langle y_i^{(k)} \rangle_{p(y_i|x_i, \hat{\boldsymbol{\xi}}_n)}$ and $\nu_k = \frac{1}{n_k} \sum_{i=1}^n \langle y_i^{(k)} \rangle_{p(y_i|x_i, \hat{\boldsymbol{\xi}}_n)} x_i$. $\overline{\sigma}_k^2 = 1/(n_k + \beta_0)$ is the posterior variance of μ_k. The generalization error is evaluated as

$$\overline{G}(\boldsymbol{x}^n) = \sum_{k=1}^{K} \tilde{a}_k^* \left\{ \log \frac{\tilde{a}_k^*}{\overline{a}_k} + \frac{M}{2} \log(1+\overline{\sigma}_k^2) - \frac{M}{2} \frac{\overline{\sigma}_k^2}{1+\overline{\sigma}_k^2} + \frac{\|\overline{\mu}_k - \tilde{\mu}_k^*\|^2}{2(1+\overline{\sigma}_k^2)} \right\}. \tag{89}$$

[3] In practice, we optimize $p(\boldsymbol{y}|\boldsymbol{x}^n, \hat{\boldsymbol{\xi}}_n)$ directly instead of calculating $\hat{\boldsymbol{\xi}}_n$ explicitly as noted at the beginning of Sect. 5.

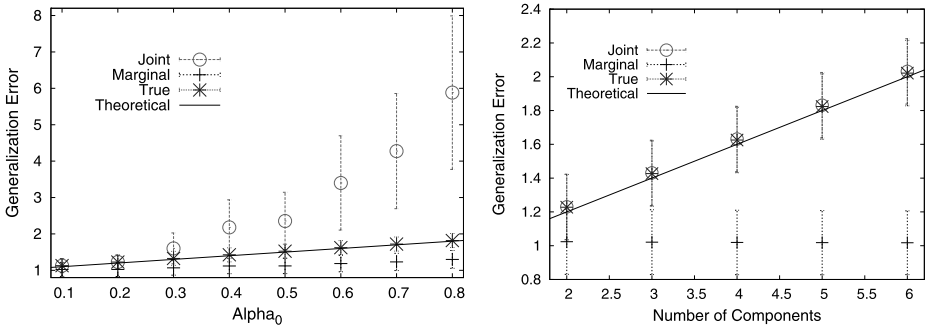

Fig. 1 Average generalization errors for $K = 2$ and different α_0 (*left*) and for $\alpha_0 = 0.2$ and different K (*right*) with 95%-confidence intervals. ∘: Average errors of the joint distribution (89). +: Average errors of the marginal distribution (91). ∗: Average errors of the joint distribution with the variational parameter substituted by the true one (78). *Solid line*: Theoretical values of the average error (90). The generalization errors are multiplied by $n = 100$ for scaling purposes

To investigate the difference between $\overline{G}(\boldsymbol{x}^n)$ and $\overline{G}^*(\boldsymbol{x}^n)$ introduced in Sect. 7.1, we also evaluated $\overline{G}^*(\boldsymbol{x}^n)$. Let $n_k^* = \sum_{i=1}^n \langle y_i^{(k)} \rangle_{p(y|x_i,\tilde{\boldsymbol{w}}^*)}$ and $\nu_k^* = \frac{1}{n_k^*} \sum_{i=1}^n \langle y_i^{(k)} \rangle_{p(y|x_i,\tilde{\boldsymbol{w}}^*)} x_i$ for $k = 1, 2, \ldots, K$. For the above choice of $\tilde{\boldsymbol{w}}^*$, we have $n_1^* = n$, $n_k^* = 0$ for $k = 2, \ldots, K$, $\nu_1^* = \frac{1}{n} \sum_{i=1}^n x_i$ and $\nu_k^* = 0$ for $k = 2, \ldots, K$. Since $\nu_1^* = \frac{1}{n} \sum_{i=1}^n x_i$ obeys the normal distribution with mean $(0, 0)^T$ and covariance matrix $\frac{1}{n} I$ where I is the unit matrix, we can show that

$$\overline{G}^*(n) \simeq \left\{ \frac{M}{2} + (K-1)\alpha_0 \right\} \frac{1}{n} + o\left(\frac{1}{n}\right), \tag{90}$$

by putting $\{n_k^*, \nu_k^*\}$ into $\{\overline{a}_k, \overline{\mu}_k\}$ and evaluating the expectation of (89) with respect to the true distribution $\prod_{i=1}^n p(x_i|\boldsymbol{w}^*)$. Note that the coefficient $\frac{M}{2} + (K-1)\alpha_0$ is exactly equal to $\overline{\lambda}$ in the inequality (49) for the case where $K_0 = 1$ and $\alpha_0 < \frac{M+1}{2}$. This means that the inequality (85) is tight in this case.

