Stochastic Complexity of Variational Bayesian Hidden Markov Models

Tikara Hosino
Department of
Computational Intelligence and
System Science,
Tokyo Institute of Technology
Mailbox R2-5, 4259 Nagatsuta, Midori-ku,
Yokohama, 226-8503 Japan.
Nihon Unisys, Ltd.
1-1-1 Toyosu, Koutou-ku,
Tokyo, 135-8560 Japan
E-mail: thosino@cs.pi.titech.ac.jp

Kazuho Watanabe
Department of
Computational Intelligence and
System Science,
Tokyo Institute of Technology
4529 Nagatsuta, Midori-ku,
Yokohama, 226-8503 Japan
E-mail: kazuho23@pi.titech.ac.jp

Sumio Watanabe
Precision and Intelligence Laboratory,
Tokyo Institute of Technology
4529 Nagatsuta, Midori-ku,
Yokohama, 226-8503 Japan
E-mail: swatanab@pi.titech.ac.jp

***Abstract*— Variational Bayesian Learning was proposed as the approximation method of Bayesian learning. In spite of efficiency and experimental good performance, their mathematical property has not yet been clarified. In this paper we analyze variational Bayesian hidden Markov models which include the true one thus the models are non-identifiable. We derive their asymptotic stochastic complexity. It is shown that, in some prior condition, the stochastic complexity is much smaller than those of identifiable models.**

I. Introduction

Hidden Markov Models (HMMs) are popular statistical models for sequence processing and used in many fields such as speech recognition, natural language processing, and bioinformatics [4], [5]. Despite successful applications in wide area, theoretical properties, for example, behavior of training and generalization error are still not clarified.

Main obstacle for analyzing HMMs is their non-identifiability. A learning model $p(x|\theta)$, where θ is a parameter, if the mapping from the parameter to the model is one-to-one, the model is identifiable, otherwise, non-identifiable. Many statistical models with hidden variables are non-identifiable. In non-identifiable models, Fisher information is degenerated and models have many singularities. Thus we cannot analyze them with conventional method.

On the Basis of the Algebraic analysis, general theory of Bayesian stochastic complexity for non-identifiable models is provided [8]. It reveals stochastic complexity and generalization error of non-identifiable models by Bayesian learning are much smaller than identifiable models. However in many cases, exact performing of Bayesian learning is computationally intractable. In practice, we need some kind of approximation schema.

There are two major approximation methods. The first is the sampling method. Most popular techniques are Markov Chain Monte Carlo (MCMC) which create Markov Chain that converge to the desired distribution [3]. In spite of generality and exactness in the limit of them, the MCMC methods are computationally intensive and have difficulty to determine convergence to desired distribution. The second is the deterministic approach. Following this line, variational Bayesian learning for statistical models with hidden variables was proposed [1].

Variational Bayesian learning approximates the posterior distribution directly via minimizing Kullback information to the true posterior. Computational efficiency for the exponential family with hidden variables and the good generalization performance for real world application were reported [2]. However little things are known about their theoretical properties. To evaluate approximation accuracy of it, we need the mathematical foundation. Recently the lower bound of stochastic complexity of variational Bayesian normal mixture models was revealed [8].

In this paper, we analyze variational Bayesian HMMs which include the true one, therefore the models are non-identifiable. We derive the asymptotic stochastic complexity.

II. HMMs and Non-Identifiability

A. Discrete HMMs

In this paper, we deal with discrete HMMs. A sequence $X = \{X_1, \ldots, X_T\}$ was observed. Each X_t is an M dimensional binary vector (M-valued finite alphabet)

$$X_t = (x_{t,1}, \ldots, x_{t,M}),$$

where if output symbol at time t is m then $x_{t,m} = 1$ otherwise 0. Moreover X_t was produced by K-valued discrete hidden state Y_t. The sequence of hidden states $Y = \{Y_1, \ldots, Y_T\}$ was generated by a first-order Markov process. Similarly, Y_t is represented by a K dimensional binary vector

$$Y_t = (y_{t,1}, \ldots, y_{t,K}),$$

where if hidden state at time t is k then $y_{t,k} = 1$ otherwise 0.

