An Analysis of the Difference of Code Lengths Between Two-Step Codes Based on MDL Principle and Bayes Codes

Masayuki Goto, *Member, IEEE*, Toshiyasu Matsushima, *Member, IEEE*, and Shigeichi Hirasawa, *Fellow, IEEE*

***Abstract*—In this paper, we discuss the difference in code lengths between the code based on the minimum description length (MDL) principle (the MDL code) and the Bayes code under the condition that the same prior distribution is assumed for both codes. It is proved that the code length of the Bayes code is smaller than that of the MDL code by $o(1)$ or $O(1)$ for the discrete model class and by $O(1)$ for the parametric model class. Because we can assume the same prior for the Bayes code as for the code based on the MDL principle, it is possible to construct the Bayes code with equal or smaller code length than the code based on the MDL principle. From the viewpoint of mean code length per symbol unit (compression rate), the Bayes code is asymptotically indistinguishable from the MDL two-stage codes.**

***Index Terms*—Asymptotic normality, Bayes code, minimum description length (MDL) principle, universal coding.**

I. Introduction

THE minimum description length (MDL) principle which was proposed by J. Rissanen has been studied [24]–[29] not only in the universal source coding but in the areas of data analysis, learning theory, and the statistical model selection [13], [14], [17], [36].

We discuss two-step coding based on the MDL principle. Since coding based on the MDL principle is the method minimizing the total description length of the data and a probabilistic model, the MDL criterion selects a particular model, and is closely related to Bayesian statistics. This is because the MDL supposes the prior distribution implicitly and is essentially equivalent to the maximization of the posterior probability.

On the other hand, Bayes coding [4], [20], whose code length is also called stochastic complexity [23], [29] is the method which uses the mixture of all models explicitly over a model class for coding function. Recently, efficient algorithms to calculate the mixture probability of the data sequence have been reported for the FSMX model class [18], [21], [38]. The Bayes code is given by the Bayes optimal solution for the code length [20], [23]. Therefore, if we can assume the same prior distribution, it is clear that the Bayes code is not worse than the two-step code based on MDL principle (MDL code[1]) [29].

The properties of these codes have been studied independently (see [4], [5], [13], [18], [20], [24]–[29], [34]–[36]). The main interest here is a quantitative evaluation between the MDL code and the Bayes code. For the same prior, the code length of the Bayes code is a lower bound on that of the MDL code for any data sequence [29], since the Bayes code is the Bayes optimal. Moreover, both of these two codes are asymptotically optimum [4], [14]. The analyses on the MDL and the Bayes code from the viewpoint of the estimator also have been studied [35]. However, the difference between the code lengths has not been analyzed directly or quantitatively.

In this paper, we analyze the difference of the code lengths between the MDL code and the Bayes code for the discrete, the parametric, and the hierarchical model classes, and show that the code length of the Bayes code is smaller than that of the MDL code by $o(1)$ or $O(1)$ for the discrete model class, and by $O(1)$ for the parametric model class. For hierarchical model classes, the difference of the code lengths between the MDL code with a mixture over parameters but a selection for the model order and the Bayes code which uses a mixture over both parameters and models is $o(1)$. The essence of the analysis for the parametric model class is that the posterior probability density of the parameter on Bayesian inference satisfies asymptotic normality. Because we assume the same prior for the Bayes code as that of the MDL code in practice, it is possible to construct the Bayes code with equal or smaller code length than the MDL code. However, from the viewpoint of mean code length per symbol unit, that is, compression rate, the Bayes code is asymptotically indistinguishable from the MDL two-stage codes.

II. Preliminaries

In this paper, we deal with the discrete, the parametric, and the hierarchical model classes. Let $\mathcal{X}$ be a discrete source alphabet and X a random variable on $\mathcal{X}$. And we denote the data sequence with length n emitted from the source by $x^n = x_1 x_2 \cdots x_n$, where $x_i \in \mathcal{X}$ $(i = 1, 2, \ldots n)$. An infinite sequence from the source is denoted by x^∞. The set of all x^n $(x_i \in \mathcal{X}, i = 1, 2, \ldots, n)$, is denoted by $\mathcal{X}^n$, where $n = 1, 2, \ldots, \infty$ and $x^1 = x$. We denote a random variable

Manuscript received March 21, 1999; revised May 17, 2000. The material in this paper was presented at the International Symposium on Information Theory and Its Application, 1997

M. Goto is with the Department of Environmental and Ocean Engineering, School of Engineering, The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-8656 Japan (e-mail: goto@biz-model.t.u-tokyo.ac.jp).

T. Matsushima and S. Hirasawa are with the Department of Industrial and Management Systems Engineering, School of Science and Engineering, Waseda University, 3-4-1, Ohkubo, Shinjuku-ku, Tokyo, 169-8555 Japan (e-mail: toshi@matsu.mgmt.waseda.ac.jp; hirasawa@hirasa.mgmt.waseda.ac.jp).

Communicated by N. Merhav, Associate Editor for Source Coding.

Publisher Item Identifier S 0018-9448(01)01351-7.

[1]We call this the MDL code, although the Bayes code was also proposed from the viewpoints of the MDL principle by Rissanen [29].

on $\mathcal{X}^n$ by X^n. The source emits a data sequence with true probability distribution

$$P^*(x^n) = P^*(X^n = x^n), \qquad x^n \in \mathcal{X}^n$$

which was previously unknown. $P(\cdot)$ is a probability function and $f(\cdot)$ is a probability density function. Throughout the paper, we suppose that the logarithm base is e and the measure of the code length is in nats.

For a $k \times 1$ matrix vector of length k, x, and a $k \times k$ matrix J, we define the norms $\|x\| = (x^T x)^{1/2}$ and $\|x\|_J^2 = x^T J x$. Here, x^T is the transpose of vector x. And, for the some differentiable function $g(\theta^k)$ with respect to k-dimensional vector θ^k, we define that

$$\frac{\partial g(\theta^k)}{\partial \theta^k} \quad \text{and} \quad \frac{\partial^2 g(\theta^k)}{\partial \theta^k (\partial \theta^k)^T}$$

are the $k \times 1$ matrix having $\frac{\partial g(\theta^k)}{\partial \theta_i}$ as the ith element and the $k \times k$ matrix having $\frac{\partial^2 g(\theta^k)}{\partial \theta_i \partial \theta_j}$ as the i-jth component, respectively. We also denote them by $g'(\theta^k)$ and $g''(\theta^k)$, respectively. For an area $\mathcal{A} \subset \mathcal{R}^k$, volume is denoted by $|\mathcal{A}|$, where $\mathcal{R}^k$ is k-dimensional Euclidean space.