Additionally, we calculated the generalization error of the marginal distribution,

$$G(\boldsymbol{x}^n) \equiv \int p(x|\boldsymbol{w}^*) \log \frac{p(x|\boldsymbol{w}^*)}{\tilde{p}(x|\boldsymbol{x}^n)} \, dx \tag{91}$$

where

$$\tilde{p}(x|\boldsymbol{x}^n) \equiv \langle p(x|\boldsymbol{w}) \rangle_{p_{\hat{\xi}_n}(\boldsymbol{w}|\boldsymbol{x}^n)} \tag{92}$$

is the marginal predictive distribution. We evaluated the expectation with respect to $p(x|\boldsymbol{w}^*)$ by generating 100000 test samples from the same true distribution. The generalization error (89) of the joint distribution is always larger than $G(\boldsymbol{x}^n)$.

Figure 1 shows the generalization errors averaged over 100 trials with different data sets. Figure 1 (left) is for the case of $K = 2$ with different values of the hyperparameter α_0. As expected, the average of $\overline{G}^*(\boldsymbol{x}^n)$ fits the theoretical line (90) well. We can see that for small α_0, the behavior of the generalization error of the joint predictive distribution is well described by that of $\overline{G}^*(n)$ and hence by the coefficient $\overline{\lambda}$ in the upper bound (49). As α_0 tends larger, the average of $\overline{G}(\boldsymbol{x}^n)$ also increases, as does that of the generalization error $G(\boldsymbol{x}^n)$ of the marginal distribution, although only slightly. This may be caused by overfitting. Figure 1 (right) shows the average of the generalization errors for the case of

$\alpha_0 = 0.2$ with different number K of components. Again, we can see that for small α_0 the generalization error of the joint predictive distribution is described by $\overline{\lambda}$ in (49) while the generalization error of the marginal distribution stays constant even when the model becomes more redundant.

8 Discussion

We presented a formula for analyzing the asymptotic behavior of the minimum variational free energy in Sect. 5 and demonstrated its application to the GMM in Sect. 6. In this section, we discuss extension of the main results to other approximation schemes than the variational Bayes approach. Then, we discuss the relationship to previous works where the free energy and the minimum variational free energy were analyzed for specific latent variable models.

8.1 Extension to other Bregman divergences

In Sect. 4, we showed that the local variational approximation with the log-sum-exp function derives an alternative view of the variational Bayes approach for latent variable models. A similar argument to that in Sect. 5 may help in evaluating the asymptotic forms of free energy approximations by the local variational approximations for other convex functions than the log-sum-exp function.

Generally, the local variational approximation forms upper and lower bounds of the free energy (Watanabe et al. 2011),

$$\overline{F}(\boldsymbol{\xi}) \equiv -\log \int p(\boldsymbol{x}|\boldsymbol{w}) \exp\{-d_\phi(\boldsymbol{h}(\boldsymbol{w}), \boldsymbol{h}(\boldsymbol{\xi}))\} p_0(\boldsymbol{w})\, d\boldsymbol{w}, \tag{93}$$

$$\underline{F}(\boldsymbol{\eta}) \equiv -\log \int p(\boldsymbol{x}|\boldsymbol{w}) \exp\{d_\psi(\boldsymbol{g}(\boldsymbol{w}), \boldsymbol{g}(\boldsymbol{\eta}))\} p_0(\boldsymbol{w})\, d\boldsymbol{w}, \tag{94}$$

where $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ are the respective variational parameters and $\boldsymbol{h}$ and $\boldsymbol{g}$ are functions transforming the parameter $\boldsymbol{w}$ to utilize the convexity of the functions ϕ and ψ. For $\boldsymbol{w}^* \in W^* = \{\boldsymbol{w}|p(\boldsymbol{x}|\boldsymbol{w}) = p(\boldsymbol{x}|\boldsymbol{w}^*)\}$, let