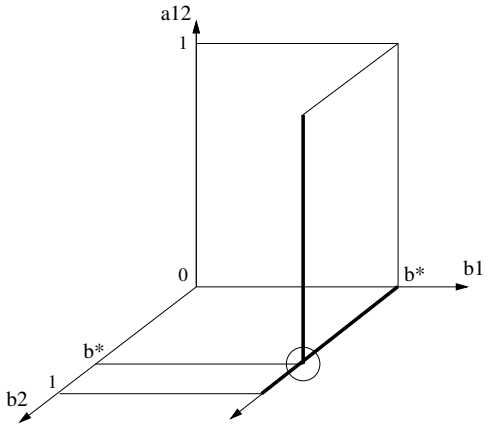

Fig. 1. Non-identifiability of HMMs. Thick lines are set of true parameters. Circle center indicates singular point.

We assume the initial state ($t = 1$) is the first one, namely $y_{1,1} = 1$. Then the probability of a sequence is given by

$$p(X|w) = \sum_{Y} \prod_{m=1}^{M} b_{1m}^{x_{1,m}} \prod_{t=2}^{T} \prod_{i=1}^{K} \prod_{j=1}^{K} a_{ij}^{y_{t,i}y_{t-1,j}} \prod_{m=1}^{M} b_{im}^{y_{t,i}x_{t,m}}, \tag{1}$$

where $\sum_Y$ is taken all over possible values of hidden variables and $w = \{A = \{a_{ij}\}, B = \{b_{im}\}\}$ are model parameters that are constrained by

$$a_{ij} \ (1 \leq i \leq K, 1 \leq j \leq K, 0 \leq a_{ij} \leq 1, \sum_{j=1}^{K} a_{ij} = 1)$$

$$b_{im} \ (1 \leq i \leq K, 1 \leq m \leq M, 0 \leq b_{im} \leq 1, \sum_{m=1}^{M} b_{im} = 1),$$

where a_{ij} represents hidden state's transition probability from jth state to ith state and b_{im} the emission probability that ith state produces alphabet m.

B. Non-identifiability of HMMs

We exemplify non-identifiability of HMM when a learning machine exceeds the true distribution. Let $p_0(x)$ be the true distribution which has one hidden state. We assume emission probability is one dimensional and the parameter is $b^* \in R$. Suppose estimating $p_0(x)$ by the learning machine $p(x|w)$ which has two hidden states and constrained $a_{22} = 1.0$. We assume initial state is the first one. Then, in the parameter space, the set of true parameters is composed by union of two sets those are crossing each other (Fig. 1), and given by

$$\{b_1 = b^*, 0 \leq b_2 \leq 1, a_{12} = 0\} \cup \{b_1 = b_2 = b^*, 0 \leq a_{12} \leq 1\}.$$

This set has a singular point $(b_1, b_2, a_{12}) = (b^*, b^*, 0)$. At the singular point, Fisher information is degenerated, therefore we cannot use the conventional quadratic (gaussian) approximation.

III. Bayesian Learning and Variational Approximation

A. Bayesian Learning

In this section we describe common procedure for Bayesian learning. Suppose n independent and identical data samples: $X^n = \{X_1, \ldots, X_n\}$, are provided which are taken from true distribution $p_0(x)$. Let $p(x|\theta)$ denote a learning machine where θ is a model parameter and $\varphi(\theta)$ is the prior distribution of the parameter.

Given data set, we compute the posterior distribution of parameters,

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

where $Z(X^n)$ is the normalizing constant. We can also predict future data $\hat{X}$ by averaging over the parameters,

$$p(\hat{X}|X^n) = \int p(\hat{X}|\theta)p(\theta|X^n)d\theta,$$

which is called predictive distribution.

Then, the stochastic complexity is defined by

$$F(X^n) = -\log Z(X^n).$$

For later use, we also define the model entropy as,

$$S = -\int p_0(x) \log p_0(x) dx,$$

which does not depend on learning machines.

In Bayesian learning, the asymptotic behavior of the stochastic complexity of a singular model was derived [7]. Averaged stochastic complexity is defined by

$$F(n) = < -\log \int \prod_{i=1}^{n} p(X_i|\theta)\varphi(\theta)d\theta >_{p_0(X^n)} - nS,$$

where we average over all training data sets and we use the notation $<>_{p(x)}$ for the value of expectation over $p(x)$.