A. The Discrete Model Class

Let λ be a model in a discrete and finite model class Λ, $\lambda \in \Lambda$. That is, Λ is a countable set. The data sequence x^n is emitted from the true distribution $P^*(\cdot)$. We do not assume in analysis that the model class includes the true distribution $P^*(\cdot)$. Let $P(x^n|\lambda) = P(X^n = x^n|\lambda)$ be the probability function for the model λ, and let $C(\lambda)$ be the description length for describing λ. Then, we can encode x^n using a model λ, and its code length is given by $-\log P(x^n|\lambda) + C(\lambda)$. The code length of the MDL code is, therefore, given by

$$\min_{\lambda \in \Lambda} \{-\log P(x^n|\lambda) + C(\lambda)\}.$$

Here, we assume that $\sum_{\lambda \in \Lambda} e^{-C(\lambda)} = 1$ [23].[2] Then, we can regard $P(\lambda) = e^{-C(\lambda)}$ as the prior probability of the model λ. Then the code length of the MDL code, $L^{\lambda}_{\mathrm{MDL}}(x^n)$,[3] is defined by[4]

$$L^{\lambda}_{\mathrm{MDL}}(x^n) = \min_{\lambda \in \Lambda} \{-\log P(x^n|\lambda) - \log P(\lambda)\}. \tag{2}$$

On the other hand, if the prior probability $P(\lambda)$ is assumed, we can construct a Bayes optimal solution for the loss function for the code length [20]. The Bayes code which is the Bayes optimal solution is given by the code using the mixture probability of all models. The code length of the Bayes code $L^{\lambda}_{\mathrm{Bayes}}(x^n)$ is given by

$$L^{\lambda}_{\mathrm{Bayes}}(x^n) = -\log \sum_{\lambda \in \Lambda} P(x^n|\lambda) P(\lambda). \tag{3}$$

B. The Parametric Model Class

We consider the parametric model class $\{P(\cdot|\theta^k)|\theta^k \in \Theta^k\}$ which has a k-dimensional continuous parameter

$$\theta^k = (\theta_1, \theta_2, \ldots, \theta_k)^T.$$

θ^k is a continuous vector in the parameter space Θ^k, where Θ^k is a compact subset of $\mathcal{R}^k$. That is, $\theta^k \in \Theta^k$ and $\Theta^k \subset \mathcal{R}^k$. The data sequence x^n is derived from the true distribution $P^*(\cdot)$. We do not assume that the model class includes the true distribution $P^*(\cdot)$.

Since θ^k is a continuous vector, we cannot encode θ^k as is. In the MDL code, the parameter set is quantized into countable cells and a quantized parameter $\overline{\theta}^k$ which is the representative point of a cell is encoded. Let $\overline{\Theta}^k_n$ be the set of all quantized parameter values $\overline{\theta}^k$, which may depend on n. That is, $\overline{\theta}^k \in \overline{\Theta}^k_n$ when x^n is encoded. Number each quantized cell in the parameter space. Let $\Theta^k_{n,l}$ be the set of all parameter in the lth quantized cell ($l = 1, 2, \ldots$). We assume $\cup_l \Theta^k_{n,l} = \Theta^k$ and $\Theta^k_{n,i} \cap \Theta^k_{n,j} = \phi$, $i \neq j$. Let $\alpha_{n,j}(\overline{\theta}^k)$ be the quantized width of the jth side of the cell represented by $\overline{\theta}^k$. The number of the cells is $O(1/\max \prod_{j=0}^{k-1} \alpha_{n,j}(\overline{\theta}^k))$. Then, it is possible to construct the prefix code which describes the quantized parameter $\overline{\theta}^k$ by the length $-\log f(\overline{\theta}^k) - \sum_j \log \alpha_{n,j}(\overline{\theta}^k)$ [23], where $f(\theta^k)$ is the prior density of the parameter θ^k, satisfying $\int_{\theta^k} f(\theta^k)\, d\theta^k = 1$. We assume that $\overline{\Theta}^k_n$ satisfies

$$\sum_{\overline{\theta}^k \in \overline{\Theta}^k_n} f(\overline{\theta}^k) \prod_j \alpha_{n,j}(\overline{\theta}^k) = 1.$$

Therefore, we can interpret $f(\overline{\theta}^k) \prod_j \alpha_{n,j}(\overline{\theta}^k)$ as the prior probability on $\overline{\Theta}^k_n$.

Here, we must consider the method for the quantization of the parameter space into cells, and an asymptotic method gives a solution. The asymptotic optimal quantized width for the code length is given by

$$\prod_{j=0}^{k-1} \alpha^*_{n,j}(\overline{\theta}^k) = \frac{1}{\sqrt{n}^k \sqrt{\det I(\overline{\theta}^k)}}. \tag{4}$$

Hence, $\alpha^*_{n,j}(\overline{\theta}^k) = O(1/\sqrt{n})$ [14], [24], [35], [37], [39], using the asymptotic normality of the maximum-likelihood estimator. Here $I(\theta^k)$ is the Fisher information matrix[5] defined by

$$I(\theta^k) = \lim_{n \to \infty} \frac{1}{n} E_\theta \left\{ -\frac{\partial^2 \log P(X^n|\theta^k)}{\partial \theta^k (\partial \theta^k)^T} \right\} \tag{5}$$

and $E_\theta\{\cdot\}$ is the expectation under $P(\cdot|\theta^k)$.

[2] If $\sum_{\lambda \in \Lambda} e^{-C(\lambda)} > 1$, then the prefix code does not exist, else if $\sum_{\lambda \in \Lambda} e^{-C(\lambda)} < 1$, then it causes the loss of the code length.

[3] The superscript λ of $L^{\lambda}_{\mathrm{MDL}}(x^n)$ means the code for the discrete model class. Later, we define the code length of the MDL code, $L^{\overline{\theta}^k}_{\mathrm{MDL}}(x^n)$, for the parametric model class.

[4] The MDL estimator $\bar{\lambda}_{\mathrm{MDL}}$ is given by

$$\begin{aligned} \bar{\lambda}_{\mathrm{MDL}} &= \arg\min_{\lambda \in \Lambda} \{-\log P(x^n|\lambda) - \log P(\lambda)\} \\ &= \arg\max_{\lambda \in \Lambda} P(\lambda|x^n). \end{aligned} \tag{1}$$

Therefore, the MDL principle is equivalent to choose the model having the maximum posterior probability for the discrete model class [30], which is the Bayes optimal solution for the 0-1 loss function.