$$\overline{H}(\boldsymbol{w}) \equiv H(\boldsymbol{w}) + \frac{1}{n}\langle d_\phi(\boldsymbol{h}(\boldsymbol{w}), \boldsymbol{h}(\tilde{\boldsymbol{w}}^*))\rangle_{p(\boldsymbol{x}|\boldsymbol{w}^*)}, \tag{95}$$

$$\underline{H}(\boldsymbol{w}) \equiv H(\boldsymbol{w}) - \frac{1}{n}\langle d_\psi(\boldsymbol{g}(\boldsymbol{w}), \boldsymbol{g}(\tilde{\boldsymbol{w}}^*))\rangle_{p(\boldsymbol{x}|\boldsymbol{w}^*)}, \tag{96}$$

where $H(\boldsymbol{w})$ is the Kullback information (11).

As we have demonstrated for the case ϕ to be the log-sum-exp function, asymptotic forms of the free energy approximations can be evaluated by

$$\overline{U}^*(n) \equiv -\log \int e^{-n\overline{H}(\boldsymbol{w})} p_0(\boldsymbol{w})\, d\boldsymbol{w}, \tag{97}$$

and

$$\underline{L}^*(n) \equiv -\log \int e^{-n\underline{H}(\boldsymbol{w})} p_0(\boldsymbol{w})\, d\boldsymbol{w}. \tag{98}$$

Corresponding zeta functions can be used for evaluating $\overline{U}^*(n)$ and $\underline{L}^*(n)$ asymptotically. Note that $\underline{L}^*(n)$ does not necessarily provide a lower bound for the mean of $\max_{\boldsymbol{g}(\boldsymbol{\eta})} \underline{F}(\boldsymbol{\eta})$

since $-\log \int \exp(\cdot) p_0(\boldsymbol{w})\, d\boldsymbol{w}$ is concave. However, it may be useful for investigating the average property of the approximation scheme.

8.2 Relation to previous works

8.2.1 Asymptotic analysis of free energy bounds

Asymptotic upper bounds of the free energy were obtained for some statistical models including the GMM, HMM and the Bayesian network (Yamazaki and Watanabe 2003a, 2003b, 2005). The upper bounds are given in such forms as

$$F^*(n) \leq \nu \log n + O(1), \tag{99}$$

where the coefficient ν was identified for each model by analyzing the largest pole of the zeta function J_H in (10). More specifically, however, these works analyzed the largest pole of $J_{\overline{H}}$ in (41) instead of J_H by using the log-sum inequality (42) (Yamazaki and Watanabe 2003a, 2003b, 2005). 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 free energy were analyzed also for the same models (Watanabe and Watanabe 2006, 2007; Hosino et al. 2005; Watanabe et al. 2009), each of which has the form

$$\overline{F}^*_{\min}(n) \leq \overline{\lambda} \log n + O(1). \tag{100}$$

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

Moreover, the previous analyses of the minimum variational free energy were based on the direct minimization of the variational free energy 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 (38) is more general and is independent of the concrete algorithm for the specific model. It does not even require that the prior distribution $p_0(\boldsymbol{w})$ be conjugate prior although in this case the variational Bayes algorithm turns out not to be practical. In fact, in the case of the mixture model, the upper bound (49) is obtained in more general cases. The mixture component $g(x|\mu)$ can be generalized as long as $K(g(x|\tilde{\mu})||g(x|\mu))$ can be approximated by $(\mu - \tilde{\mu})^T J(\mu - \tilde{\mu})$ for some positive definite matrix J while in Watanabe and Watanabe (2007) it was generalized only to the exponential family.