We define Kullback information as,

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

Further, we define meromorphic function of z by

$$J(z) = \int H(\theta)^z \varphi(\theta) d\theta.$$

Then, it was proved that $F(n)$ has the following asymptotic form,

$$F(n) = \lambda \log n - (m-1) \log \log n + O(1),$$

where $-\lambda$ is a rational number which is the maximal pole of $J(z)$ and m is multiplicity of it. In identifiable models, it was shown that $\lambda = \frac{d}{2}$ where d is the number of model parameters. Whereas, in many non-identifiable models, λ is much smaller than $\frac{d}{2}$.

In the case of Bayesian learning HMMs, we assume following conditions.

(A1) The true HMM has K_0 hidden states.

(A2) The learning machine is given by eq. (1).

(A3) The prior is uniform over the parameters.

Then, by using division of Kullback information and blowing up which is algebraic geometrical method, one pole of $J(z)$ was obtained [9]. Hence, the coefficient λ is upper bounded by

$$\lambda \leq \frac{K_0(K_0-1)+K_0(M-1)}{2}+K_0(K-K_0), \quad (2)$$

which is much smaller than half of the number of model parameters:

$$\frac{K(K-1)+K(M-1)}{2}.$$

In particular, if HMMs have simple left-to-right structure, utilizing sparsity of the transition probability, the upper bound of the coefficient λ is given by

$$\lambda \leq \frac{(K_0-1)+K_0(M-1)}{2}+\frac{1}{2} \quad (3)$$

B. Variational Bayesian Learning

Using complete data $\{X^n, Y^n\}$ where $X^n = \{X_1, \ldots, X_n\}$ is a sample dataset and $Y^n = \{Y_1, \ldots, Y_n\}$ is a corresponding hidden dataset, we can write stochastic complexity as

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

where $\sum_{Y^n}$ is taken all over possible values of hidden variables. The variational Bayesian learning approximates the stochastic complexity by using an arbitrary trial distribution $q(Y^n, \theta)$. By using Jensen's inequality, the stochastic complexity is upper bounded as

$$\begin{aligned} F(X^n) &= -\log \int \sum_{Y^n} q(Y^n, \theta)\frac{p(X^n, Y^n, \theta)}{q(Y^n, \theta)}d\theta \\ &\leq -\int \sum_{Y^n} q(Y^n, \theta)\log \frac{p(X^n, Y^n, \theta)}{q(Y^n, \theta)}d\theta \\ &\equiv \overline{F}(X^n). \end{aligned}$$

We rewrite it to

$$\overline{F}(X^n) = F(X^n) + \int \sum_{Y^n} q(Y^n, \theta)\log \frac{q(Y^n, \theta)}{p(Y^n, \theta|X^n)}d\theta, \quad (4)$$

where we use $p(X^n, Y^n, \theta) = p(Y^n, \theta|X^n)p(X^n)$. Equation (4) shows that minimizing $\overline{F}(X^n)$ is equivalent to minimizing Kullback information from the trial posterior $q(Y^n, \theta)$ to the true posterior $p(Y^n, \theta|X^n)$ and the equality condition only is fulfilled when $p(Y^n, \theta|X^n) = q(Y^n, \theta)$. Namely the trial distribution equals the true posterior distribution.

Additionally for the computational efficiency we constrain the trial distribution $q(Y^n, \theta)$ to the factorized form $q(Y^n)r(\theta)$. Then constrained $\overline{F}(Y^n)$ is given by

$$\begin{aligned} F(X^n) &\leq -\int \sum_{Y^n} q(Y^n)r(\theta)\log \frac{p(X^n, Y^n|\theta)\varphi(\theta)}{q(Y^n)r(\theta)}d\theta \\ &= -\int \sum_{Y^n} q(Y^n)r(\theta)\log \frac{p(X^n, Y^n|\theta)}{q(Y^n)}d\theta \\ &+ \int r(\theta)\log \frac{r(\theta)}{\varphi(\theta)}d\theta \equiv \overline{F}(X^n). \end{aligned}$$