[5] Later, we will define another information matrix $I^*(\theta^k)$ for analysis.

Then, the code length of the MDL code $L^{\overline{\theta}^k}_{\mathrm{MDL}}(x^n)$ is given by

$$L^{\overline{\theta}^k}_{\mathrm{MDL}}(x^n) = \min_{\overline{\theta}^k \in \overline{\Theta}^k_n} \left\{ -\log P(x^n|\overline{\theta}^k) - \log \frac{f(\overline{\theta}^k)}{\sqrt{n}^k \sqrt{\det I(\overline{\theta}^k)}} \right\}. \quad (6)$$

On the other hand, the code length of the Bayes code $L^{\theta^k}_{\mathrm{Bayes}}(x^n)$ is given by

$$L^{\theta^k}_{\mathrm{Bayes}}(x^n) = -\log \int_{\theta^k} P(x^n|\theta^k) f(\theta^k)\, d\theta^k \quad (7)$$

where the integral is calculated over the parameter set Θ^{k}[6] [4], [5], [20]. Equation (7) is the Bayes optimal decision for the code length loss function (the logarithmic loss function) for the parametric model class [20].

C. The Hierarchical Model Class

We consider the (partial) hierarchical model class

$$\{P(\cdot|m, \theta^{k_m})|m \in \mathcal{M},\ \theta^{k_m} \in \Theta^{k_m}\}.$$

m is a discrete label for models in the discrete and finite model class $\mathcal{M}$. That is, $\mathcal{M}$ is a finite countable set. If each model m has a k_m-dimensional parameter $\theta^{k_m} = (\theta^{k_m}_1, \theta^{k_m}_2, \ldots, \theta^{k_m}_{k_m})^T$ in a parameter space Θ^{k_m} which is a compact subset of $\mathcal{R}^{k_m}$, then m specifies a parametric model class.

Let $\mathcal{H}^{k_m}$ be the class of the probability distribution of the model m. Then the (partial) hierarchical model class $\mathcal{H}$ is defined by

$$\mathcal{H} = \cup_m \mathcal{H}^{k_m}. \quad (8)$$

We also denote the model class by

$$\mathcal{H} = \{P(\cdot|m, \theta^{k_m})|m \in \mathcal{M}, \theta^{k_m} \in \Theta^{k_m}\}.$$

Here the nested structure

$$\mathcal{H}^{k_{m1}} \subset \mathcal{H}^{k_{m2}} \subset \mathcal{H}^{k_{m3}} \subset \cdots \quad (9)$$

may be satisfied for $m_1, m_2, m_3, \ldots \in \mathcal{M}$ and $k_{m_1} < k_{m_2} < k_{m_3} \cdots$. This nested structure may be linear order or partial order.[7]

Let $-\log P(m)$ be the description length to describe the label m. We assume the prior distribution $f(\theta^{k_m}|m)$ on Θ^{k_m}. For the hierarchical model class, two types of MDL codes can be defined.

The MDL Code Type1: At first, we consider the MDL code which uses an operation of parameter quantization to describe both a quantized parameter $\overline{\theta}^{k_m}$ and a discrete label m. Similarly, with the parametric model class, the code length of the MDL code with parameter quantization is given by

$$L^{m,\overline{\theta}^{k_m}}_{\mathrm{MDL}}(x^n) = \min_{m \in \mathcal{M},\, \overline{\theta}^{k_m} \in \overline{\Theta}^{k_m}_n} \cdot \left\{ -\log P(x^n|m, \overline{\theta}^{k_m}) - \log f(\overline{\theta}^{k_m}|m) \cdot \frac{1}{\sqrt{n}^{k_m} \sqrt{\det I(\overline{\theta}^{k_m}|m)}} - \log P(m) \right\} \quad (10)$$

where $\overline{\Theta}^{k_m}_n$ is a set of all quantized parameter values $\overline{\theta}^{k_m}$ of a model m, and $I(\theta^{k_m}|m)$ is the Fisher information matrix of a model m defined by

$$I(\theta^{k_m}|m) = \lim_{n\to\infty} \frac{1}{n} E_\theta \left\{ -\frac{\partial^2 \log P(X^n|m, \theta^{k_m})}{\partial \theta^{k_m} (\partial \theta^{k_m})^T} \right\}. \quad (11)$$

Here, E_θ is the expectation under $P(X^n|m, \theta^{k_m})$. That is, $\overline{\theta}^{k_m} \in \overline{\Theta}^{k_m}_n$ when x^n is encoded and we assume

$$\sum_{\overline{\theta}^k \in \overline{\Theta}^k_n} f(\overline{\theta}^{k_m}|m) \frac{1}{\sqrt{n}^{k_m} \sqrt{\det I(\overline{\theta}^{k_m}|m)}} = 1.$$

The MDL Code Type2: Second, we define the MDL code which uses the mixture for parameter and selects only a discrete label m

$$L^{m,\theta^{k_m}}_{\mathrm{MDL}}(x^n) = \min_{m \in \mathcal{M}} \{-\log P(x^n|m) - \log P(m)\} \quad (12)$$

$$-\log P(x^n|m) = -\log \int_{\theta^{k_m}} P(x^n|m, \theta^{k_m}) f(\theta^{k_m}|m)\, d\theta^{k_m}. \quad (13)$$

The latter is the minimization of the Bayes code for the parameter and excludes the quantization of the parameter.

The Bayes Code: On the other hand, the Bayes code is given by

$$L^{m,\theta^{k_m}}_{\mathrm{Bayes}}(x^n) = -\log \sum_m \int_{\theta^{k_m}} P(x^n|m, \theta^{k_m}) \cdot p(\theta^{k_m}|m) p(m)\, d\theta^{k_m}. \quad (14)$$

Here, the integral is calculated over the parameter space Θ^{k_m}.

As we have mentioned above, the Bayes code is characterized using the mixture distribution over all models for the coding function! In the following section, we analyze the difference of the code lengths between the MDL code and the Bayes code for the discrete, the parametric, and the hierarchical model classes.

III. Analysis for the Discrete Model Class

In this section, we analyze the difference of code lengths, (2) and (3), for the discrete model class. The data sequence is emitted from the true distribution $P^*(\cdot)$ and we do not assume that $P^*(\cdot)$ exists in the model class.