8.2.2 Accuracy of approximation

For several statistical models, tighter bounds or exact evaluation of the coefficient λ of the free energy (9) has been obtained recently (Aoyagi and Watanabe 2005; Yamazaki et al. 2010). If the free energy and the minimum variational free energy have the asymptotic forms, $F^*(n) \simeq \lambda \log n + o(\log n)$ and $\overline{F}^*_{\min}(n) \simeq \overline{\lambda} \log n + o(\log n)$ respectively and $\lambda < \overline{\lambda}$, the approximation accuracy of the variational Bayes approach is evaluated as

$$\overline{F}^*_{\min}(n) - F(n) = (\overline{\lambda} - \lambda) \log n + o(\log n). \tag{101}$$

This turns out to be the Kullback information from the approximating posterior to the true posterior from (14). Such a comparison was first done for GMMs (Watanabe and Watanabe 2006). It was proved that $\overline{\lambda}$ can be strictly greater than λ while $\overline{\lambda}$ is not so large as the number of parameters divided by two, as in the BIC. The results presented in Sect. 5 imply that such a comparison can be done for general latent variable models by examining the difference between $\inf_{\tilde{\boldsymbol{w}}^* \in W^*} \overline{U}(n)$ and $F^*(n)$, which corresponds to that between the poles of $J_{\overline{H}}$ and J_H. The discussion in Sect. 8.1 implies the possibility of assessing the approximation accuracy of other approximation schemes in a similar way.

8.2.3 Average generalization error

Previous results on the variational Bayesian approximation have mentioned little about its generalization ability except for a few limited models such as the reduced rank regression and the matrix factorization models (Nakajima and Watanabe 2007; Nakajima and Sugiyama 2010). In Sect. 7.1, we derived an inequality which implies the relationship between the generalization error and the variational free energy for general latent variable models. In Sect. 7.2, we empirically demonstrated that the coefficient of the minimum variational free energy partly describes the behavior of the generalization error. Thorough investigation of the generalization ability of the variational Bayes algorithm including the case for large α_0 and for the marginal predictive distribution will be left for future work.

In the original (not approximate) Bayesian estimation, the universal relation among the quartet, Bayes and Gibbs generalization errors and Bayes and Gibbs training errors, was proved (Watanabe 2009). It is an important undertaking to explore such relationships among the quantities introduced in Sect. 7 for the approximate Bayesian estimation.

9 Conclusion

In this paper, we provided an alternative view of variational Bayes for latent variable models as an application of local variational approximation. Combining this view with the asymptotic theory of Bayesian estimation, we derived a formula for asymptotic analysis of the minimum variational free energy. As a byproduct of this formula, we also obtained an inequality that relates the minimum variational free energy to the generalization error.

It is an important undertaking to elucidate the condition under which the upper bound evaluated by the formula gives the exact asymptotic form of the minimum variational free energy. The approach presented in this paper is applicable for evaluating the asymptotic approximation accuracy of other models and other choices of the convex function, that is, approximation scheme. These will be pursued in the future.

Acknowledgements We would like to thank the anonymous reviewers for their helpful comments to improve the manuscript. In this research, we used the supercomputer of ACCMS, Kyoto University.

Appendix A: Derivation of (23)

From the definitions of $Z(\boldsymbol{x})$, $\underline{Z}(\boldsymbol{\xi})$ and $\underline{p}_{\boldsymbol{\xi}}(\boldsymbol{w}, \boldsymbol{x})$, we have

$$\begin{aligned}\overline{F}(\boldsymbol{\xi}) - F &= \log \frac{Z(\boldsymbol{x})}{\underline{Z}(\boldsymbol{\xi})} = \log \left(\frac{p(\boldsymbol{w}, \boldsymbol{x})}{p(\boldsymbol{w}|\boldsymbol{x})} \frac{p_{\boldsymbol{\xi}}(\boldsymbol{w}|\boldsymbol{x})}{\underline{p}_{\boldsymbol{\xi}}(\boldsymbol{w}, \boldsymbol{x})} \right) \\ &= \log \left(\frac{p_{\boldsymbol{\xi}}(\boldsymbol{w}|\boldsymbol{x})}{p(\boldsymbol{w}|\boldsymbol{x})} \right) - d_{\phi}(\boldsymbol{h}(\boldsymbol{w}), \boldsymbol{h}(\boldsymbol{\xi})). \end{aligned} \tag{102}$$

Taking the expectations of both sides with respect to $p_{\xi}(\boldsymbol{w}|\boldsymbol{x})$ yields (23).