$\overline{F}(X^n)$ is the functional of the free distributions $q(Y^n)$ and $r(\theta)$. Minimization of $\overline{F}(X^n)$ is achieved by variational methods. Under the constraint of $\int r(\theta)d\theta = 1$, We take a functional derivatives of $\overline{F}(X^n)$ with respect to $r(\theta)$, and equal to zero. Then, we obtain the optimal variational posterior of parameters as

$$r(\theta) = C_r \exp < \log p(X^n, Y^n|\theta)\varphi(\theta) >_{q(Y^n)}, \quad (5)$$

where C_r is the normalizing constant. In the similar fashion, we obtain the optimal variational posterior of the hidden variable as

$$q(Y^n) = C_q \exp < \log p(X^n, Y^n|\theta) >_{r(\theta)}, \quad (6)$$

where C_q is the normalizing constant. Variational Bayesian learning iteratively minimizes $\overline{F}(X^n)$ by using eq. (5) and eq. (6). If we select $p(X^n, Y^n)$ from an exponential family with hidden variables which includes discrete HMMs and a corresponding conjugate prior, the very efficient Expectation-Maximization like algorithm is derived. During the iteration, $\overline{F}(X^n)$ decreases monotonically and the convergence to a local minimum is guaranteed.

IV. Main Results

Define the averaged variational stochastic complexity as

$$\begin{aligned} \overline{F}(n) &= \langle \overline{F}(X^n) \rangle_{p_0(X^n)} - nS \\ &= \langle -\sum_{i=1}^{n} \{ \langle \log \frac{p(X_i, Y_i|\theta)}{q(Y_i)} \rangle_{q(Y_i)r(\theta)} \\ &+ \log p_0(X_i) \} \rangle_{p_0(X^n)} + \int r(\theta)\log \frac{r(\theta)}{\varphi(\theta)}d\theta \end{aligned} \quad (7)$$

and assume the following conditions (A1), (A2), (A3), (A4).

(A1) The true distribution $p_0(x)$ has K_0 hidden states and emits M-valued discrete symbols. The true distribution $p_0(x)$ is given by

$$p(X|\theta^*) = \sum_{Y} \prod_{m=1}^{M} b_{1m}^{*x_{1,m}} \prod_{t=2}^{T} \prod_{i=1}^{K_0} \prod_{j=1}^{K_0} a_{ij}^{*y_{t,i}y_{t-1,j}} \prod_{m=1}^{M} b_{im}^{*y_{t,i}x_{t,m}}$$

where $\sum_Y$ is taken all over possible values of hidden variables. Moreover the true parameter is defined by

$$\theta^* = \{\{a_{ij}^*\}, \{b_{im}^*\}\} \ (1 \leq i, j \leq K_0, 1 \leq m \leq M).$$

(A2) All parameters $\{a_{ij}^*, b_{im}^*\}$ are strictly positive:

$$\theta^* = \{\{a_{ij}^* > 0\}, \{b_{im}^* > 0\}\} \ (1 \leq i, j \leq K_0, 1 \leq m \leq M).$$

(A3) The learning machine eq. (1) can attain the true distribution, thus the learning machine has K hidden states that satisfy $K_0 \leq K$.

(A4) The prior distributions of the transition probability $A = \{a_{ij}\}$ and the emission probability $B = \{b_{im}\}$ are the Dirichlet distribution and hyperparameters are denoted by $\phi_0 > 0, \xi_0 > 0$:

$$\varphi(A) = \prod_{i=1}^{K} \frac{\Gamma(K\phi_0)}{\Gamma(\phi_0)^K} \prod_{j=1}^{K} a_{ij}^{\phi_0 - 1},$$

$$\varphi(B) = \prod_{i=1}^{K} \frac{\Gamma(M\xi_0)}{\Gamma(\xi_0)^M} \prod_{m=1}^{M} b_{im}^{\xi_0 - 1}.$$

Theorem 1

Assume the conditions from (A1) to (A4). Then the averaged variational stochastic complexity satisfies

$$\frac{\overline{F}(n)}{\log n} \to \overline{\lambda},$$

as n tends to infinity where

$$\overline{\lambda} = \begin{cases} \frac{K_0(K_0-1)+K_0(M-1)}{2} + K_0(K-K_0)\phi_0 \\ \quad (\text{if } \phi_0 \leq \frac{K_0+K+M-2}{2K_0}), \\ \frac{K(K-1)+K(M-1)}{2} \\ \quad (\text{otherwise } \phi_0 > \frac{K_0+K+M-2}{2K_0}). \end{cases} \tag{8}$$

(Proof of Theorem 1).