[6]The mixture code may be also defined by the mixture for the quantized parameters which is given by

$$-\log \sum_{\overline{\theta}^k} P(x^n|\overline{\theta}^k) f(\overline{\theta}^k) \prod_{j=0}^{k-1} \alpha_j(\overline{\theta}^k).$$

However, since the prior density is assumed, the Bayes code as Bayes optimal solution is given by (7) which is the limitation of the above equation as $\alpha_j(\overline{\theta}^k) \to 0$ [23].

[7]For example, the finite Markov sources have linear order structure. The FSMX sources have partial order structure.

A. Assumptions

Let D_0 be the set of λ minimizing $D^*(P^*||P_\lambda)$, where $D^*(P^*||P_\lambda)$ is given by

$$D^*(P^*||P_\lambda) = \min_{n\to\infty} \frac{1}{n} E^* \log \frac{P^*(X^n)}{P(X^n|\lambda)} = \lim_{n\to\infty} \frac{1}{n} \sum_{x^n} P^*(x^n) \log \frac{P^*(x^n)}{P(x^n|\lambda)} \quad (15)$$

where E^* is the expectation under $P^*(\cdot)$. Minimizing $D^*(P^*||P_\lambda)$ is equivalent to minimizing $H^*(\lambda)$, where

$$H^*(\lambda) = \lim_{n\to\infty} \frac{1}{n} E^*\{-\log P(X^n|\lambda)\}. \quad (16)$$

We denote a λ minimizing $D^*(P^*||P_\lambda)$ as λ^*, which may not be unique. That is, $|D_0| \geq 1$. This is so because there may exist λ_1 and λ_2 such that $P(\cdot|\lambda_1) \neq P(\cdot|\lambda_2)$ and $H^*(\lambda_1) = H^*(\lambda_2)$. We can also show a case in which $P(\cdot|\lambda_1) = P(\cdot|\lambda_2)$, which leads to $H^*(\lambda_1) = H^*(\lambda_2)$. In these cases, $|D_0| \geq 2$ if λ_1 and λ_2 minimize $H^*(\lambda)$. We call λ^* the optimal model. Let $D_1 \subset \Lambda$ be the set of λ not minimizing $D^*(P^*||P_\lambda)$. That is, $\mathcal{M} = D_0 \cup D_1$.

We assume the following condition which will be needed in our derivations.

Condition 1:

i) $|D_0| \geq 1$.

ii) For $\forall \lambda \in \Lambda$, $P(\lambda) > 0$.

iii) (The strong law of large numbers) For $\forall \lambda \in \Lambda$

$$-\frac{1}{n} \log P(x^n|\lambda) \to H^*(\lambda) \text{ a.s.} \quad (17)$$

That is, for $\forall \lambda \in \Lambda$ and $\forall \epsilon > 0$

$$\left| -\frac{1}{n} \log P(x^n|\lambda) - H^*(\lambda) \right| < \epsilon \quad (18)$$

is satisfied for all sufficiently large n with probability one. □

We also denote almost sure convergence as, for example,

$$\frac{1}{n}\{-\log P(x^n|\lambda)\} \to H^*(\lambda), \text{ a.s.}$$

or

$$\left| -\frac{1}{n} \log P(x^n|\lambda) - H^*(\lambda) \right| < \epsilon \text{ a.s.}, \qquad \text{when } n \to \infty.$$

If (17) is satisfied, then

$$\frac{1}{n} \log \frac{P(x^n|\lambda_1)}{P(x^n|\lambda_2)} \to D^*(P_{\lambda_1}||P_{\lambda_2}) \text{ a.s.} \quad (19)$$

as $n \to \infty$ is satisfied for $\forall \lambda_1, \lambda_2 \in \Lambda$, where

$$D^*(P_{\lambda_1}||P_{\lambda_2}) = \lim_{n\to\infty} E^* \log \frac{P(X^n|\lambda_1)}{P(X^n|\lambda_2)}. \quad (20)$$

Because $D^*(P_{\lambda^*}||P_\lambda) > 0$ for $\forall \lambda \in D_1$, then we have

$$\frac{P(x^n|\lambda)}{P(x^n|\lambda^*)} \to 0 \text{ a.s.} \quad (21)$$

when $n \to \infty$.

B. Examples for Model Class

Example 1 [Multinomial Independent and Identically Distributed (i.i.d.) Source]: Consider the multinomial i.i.d. source with $\mathcal{X} = \{0, 1, 2, \ldots, \beta\}$. Let p_i be the probability of the symbol i, $i \in \mathcal{X}$ and $p_\beta = 1 - \sum_{i=0}^{\beta-1} p_i$. Then the vector $p^\beta = (p_0, p_1, \ldots, p_{\beta-1})^T$ specifies a probabilistic model.

Let λ_1 mean

$$p^\beta(\lambda_1) = (p_0(\lambda_1), p_1(\lambda_1), \ldots, p_{\beta-1}(\lambda_1))^T$$

where

$$0 < p_0(\lambda_1), p_1(\lambda_1), \ldots, p_{\beta-1}(\lambda_1) < 1.$$

In the same way, let $\lambda_2. \lambda_3, \ldots$ be

$$p^\beta(\lambda_2) = (p_0(\lambda_2), p_1(\lambda_2), \ldots, p_{\beta-1}(\lambda_2))^T$$
$$p^\beta(\lambda_3) = (p_0(\lambda_3), p_1(\lambda_3), \ldots, p_{\beta-1}(\lambda_3))^T$$
$$\ldots.$$

Thus, we can define the model class to be $\Lambda = \{\lambda_1, \lambda_2, \ldots\}$. The true probability is denoted by $p^{\beta *} = (p_0^*, p_1^*, \ldots, p_{\beta-1}^*)^T$, where p_i^* is the true probability of the symbol i. The optimal model $\lambda^* \in \Lambda$ may not be unique in this case.