Appendix B: Proof of the inequality (38)

$$
\begin{aligned}
\overline{F}^{*}_{\min}(n) &\leq U^{*}(n) \\
&= \langle U(\boldsymbol{x}) \rangle_{p(\boldsymbol{x}|\boldsymbol{w}^{*})} - nS \\
&= -\left\langle \log \int \exp\left\{ \sum_{y} p(y|\boldsymbol{x}, \tilde{\boldsymbol{w}}^{*}) \log \frac{p(\boldsymbol{x}, y|\boldsymbol{w})}{p(\boldsymbol{x}, y|\tilde{\boldsymbol{w}}^{*})} \right\} p_0(\boldsymbol{w})\, d\boldsymbol{w} \right\rangle_{p(\boldsymbol{x}|\boldsymbol{w}^{*})} \\
&\leq -\log \int \exp\left\{ \left\langle \sum_{y} p(y|\boldsymbol{x}, \tilde{\boldsymbol{w}}^{*}) \log \frac{p(\boldsymbol{x}, y|\boldsymbol{w})}{p(\boldsymbol{x}, y|\tilde{\boldsymbol{w}}^{*})} \right\rangle_{p(\boldsymbol{x}|\boldsymbol{w}^{*})} \right\} p_0(\boldsymbol{w})\, d\boldsymbol{w} \\
&= -\log \int e^{-n\overline{H}(\boldsymbol{w})} p_0(\boldsymbol{w})\, d\boldsymbol{w}. \qquad (103)
\end{aligned}
$$

The first inequality follows from the fact, $\overline{F}_{\min}(\boldsymbol{x}) \leq U(\boldsymbol{x})$. The first equality is the definition of $U^{*}(n)$. The second equality follows from the definitions of $U(\boldsymbol{x})$ and the entropy S. We have applied Jensen's inequality to the convex function $\log \int \exp(\cdot) p_0(\boldsymbol{w})\, d\boldsymbol{w}$ to obtain the last inequality. Finally, the last equality follows from the fact that $p(\boldsymbol{x}|\boldsymbol{w}^{*}) p(y|\boldsymbol{x}, \tilde{\boldsymbol{w}}^{*}) = p(\boldsymbol{x}, y|\tilde{\boldsymbol{w}}^{*})$ and the i.i.d. assumption.

Appendix C: Proof of the inequality (81)

$$
\begin{aligned}
&U^{*}(\boldsymbol{x}^{n+1}) - U^{*}(\boldsymbol{x}^{n}) \\
&= -\log \frac{\int \prod_{i=1}^{n+1} \exp\{\sum_{y} p(y|x_i, \tilde{\boldsymbol{w}}^{*}) \log \frac{p(x_i, y|\boldsymbol{w})}{p(x_i, y|\tilde{\boldsymbol{w}}^{*})}\} p_0(\boldsymbol{w})\, d\boldsymbol{w}}{\int \prod_{i=1}^{n} \exp\{\sum_{y} p(y|x_i, \tilde{\boldsymbol{w}}^{*}) \log \frac{p(x_i, y|\boldsymbol{w})}{p(x_i, y|\tilde{\boldsymbol{w}}^{*})}\} p_0(\boldsymbol{w})\, d\boldsymbol{w}} \\
&= -\log \int \exp\left\{ \sum_{y} p(y|x_{n+1}, \tilde{\boldsymbol{w}}^{*}) \log \frac{p(x_{n+1}, y|\boldsymbol{w})}{p(x_{n+1}, y|\tilde{\boldsymbol{w}}^{*})} \right\} p_{\tilde{\boldsymbol{w}}^{*}}(\boldsymbol{w}|\boldsymbol{x}^{n})\, d\boldsymbol{w} \\
&= \sum_{y} p(y|x_{n+1}, \tilde{\boldsymbol{w}}^{*}) \log p(x_{n+1}, y|\tilde{\boldsymbol{w}}^{*}) \\
&\quad - \log \int \exp\left\{ \langle \log p(x_{n+1}, y|\boldsymbol{w}) \rangle_{p(y|x_{n+1}, \tilde{\boldsymbol{w}}^{*})} \right\} p_{\tilde{\boldsymbol{w}}^{*}}(\boldsymbol{w}|\boldsymbol{x}^{n})\, d\boldsymbol{w} \\
&\geq \sum_{y} p(y|x_{n+1}, \tilde{\boldsymbol{w}}^{*}) \log \frac{p(x_{n+1}, y|\tilde{\boldsymbol{w}}^{*})}{\langle p(x_{n+1}, y|\boldsymbol{w}) \rangle_{p_{\tilde{\boldsymbol{w}}^{*}}(\boldsymbol{w}|\boldsymbol{x}^{n})}}. \qquad (104)
\end{aligned}
$$