In the case of HMMs, the log likelihood of a sequence of complete data $\{X, Y\}$ is defined by

$$\log p(X, Y|\theta) = \sum_{t=2}^{T} \sum_{i=1}^{K} \sum_{j=1}^{K} y_{t,i} y_{t-1,i} \log a_{ij} + \sum_{t=1}^{T} \sum_{i=1}^{K} \sum_{m=1}^{M} y_{t,i} x_{t,m} \log b_{im}.$$

Moreover, we define the expected sufficient statistics by

$$n_i = \sum_{t=1}^{T} < y_{t,i} >_{q(Y)},$$

$$n_{ij} = \sum_{t=2}^{T} < y_{t,i} y_{t-1,i} >_{q(Y)},$$

$$n_{im} = \sum_{t=1}^{T} < y_{t,i} >_{q(Y)} x_{t,m},$$

where the expected count n_i is constrained by $n_i = \sum_j n_{ij}$.

Then, the posterior distribution of parameters $r(\theta)$ is given by

$$r(A) = \prod_{i=1}^{K} \frac{\Gamma(n_i + K\phi_0)}{\prod_{j=1}^{K} \Gamma(n_{ij} + \phi_0)} \prod_{j=1}^{K} a_{ij}^{n_{ij} + \phi_0 - 1},$$

$$r(B) = \prod_{i=1}^{K} \frac{\Gamma(n_i + M\xi_0)}{\prod_{m=1}^{M} \Gamma(n_{im} + \xi_0)} \prod_{m=1}^{M} b_{im}^{n_{im} + \xi_0 - 1}.$$

At the same time, the posterior distribution of hidden variables is given by

$$q(Y|X) = C_q \exp(\sum_{t=2}^{T} \sum_{i=1}^{K} \sum_{j=1}^{K} y_{t,i} y_{t-1,j} < \log a_{ij} >_{r(\theta)} + \sum_{t=1}^{T} \sum_{i=1}^{K} \sum_{m=1}^{M} y_{t,i} x_{t,m} < \log b_{im} >_{r(\theta)}), \tag{9}$$

where C_q is the normalizing constant.

First, we consider the first term of $\overline{F}(n)$. Using Jensen's inequality,

$$\int \sum_{Y^n} q(Y^n) r(\theta) \log \frac{P(X^n, Y^n)}{q(Y^n)} d\theta \leq \log < p(X^n|\theta) >_{r(\theta)} \leq \log p(X^n|\theta^*)$$

where θ^* is the maximum likelihood parameter and using the predictive distribution $< p(X^n|\theta) >_{r(\theta)}$ belongs to the same distribution as $p(X^n|\theta^*)$. When the stochastic model is the finite dimension, under weak assumptions, the likelihood ratio statistics converges $O(1)$ as n tends to infinity and discrete HMMs satisfy the assumption. Hence, the first term of $\overline{F}$ satisfies inequality:

$$\sum_{i=1}^{n} \{\langle -\log \frac{p(X_i, Y_i|\theta)}{q(Y_i)} \rangle_{q(Y_i) r(\theta)} + p_0(X_i)\}$$
$$\geq \sum_{i=1}^{n} \log \frac{p_0(X_i)}{\langle p(X_i|\theta) \rangle_{r(\theta)}}$$
$$\geq \sum_{i=1}^{n} \log \frac{p_0(X_i)}{p(X_i|\theta^*)} \to O(1) \quad (n \to \infty)$$

where the right hand side is the likelihood ratio statistics.