For example, when $\beta = 2$, we may define discrete models such as

$$p^2(\lambda_1) = (1/10, 1/10)^T$$
$$p^2(\lambda_2) = (1/10, 2/10)^T$$
$$p^2(\lambda_3) = (1/10, 3/10)^T, \ldots, p^2(\lambda_7) = (1/10, 8/10)^T$$
$$p^2(\lambda_8) = (2/10, 1/10)^T, \ldots, p^2(\lambda_{36}) = (8/10, 1/10)^T.$$

Then the model class is $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_{36}\}$. If

$$p_0^* = p_1^* = p_2^* = 1/3$$

then

$$H(\lambda_{18}) = H(\lambda_{19}) = H(\lambda_{25})$$

and

$$D_0 = \{\lambda_{18}, \lambda_{19}, \lambda_{25}\}$$

where

$$p^2(\lambda_{18}) = (3/10, 3/10)$$
$$p^2(\lambda_{19}) = (3/10, 4/10)$$

and

$$p^2(\lambda_{25}) = (4/10, 3/10).$$

For the multinomial i.i.d. source, the strong law of large number is satisfied [8]. That is,

$$\frac{n_i}{n} \to p_i^* \text{ a.s.} \quad (22)$$

as $n \to \infty$ where n_i is the number of times the symbol i appears in x^n. On the other hand, $\log P(x^n|\lambda)$ is given by

$$\log P(x^n|\lambda) = \sum_{i=0}^{\beta} n_i \log p_i(\lambda). \quad (23)$$

From (22), we have

$$-\frac{1}{n} \log P(x^n|\lambda) \to -\sum_{i=0}^{\beta} p_i^* \log p_i(\lambda) \text{ a.s.} \quad (24)$$

when $n \to \infty$ for $\forall \lambda \in \Lambda$, where

$$H^*(\lambda) = -\sum_{i=1}^{\beta} p_i^* \log p_i(\lambda).$$

This model class satisfies Condition 1, iii).

We have, therefore,

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{P(x^n|\lambda^*)}{P(x^n|\lambda)} = D^*(P_{\lambda^*}||P_\lambda) \text{ a.s.} \quad (25)$$

where $D^*(P_{\lambda^*}||P_\lambda)$ is given by

$$D^*(P_{\lambda^*}||P_\lambda) = \sum_{i=0}^{\beta} p_i^* \log \frac{p_i(\lambda^*)}{p_i(\lambda)}.$$
$$= D^*(P^*||P_\lambda) - D^*(P^*||P_{\lambda^*}). \quad (26)$$

On the other hand, $D^*(P^*||P_\lambda) - D^*(P^*||P_{\lambda^*}) > 0$ is satisfied for all $\lambda^* \in D_0$ and all $\lambda \in D_1$ by definition. □

Example 2 (Finite Ergodic Markov Source): We consider a finite ergodic Markov source on $\mathcal{X} = \{0, 1, 2, \ldots, \beta\}$. Let $p_{i,j}$ be the probability of the symbol i at the jth state s_j, where $i \in \mathcal{X}$ and $j = 0, 1, \ldots, S$. Let q_{s_j} be the stationary probability of the state s_j, where $j = 0, 1, \ldots, S$, which are uniquely decided by $p_{i,j}$, $i \in \mathcal{X}$ and $j = 0, 1, \ldots, S$. Those of model λ are denoted by $p_{i,j}(\lambda)$ and $q_{s_j}(\lambda)$. Let a model λ be

$$\lambda = \{p_{i,j}(\lambda) | i = 0, \ldots, \beta - 1, j = 0, \ldots, S\}$$

and the model λ^*

$$\lambda^* = \{p_{i,j}(\lambda^*) | i = 0, \ldots, \beta - 1, j = 0, \ldots, S\}$$

where $\lambda^* \in D_0$.

We assume the initial state is known. Let $n_0, \ldots, n_S$ be the numbers of times the states $s_0, \ldots s_S$ appear in x^n, respectively. Let $n_{0,j}, n_{1,j}, \ldots, n_{\beta,j}$ be the numbers of times the symbols $0, 1, \ldots, \beta$ appear conditioned by the state s_j in data sequence x^n, respectively. That is, $n = \sum_j n_j$ and $n_j = \sum_i n_{i,j}$. The true probability of the symbol i at the jth state s_j is denoted by $p_{i,j}^*$, $i = 0, \ldots, \beta - 1$, $j = 0, \ldots, S$, where $p_{i,j}^*$ is given by

$$p_{i,j}^* = \lim_{n\to\infty} \frac{1}{n} E^* N_{i,j}. \quad (27)$$

That is,

$$\{p_{i,j}^* | i = 0, \ldots, \beta - 1, j = 0, \ldots, S\}$$

is given by

$$\arg \min_{\{p_{i,j}\}} \lim_{n\to\infty} E^* \sum_{i,j} \frac{N_{i,j}}{n} \log p_{i,j} \quad (28)$$

where $N_{i,j}$ are random variables representing the numbers of times each symbols $0, 1, \ldots, \beta$, appear conditioned by the states s_j, $j = 0, \ldots, S$ in the random variable X^n. The stationary probabilities on the states calculated by $p_{i,j}^*$ are denoted by $q_{s_0}^* \cdots q_{s_S}^*$. In this case, it is known that the strong law of large number is satisfied, that is,

$$\frac{n_j}{n} \to q_{s_j}^* \text{ a.s.} \quad (29)$$

$$\frac{n_{i,j}}{n_j} \to p_{i,j}^* \text{ a.s.} \quad (30)$$

when $n \to \infty$.

We may define model class Λ in the same way as in Example 1. That is, the model may be defined by

$$p_{i,j} \in \{1/2, 1/3, \ldots\}$$

and $\log P(x^n|\lambda)$ is given by

$$\log P(x^n|\lambda) = \sum_{i,j} n_{i,j} \log p_{i,j}(\lambda). \quad (31)$$

Therefore, we have

$$\frac{1}{n} \log P(x^n|\lambda) = \sum_{i,j} \frac{n_j}{n} \frac{n_{i,j}}{n_j} \log p_{i,j}(\lambda)$$
$$\to \sum_{i,j} q_{s_j}^* p_{i,j}^* \log p_{i,j}(\lambda) \text{ a.s.} \quad (32)$$

when $n \to \infty$. This model class satisfies Condition 1, iii).

We have, therefore,

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{P(x^n|\lambda^*)}{P(x^n|\lambda)} = D^*(P_{\lambda^*}||P_\lambda) \text{ a.s.} \quad (33)$$

where $D^*(P_{\lambda^*}||P_\lambda)$ is given by

$$D^*(P_{\lambda^*}||P_\lambda) = \sum_{j=0}^{S} q_{s_j}^* \sum_{i=0}^{\beta} p_{i,j}^* \log \frac{p_{i,j}^*}{p_{i,j}(\lambda)}. \quad (34)$$

$D^*(P^*||P_\lambda) - D^*(P^*||P_{\lambda^*}) > 0$ is satisfied for $\forall \lambda^* \in D_0$ and $\forall \lambda \in D_1$. □

C. Main Results for the Discrete-Model Class

First, we show the following lemma.

Lemma 1: Under Condition 1, the asymptotic code length of the Bayes code is given by

$$L_{\text{Bayes}}^{\lambda}(x^n) = -\log \sum_{\lambda^* \in D_0} P(x^n|\lambda^*) P(\lambda^*) - o^+(1) \text{ a.s.} \quad (35)$$

where $o^+(1)$ is the positive term such that $o^+(1) \to +0$, a.s., when $n \to \infty$.