In the last inequality, we have applied Jensen's inequality due to the convexity of the function $\log \int \exp(\cdot) p(\boldsymbol{w})\, d\boldsymbol{w}$. Taking expectation with respect to $\prod_{i=1}^{n+1} p(x_i|\boldsymbol{w}^{*})$ in both sides of the above inequality yields the inequality (81).

References

Aoyagi, M., & Watanabe, S. (2005). Stochastic complexities of reduced rank regression in Bayesian estimation. *Neural Networks*, *18*, 924–933.

Attias, H. (1999). Inferring parameters and structure of latent variable models by variational Bayes. In *Uncertainty in artificial intelligence* (pp. 21–30).

Banerjee, A., Merugu, S., Dhillon, I. S., & Ghosh, J. (2005). Clustering with Bregman divergences. *Journal of Machine Learning Research*, *6*, 1705–1749.

Beal, M. J. (2003). *Variational algorithms for approximate Bayesian inference*. Ph.D. thesis, University College London.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Berlin: Springer.

Hosino, T., Watanabe, K., & Watanabe, S. (2005). Stochastic complexity of variational Bayesian hidden Markov models. In *Proc. of IEEE international joint conference on neural networks* (pp. 1114–1119).

Jaakkola, T., & Jordan, M. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing*, *10*, 25–37.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., & Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning*, *37*, 183–233.

Nakajima, S., & Sugiyama, M. (2010). Implicit regularization in variational Bayesian matrix factorization. In *Proc. of the 27th international conference on machine learning*.

Nakajima, S., & Watanabe, S. (2007). Variational Bayes solution of linear neural networks and its generalization performance. *Neural Computation*, *19*, 1112–1153.

Rusakov, D., & Geiger, D. (2005). Asymptotic model selection for naive Bayesian networks. *Journal of Machine Learning Research*, *6*(1), 1–35.

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

Seeger, M. (2008). Bayesian inference and optimal design for the sparse linear model. *Journal of Machine Learning Research*, *9*, 759–813.

Seeger, M. (2009). Sparse linear models: variational approximate inference and Bayesian experimental design. *Journal of Physics. Conference Series*, *197*, 012001.

Watanabe, S. (2009). *Algebraic geometry and statistical learning theory*. Cambridge: Cambridge University Press.

Watanabe, K. (2010). An alternative view of variational Bayes and minimum variational stochastic complexity. In *Proc. of 3rd workshop on information theoretic methods in science and engineering (WITMSE-10)*. Tampere International Center for Signal Processing.

Watanabe, K., & Watanabe, S. (2006). Stochastic complexities of Gaussian mixtures in variational Bayesian approximation. *Journal of Machine Learning Research*, *7*, 625–644.

Watanabe, K., & Watanabe, S. (2007). Stochastic complexities of general mixture models in variational Bayesian learning. *Neural Networks*, *20*, 210–219.

Watanabe, K., Shiga, M., & Watanabe, S. (2009). Upper bound for variational free energy of Bayesian networks. *Machine Learning*, *75*, 199–215.

Watanabe, K., Okada, M., & Ikeda, K. (2011). Divergence measures and a general framework for local variational approximation. *Neural Networks*. doi:10.1016/j.neunet.2011.06.004.

Yamazaki, K., & Watanabe, S. (2003a). Singularities in mixture models and upper bounds of stochastic complexity. *Neural Networks*, *16*, 1029–1038.

Yamazaki, K., & Watanabe, S. (2003b). Stochastic complexity of Bayesian networks. In *Uncertainty in artificial intelligence* (pp. 592–599).

Yamazaki, K., & Watanabe, S. (2005). Algebraic geometry and stochastic complexity of hidden Markov models. *Neurocomputing*, *69*, 62–84.

Yamazaki, K., Aoyagi, M., & Watanabe, S. (2010). Asymptotic analysis of Bayesian generalization error with Newton diagram. *Neural Networks*, *23*, 35–43.