Further, using the law of large numbers,

$$\frac{1}{n} \sum_{i=1}^{n} \log \frac{p_0(X_i)}{\langle p(X_i|\theta) \rangle_{r(\theta)}}$$
$$\to \int p_0(X) \log \frac{p_0(X)}{\langle p(X|\theta) \rangle_{r(\theta)}} dX \quad (n \to \infty),$$

Since the right hand side is Kullback information, it is positive. When the learning machine uses the fewer hidden states than the true one, in other words, the number of i which $n_i \to \infty$ is smaller than K_0, Kullback information is lower bounded by positive constant. Therefore, first term of $\overline{F}(n)$ diverges to plus infinity. Hence, we can assume the number of hidden states which $n_i \to \infty$ is at least K_0.

Next, we consider the second term of $\overline{F}(n)$:

$$\int r(\theta) \log \frac{r(\theta)}{\varphi(\theta)} d\theta = K(r(\theta)||\varphi(\theta)),$$

which is Kullback information from the posterior distribution

to the prior distribution of parameters.

$$
\begin{aligned}
&K(r(\theta)||\varphi(\theta)) \\
&= \sum_{i=1}^{K}\Big\{\log\Gamma(n_i+K\phi_0) - n_i\Psi(n_i+K\phi_0) \\
&- \sum_{j=1}^{K}\{\log\Gamma(n_{ij}+\phi_0) - n_{ij}\Psi(n_{ij}+\phi_0)\}\Big\} \\
&+ \sum_{i=1}^{K}\Big\{\log\Gamma(n_i+M\xi_0) - n_i\Psi(n_i+M\xi_0) \\
&- \sum_{m=1}^{M}\{\log\Gamma(n_{im}+\xi_0) - n_{im}\Psi(n_{im}+\xi_0)\}\Big\} \\
&+ const. \qquad (10)
\end{aligned}
$$

Using the asymptotic forms of the psi function $\Psi(x)$ and the log gamma function $\log\Gamma(x)$ as $x\to\infty$

$$
\begin{aligned}
&\Psi(x) = \log x + O(1), \\
&\log\Gamma(x) = (x-\frac{1}{2})\log x - x + O(1).
\end{aligned}
$$

Then, the asymptotic form of eq. (10) is given by

$$
\begin{aligned}
K(r(\theta)||\varphi(\theta)) = &\sum_{i=1}^{K}\Big\{(K\phi_0-\frac{1}{2})\log(n_i+K\phi_0) \\
&- \sum_{j=1}^{K}\{(\phi_0-\frac{1}{2})\log(n_{ij}+\phi_0)\}\Big\} \\
&+ \sum_{i=1}^{K}\Big\{(M\xi_0-\frac{1}{2})\log(n_i+M\xi_0) \\
&- \sum_{m=1}^{M}\{(\xi_0-\frac{1}{2})\log(n_{im}+\xi_0)\}\Big\} + O(1). \qquad (11)
\end{aligned}
$$

To evaluate the coefficient of the leading term, set each n_i to be the α_ith order of n and n_{ij} the β_{ij}th order of n:

$$
\begin{aligned}
&n_i = p_i n^{\alpha_i} + o(n^{\alpha_i}), (0<p_i, 0\le\alpha_i\le 1, i=1,\ldots,K), \\
&n_{ij} = p_{ij} n^{\beta_{ij}} + o(n^{\beta_{ij}}), \\
&(0<p_{ij}, 0\le\beta_{ij}\le 1, i,j=1,\ldots,K).
\end{aligned}
$$

Then, by using (A2),

$$
\begin{aligned}
&\log(n_i+K\phi_0) = \log(n_i+M\xi_0) = \alpha_i\log n + o(\log n), \\
&\log(n_{ij}+\phi_0) = \beta_{ij}\log n + o(\log n), \qquad (12) \\
&\log(n_{im}+\xi_0) = \alpha_i\log n + o(\log n)
\end{aligned}
$$

hold. Hence substitute eq. (12) into eq. (11),

$$
\begin{aligned}
\frac{K(r(\theta)||\varphi(\theta))}{\log n} &= \sum_{i=1}^{K}\{(K\phi_0-\frac{1}{2})\alpha_i - \sum_{j=1,\alpha_i\neq 0}^{K}(\phi_0-\frac{1}{2})\beta_{ij}\} \\
&+ \sum_{i=1}^{K}\{(\frac{M}{2}-\frac{1}{2})\alpha_i\} + o(1) \quad (n\to\infty), \qquad (13)
\end{aligned}
$$

where the inner sum is taken only $\alpha_i\neq 0$ which reflects the constraint $n_i = \sum_{j=1}^{K} n_{ij}$.