Proof: From the definition of D_0, we have

$$L_{\text{Bayes}}^{\lambda}(x^n) = -\log \left\{ \sum_{\lambda^* \in D_0} P(x^n|\lambda^*) P(\lambda^*) + \sum_{\lambda \in D_1} P(x^n|\lambda) P(\lambda) \right\}$$
$$= -\log \sum_{\lambda^* \in D_0} P(x^n|\lambda^*) P(\lambda^*) - \log \left\{ 1 + \frac{\sum_{\lambda \in D_1} P(x^n|\lambda) P(\lambda)}{\sum_{\lambda^* \in D_0} P(x^n|\lambda^*) P(\lambda^*)} \right\}. \quad (36)$$

Because Λ is finite and $\frac{P(x^n|\lambda)}{P(x^n|\lambda^*)} \to 0$, a.s., when $n \to \infty$ for all $\lambda^* \in D_0$ and all $\lambda \in D_1$ from (21), we have

$$\frac{\sum_{\lambda \in D_1} P(x^n|\lambda) P(\lambda)}{\sum_{\lambda^* \in D_0} P(x^n|\lambda^*) P(\lambda^*)} = o^+(1) \text{ a.s.} \quad (37)$$

Therefore, we have

$$\log\left\{1+\frac{\sum_{\lambda\in D_1}P(x^n|\lambda)P(\lambda)}{\sum_{\lambda^*\in D_0}P(x^n|\lambda^*)P(\lambda^*)}\right\}=o^+(1)\ \text{ a.s.} \quad (38)$$

Thus, the proof is completed. □

Lemma 1 is used in the proof of the following main results. In some cases, we can assume that two or more identical models are not included in the model class and the optimal model is unique. When $|D_0|=1$, the code length is given by

$$L^{\lambda}_{\text{Bayes}}(x^n)=-\log P(x^n|\lambda^*)-\log P(\lambda^*)-o^+(1)\ \text{ a.s.} \quad (39)$$

Next, we show the key theorem in order to analyze the difference of the code lengths between the MDL codes and the Bayes codes.

Theorem 1: For the same prior probability, the relation between (2) and (3) is given as follows:

$$L^{\lambda}_{\text{MDL}}(x^n)=L^{\lambda}_{\text{Bayes}}(x^n)-\log P(\tilde{\lambda}|x^n). \quad (40)$$

Here, $\tilde{\lambda}$ represents the model which maximizes the posterior probability $P(\lambda|x^n)$.

Proof: From Bayes rule

$$P(x^n|\lambda)P(\lambda)=P(\lambda|x^n)P(x^n)$$

(40) is obviously obtained. □

This theorem shows that the code length of the Bayes code is smaller than that of the MDL code by the factor $-\log P(\tilde{\lambda}|x^n)$ on the same prior, and the Bayes code is effective for the finite length of the data sequence. Next, we consider the order of the term $-\log P(\tilde{\lambda}|x^n)$.

Usually, the optimal model λ^* is unique. First, we show the convergence rate of the difference of the code lengths in this case.

Theorem 2: Under Condition 1, if λ^* is unique, that is, $|D_0|=1$, then the relation between (2) and (3) is given by

$$L^{\lambda}_{\text{MDL}}(x^n)=L^{\lambda}_{\text{Bayes}}(x^n)+o^+(1)\ \text{ a.s.} \quad (41)$$

where $o^+(1)$ is the positive term such that $o^+(1)\to+0$, a.s., as $n\to\infty$.

Proof: From

$$P(\lambda|x^n)\propto P(x^n|\lambda)P(\lambda) \quad (42)$$

and (21), we have

$$\frac{P(\lambda|x^n)}{P(\lambda^*|x^n)}=o^+(1)\ \text{ a.s.} \quad (43)$$

for $\forall\,\lambda^*\in D_0$ and $\forall\,\lambda\in D_1$. Therefore, we have

$$P(\lambda^*|x^n)=1-o^+(1)\ \text{ a.s.} \quad (44)$$

Hence, the model λ which maximizes the posterior probability $P(\lambda|x^n)$ almost surely corresponds to the true model λ^* asymptotically, that is, $\tilde{\lambda}=\lambda^*$ for all sufficiently large n with probability one (strong consistency, see [15] and [32]). Then we can substitute $\tilde{\lambda}$ for λ^* when $n\to\infty$. Therefore, we have

$$-\log P(\tilde{\lambda}|x^n)=o^+(1)\ \text{ a.s.} \quad (45)$$

The proof is completed. □

This theorem implies that the difference of the code lengths between the MDL code and the Bayes code converges to 0 when the optimal model is unique. Then, the code lengths for both codes are asymptotically equal.

Next, we consider the case $|D_0|\geq 2$.

Theorem 3: Under Condition 1, if $|D_0|\geq 2$ and

$$P(x^n|\lambda_1^*)=P(x^n|\lambda_2^*),\qquad \text{for}\ \forall\,\lambda_1^*,\ \lambda_2^*\in D_0$$

then, the relation between (2) and (3) is

$$L^{\lambda}_{\text{MDL}}(x^n)=L^{\lambda}_{\text{Bayes}}(x^n)+O^+(1)\ \text{ a.s.} \quad (46)$$

where $O^+(1)$ is the positive term such that $O^+(1)\to C$, a.s., when $n\to\infty$, and the positive constant C is given by

$$C=-\log\frac{\max_{\lambda^*\in D_0}P(\lambda^*)}{\sum_{\lambda^*\in D_0}P(\lambda^*)}. \quad (47)$$

Proof: From the equation

$$-\log P(\lambda|x^n)=-\log P(x^n|\lambda)-\log P(\lambda)+\log P(x^n) \quad (48)$$

and Lemma 1, we have

$$-\log P(\lambda|x^n)=-\log\frac{P(x^n|\lambda)P(\lambda)}{\sum_{\lambda^*\in D_0}P(x^n|\lambda^*)P(\lambda^*)}+o^+(1)\ \text{ a.s.} \quad (49)$$

From (43), we have $-\log P(\lambda|x^n)\to+\infty$, a.s., for $\forall\,\lambda\in D_1$.