We divide the sum (13) to the true K_0 and redundant $K-K_0$ terms. Moreover, we assume additional $l(0\le l\le K-K_0)$ hidden states are used and β_{ij} is the same value of α_i,

$$
\begin{aligned}
&\sum_{i=1}^{K_0}\{(K\phi_0+\frac{M}{2}-1)\alpha_i + \sum_{j=1,\alpha_i\neq 0}^{K_0}(\frac{1}{2}-\phi_0)\beta_{ij}\} \\
&+ \sum_{i=K_0+1}^{K}\{(K\phi_0+\frac{M}{2}-1)\alpha_i - \sum_{j=K_0+1,\alpha_i\neq 0}^{K}(\phi_0-\frac{1}{2})\beta_{ij}\} \\
&= \sum_{i=1}^{K_0}\{(K\phi_0+\frac{M}{2}-1)\alpha_i - \sum_{j=1,\alpha_i\neq 0}^{K_0}(\phi_0-\frac{1}{2})\beta_{ij}\} + g(l),
\end{aligned}
$$

where $g(l)$ is given by

$$
g(l) = (K\phi_0+\frac{M}{2}-1)l - (\phi_0-\frac{1}{2})(2K_0 l + l^2).
$$

Minimizing $g(l)$ is achieved by

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

In the case $l = K-K_0$, when $\alpha_i = 1$ and $\beta_{ij} = 0$, then eq. (13) surely increase. Thus, the minimum of eq. (13) is obtained by $g(l)$. On the upper bound, in these conditions, learning machine contains the true one, therefore we can set n_i and n_{ij} appropriately so that the first term of $\overline{F}(n)\le O(1)$ as $n\to\infty$. Finally, plug (K_0+l) into eq. (13), which completes the proof.

(End of proof of Theorem 1).

Next, we consider the simple left-to-right HMMs.

(A5) In the simple left-to-right HMMs whose hidden state's transition is constrained in itself or the next state:

$$
\{a_{ij} = 0, i\neq\{j, j+1\}\}.
$$

Thus only $a_{j+1,j}$ is a substantial parameter in transition probability.

Theorem 2.

Assume the conditions from (A1) to (A5). Then the averaged variational stochastic complexity satisfies

$$
\frac{\overline{F}(n)}{\log n} \to \overline{\lambda}
$$

as n tends to infinity where

$$
\overline{\lambda} = \begin{cases}
\frac{(K_0-1)+K_0(M-1)}{2} & (\text{if } K=K_0) \\
\frac{(K_0-1)+K_0(M-1)}{2} + \frac{1}{2} & (\text{if } K>K_0).
\end{cases} \qquad (14)
$$

(Proof of Theorem 2).

Using the constraints of the transition probability, the asymptotic form of Kullback information is given by

$$K(r(\theta)||\varphi(\theta)) = \sum_{i=1}^{K-1}\Big\{(2\phi_0 - \frac{1}{2})\log(n_i + 2\phi_0) - \sum_{j=1}^{2}\{(\phi_0 - \frac{1}{2})\log(n_{ij} + \phi_0)\}\Big\} + \sum_{i=1}^{K}\Big\{(M\xi_0 - \frac{1}{2})\log(n_i + M\xi_0) - \sum_{m=1}^{M}\{(\xi_0 - \frac{1}{2})\log(n_{im} + \xi_0)\}\Big\} + O(1).$$

According to the same argument of theorem 1, Kullback information satisfies

$$\frac{K(r(\theta)||\varphi(\theta))}{\log n} = \sum_{i=1}^{K-1}\{(2\phi_0 - \frac{1}{2})\alpha_i - \sum_{j=1,\alpha_i\neq 0}^{2}(\phi_0 - \frac{1}{2})\beta_{ij}\} + \sum_{i=1}^{K}\{(\frac{M}{2} - \frac{1}{2})\alpha_i\} + o(1) \quad (n\to\infty). \tag{15}$$

Then, the similar function $g(l)$ $(0 \le l \le K - K_0)$ is given by

$$g(l) = \frac{M}{2}l.$$

Hence, the minimum of $g(l)$ is obviously obtained by $l = 0$. Finally, in the eq. (15), when $K = K_0$ then substitutes the first term K to K_0 otherwise $K_0 + 1$ and the second term K to K_0, which completes the proof.

(End of proof of Theorem 2).

V. Discussion

First, theorem 1 shows that the asymptotic stochastic complexity is divided into two cases of the hyperparameter in the transition probability ϕ_0. When ϕ_0 is large, the coefficient $\overline{\lambda}$ is equal to the number of model parameters. Then stochastic complexity, as suggested by Attias [1], coincides with the BIC [6]. However when ϕ_0 is small, $\overline{\lambda}$ is much smaller than the number of model parameters (Fig. 2). In particular, the growth of stochastic complexity not depends on the dimension of the emission probability.

Second, using the stochastic complexity as a model selection criterion, when ϕ_0 is small, the stochastic complexity is not sharply peaked at the true model size. However, in that condition, redundant hidden states are effectively eliminated. Therefore, there is no deficit for the use of the prediction. On the contrary, when ϕ_0 is large, the stochastic complexity has a clear minimum at the true model size. Hence we can easily select an optimal model. An analysis of a trade-off between them and finding the optimal hyperparameter is a future research.

Third, to evaluate the approximation accuracy of variational Bayesian HMMs, we compare the variational stochastic complexity to the exact Bayes upper bound. In the Dirichlet distribution, the uniform prior corresponds to the hyperparameter $\phi_0 = 1.0$. By substituting $\phi_0 = 1.0$ into eq. (8), eq. (14) and comparing them to eq. (2), eq. (3), we get the accordance of them. Hence, for the detailed comparison, we need more accurate estimate of the exact Bayes stochastic complexity.

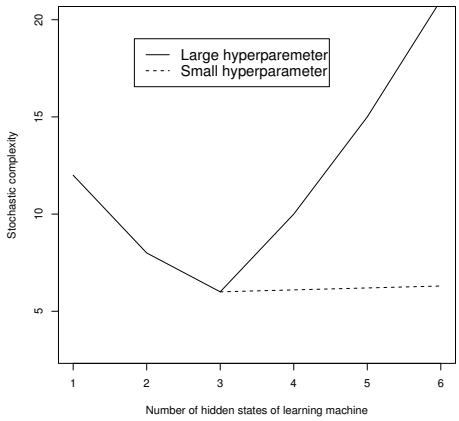

Fig. 2. The schematic behavior of the variational stochastic complexity scaled by $\log n$. The true distribution has 3 hidden states, emits 3-valued alphabet and the small hyperparameter is $\phi_0 = 0.1$.

VI. Conclusions

The asymptotic behavior of the stochastic complexity in non-identifiable variational Bayesian HMMs was derived. The result shows that, in some prior condition, the stochastic complexity is much smaller than identifiable models. Further, in those condition, the stochastic complexity satisfies eliminating redundant hidden states. This indicates the reason why variational Bayesian HMMs enjoy discovering a model structure.

Acknowledgment

We thank Keisuke Yamazaki for helpful discussions. This work was supported by the Ministry of Education, Science, Sports, and Culture in Japan, Grant-in-aid for scientific research 15500130.

References

[1] H. Attias, "Inferring parameters and structure of latent variable models by variational Bayes", in *Proc. 15th Conference on Uncertainty in Artificial Intelligence*, 1999, pp. 21-20.

[2] M. J. Beal, *Variational Algorithms for Approximate Bayesian Inference*, PhD thesis, University College London, 2003.

[3] W. R. Gilks, S. Richardson, D. J. Speigelhalter (eds), *Markov Chain Monte Carlo in Practice*, Chapman and Hall, 1996.

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

[5] P. Baldi, S. Brunak, *Bioinformatics*, The MIT Press, 1998.

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

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

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

[9] K. Yamazaki, S. Watanabe, "Stochastic Complexities of Hidden Markov Models," *IEEE International Workshop On Neural Networks For Signal Processing*, 2003.