On the other hand, for $\lambda^*\in D_0$, we have

$$-\log P(\lambda|x^n)=-\log\frac{P(\lambda)}{\sum_{\lambda^*\in D_0}P(\lambda^*)}+o(1)\ \text{ a.s.} \quad (50)$$

because

$$P(x^n|\lambda_1^*)=P(x^n|\lambda_2^*),\qquad \text{for}\ \forall\,\lambda_1^*,\ \lambda_2^*\in D_0.$$

The first term of the right-hand side (RHS) of (50) is independent of the sample size n and dominates the equation. Then, the code length of the MDL code in this case is given by

$$L^{\lambda}_{\text{MDL}}(x^n)=L^{\lambda}_{\text{Bayes}}(x^n)-\log\max_{\lambda^*\in D_0}P(\lambda^*)+\log\left\{\sum_{\lambda^*\in D_0}P(\lambda^*)\right\}+o(1)\ \text{ a.s.} \quad (51)$$

So the following equation is obtained:

$$L^{\lambda}_{\text{MDL}}(x^n)=L^{\lambda}_{\text{Bayes}}(x^n)+O^+(1)\ \text{ a.s.} \quad (52)$$

where $O^+(1)$ is the term such that $O^+(1) \to C$, a.s., when $n \to \infty$.

From Condition 1, ii), we have

$$-\log \max_{\lambda^* \in D_0} P(\lambda^*) > -\log \sum_{\lambda^* \in D_0} P(\lambda^*). \tag{53}$$

Therefore, the constant C is positive, and the proof is completed. □

This result may be interpreted as follows. The model selection by the MDL principle for the discrete model class is essentially equivalent to that of maximization of the posterior probability. If $\lambda^* \in D_0$ is unique, then λ^* will be asymptotically obtained by maximization of $P(\lambda|x^n)$. However, if there are models λ_1^* and λ_2^* satisfying $\lambda_1^* \neq \lambda_2^*$, $\lambda_1^*, \lambda_2^* \in D_0$, and $P(\cdot|\lambda_1^*) = P(\cdot|\lambda_2^*)$, then the model which has maximum prior probability in D_0 will be asymptotically selected by maximization of $P(\lambda|x^n)$. That is, when the optimal model is not unique, the posterior probability of an optimal model $P(\lambda^*|x^n)$ does not approach 1, so that uncertainty asymptotically remains for model selection. This uncertainty in the model selection makes the code length of the MDL code larger.

On the other hand, the Bayes code uses the mixture $\sum_\lambda P(x^n|\lambda)P(\lambda)$. Consider a case such that there is a model λ_2^* satisfying

$$P(\cdot|\lambda_1^*) = P(\cdot|\lambda_2^*), \qquad \lambda_1^* \neq \lambda_2^*, \lambda_1^*, \lambda_2^* \in D_0.$$

If we define λ_c^* as

$$P(\lambda_c^*) = \sum_{\lambda^* \in D_0} P(\lambda^*)$$

and

$$P(x^n|\lambda_c^*) = P(x^n|\lambda_1^*)$$

for λ_1^* in D_0, remove $\forall \lambda^*$ satisfying $\lambda^* \neq \lambda_1^*$ and $\lambda^* \in D_0$, and construct the Bayes code for such a reconstructed model class, then its code length is asymptotically equivalent to that of the original Bayes code and we can regard λ_c^* as the unique optimal model. Thus, the code length of the Bayes code is not made larger even if there exist models $\lambda_1^* \neq \lambda_2^*$, $\lambda_1^*, \lambda_2^* \in D_0$, and $P(\cdot|\lambda_1^*) = P(\cdot|\lambda_2^*)$.

Next we consider the case that there are models, λ_1^* and λ_2^*, satisfying $P(\cdot|\lambda_1^*) \neq P(\cdot|\lambda_2^*)$ and $\lambda_1^* \neq \lambda_2^*$, $\lambda_1^*, \lambda_2^* \in D_0$. For most practical model classes, if $P(\cdot|\lambda_1^*) \neq P(\cdot|\lambda_2^*)$ for $\lambda_1^* \neq \lambda_2^*$, $\lambda_1^*, \lambda_2^* \in D_0$, then

$$\log \frac{P(x^n|\lambda_1^*)}{P(x^n|\lambda_2^*)} \to +\infty \quad \text{a.s.} \tag{54}$$

or

$$\log \frac{P(x^n|\lambda_1^*)}{P(x^n|\lambda_2^*)} \to -\infty \quad \text{a.s.} \tag{55}$$

Example 3 (Multinomial i.i.d. Source): Again consider Example 1. Since $\frac{n_i}{n} \to p_i^*$, a.s., we have $n_i \to \infty$, almost surely. On the other hand, for all $\lambda_1, \lambda_2 \in \Lambda$we have

$$\log \frac{P(x^n|\lambda_1)}{P(x^n|\lambda_2)} = \sum_{i=0}^{\beta} n_i \log \frac{p_i(\lambda_1)}{p_i(\lambda_2)}. \tag{56}$$

Because $n_i \to \infty$, a.s., if $p_i(\lambda_1) \neq p_i(\lambda_2)$ then $\log \frac{P(x^n|\lambda_1)}{P(x^n|\lambda_2)} \to \pm\infty$.

Of course,

$$\frac{1}{n} \log \frac{P(x^n|\lambda_1^*)}{P(x^n|\lambda_2^*)} \to 0, \text{ a.s.} \tag{57}$$

holds for $\lambda_1^*, \lambda_2^* \in D_0$. Nevertheless, $\log \frac{P(x^n|\lambda_1^*)}{P(x^n|\lambda_2^*)} \to 0$, a.s. does not hold. $\log \frac{P(x^n|\lambda_1^*)}{P(x^n|\lambda_2^*)}$ diverges almost surely with an order $O(\sqrt{\frac{\log\log n}{n}})$. Similarly, finite ergodic Markov sources also satisfy (54) or (55). See [8]. □

Theorem 4: Assume Condition 1. If $|D_0| \geq 2$ and

$$P(x^n|\lambda_1^*) \neq P(x^n|\lambda_2^*), \qquad \text{for } \forall \lambda_1^*, \lambda_2^* \in D_0$$

and (54) or (55) hold, then the relation between (2) and (3) is given as follows:

$$L_{\text{MDL}}^{\lambda}(x^n) = L_{\text{Bayes}}^{\lambda}(x^n) + o^+(1) \quad \text{a.s.} \tag{58}$$

Proof: Since (54) or (55) hold, we have

$$\frac{P(\lambda|x^n)}{P(\tilde{\lambda}|x^n)} = o^+(1) \quad \text{a.s.} \tag{59}$$

for $\lambda \neq \tilde{\lambda}, \forall \lambda \in \Lambda$. So analogy with Theorem 2 leads to (58). □

IV. Analysis for the Parametric Model Class

In this section, we discuss the difference between the code lengths from (6) and (7), for the parametric model class $\{P(\cdot|\theta^k)|\theta^k \in \Theta^k\}$. The data sequence x^n is emitted from $P^*(x^n)$ and we do not assume that $P^*(\cdot)$ is in the model class. We define the information matrix $I^*(\theta^k)$ as follows:

$$I^*(\theta^k) = \lim_{n\to\infty} \frac{1}{n} E^* \left\{ -\frac{\partial^2 \log P(X^n|\theta^k)}{\partial\theta^k (\partial\theta^k)^T} \right\} \tag{60}$$

in which $E^*\{\cdot\}$ is the expectation under $P^*(\cdot)$. $I^*(\theta^{k*}) = I(\theta^{k*})$ is not generally satisfied. From the definition $I^*(\theta^{k*}) = I(\theta^{k*})$ when $P^*(\cdot) \in \{P(\cdot|\theta^k)|\theta^k \in \Theta^k\}$. However, for most of practical model classes for source coding, $I^*(\theta^{k*}) = I(\theta^{k*})$ is satisfied even if $P^*(\cdot) \neq P(\cdot|\theta^{k*})$ because $\mathcal{X}$ is discrete. See Examples 4 and 5.

We also define $H^*(\theta^k)$ as follows:

$$H^*(\theta^k) = \lim_{n\to\infty} \frac{1}{n} E^*\{-\log P(X^n|\theta^k)\}. \tag{61}$$

Let θ^{k*} be the optimal parameter given by

$$\theta^{k*} = \arg\min_{\theta^k} H^*(\theta^k). \tag{62}$$

We denote the maximum-likelihood estimator and the maximum posterior estimator given x^n by $\hat{\theta}^k$ and $\tilde{\theta}^k$, respectively,

$$\hat{\theta}^k = \arg\max_{\theta^k} \log P(x^n|\theta^k) \tag{63}$$

$$\tilde{\theta}^k = \arg\max_{\theta^k} \log f(\theta^k|x^n). \tag{64}$$

A. Assumptions

Defining

$$B_\delta(\theta^{k*}) = \{\theta^k \in \Theta^k | \|\theta^k - \theta^{k*}\| < \delta\}$$

we assume the following conditions for the parametric model class and Bayesian inference.

Condition 2:

i) (Existence of θ^{k*}) The function $-H^*(\theta^k)$ is a unimodal function with a maximum point in the interior of Θ^k. That is, the optimal parameter θ^{k*} uniquely exists in the interior of Θ^k.

ii) (Smoothness of class) The Fisher information matrix $I(\theta^k)$ satisfies $0 < C_0 < \det I(\theta^k) < \infty$ for $\forall\, \theta^k \in \Theta^k$, where C_0 is a positive constant. When $\theta^k = \theta^{k*}$, $I(\theta^{k*}) = I^*(\theta^{k*})$. Moreover, $\det I(\theta^k)$ and $\det I^*(\theta^k)$ satisfy

$$\left\| \frac{\partial \det I(\theta^k)}{\partial \theta^k} \right\| < \infty \tag{65}$$

$$\left\| \frac{\partial \det I^*(\theta^k)}{\partial \theta^k} \right\| < \infty \tag{66}$$

for $\forall\, \theta^k \in \Theta^k$.

iii) (Smoothness of the prior) For $\forall\, \theta^k \in \Theta^k$, $f(\theta^k) > 0$ and $f(\theta^k)$ is three times continuously differentiable for θ^k. That is, we have

$$\left| \sum_{i=0}^{k-1} \sum_{j=0}^{k-1} \sum_{l=0}^{k-1} \frac{\partial f(\theta^k)}{\partial \theta_i \partial \theta_j \partial \theta_l} \right| < c_f(\theta^k) \tag{67}$$

where $c_f(\theta^k)$ is a finite constant depending only on θ^k.

iv) (Existence of estimators) The likelihood function $P(x^n|\theta^k)$ given x^n is unimodal or monotonic with respect to $\theta^k \in \Theta^k$ for $\forall\, x^n \in \mathcal{X}^n$. The posterior density function $f(\theta^k|x^n)$ has a unique maximum in Θ^k for $\forall\, x^n \in \mathcal{X}^n$. That is, both of the maximum-likelihood estimator $\hat{\theta}^k$ and the maximum posterior estimator $\tilde{\theta}^k$ uniquely exist in Θ^k.

v) (Consistency of the maximum posterior estimator) The maximum posterior estimator $\tilde{\theta}^k$ is strongly consistent. That is,

$$\left\| \tilde{\theta}^k - \theta^{k*} \right\| \to 0 \quad \text{a.s.} \tag{68}$$

when $n \to \infty$.

vi) (Consistency of the information matrix) There exists $\delta > 0$, such that

$$-\frac{1}{n} \log P(x^n|\theta^k) \to H^*(\theta^k) \quad \text{a.s.} \tag{69}$$

$$-\frac{1}{n} \frac{\partial^2 \log P(x^n|\theta^k)}{\partial \theta^k (\partial \theta^k)^T} \to I^*(\theta^k) \quad \text{a.s.} \tag{70}$$

uniformly for all $\theta^k \in B_\delta(\theta^{k*})$. □

Because

$$\frac{1}{n} \log f(\theta^k|x^n) \propto \frac{1}{n} \log P(x^n|\theta^k) + \frac{1}{n} f(\theta^k) \to \frac{1}{n} \log P(x^n|\theta^k)$$

for all $\theta^k \in \Theta^k$, and

$$-\frac{1}{n} \log P(x^n|\theta^k) \to H^*(\theta^k) \quad \text{a.s.}$$

uniformly for all $\theta^k \in B_\delta(\theta^k)$, if $\tilde{\theta}^k \to \hat{\theta}^k$, a.s., is not satisfied, so (68) is also not satisfied. Therefore, (68) means $\hat{\theta}^k \to \theta^{k*}$ a.s.

From Condition 2, i), iv), and v), the likelihood function $P(x^n|\theta^k)$ and the posterior density function $f(\theta^k|x^n)$ given x^n are almost surely unimodal functions when $n \to \infty$.

Here, we just show the asymptotic normality of the maximum-likelihood estimator since it is used in the discussion in mean code length of the MDL codes [14]. However, it will not be directly used in the proof of the results in this paper.

Condition 3: The distribution of $\eta^k = \sqrt{n}(\hat{\theta}^k - \theta^{k*})$ converges to a normal distribution with mean zero and covariance matrix $\{I(\theta^{k*})\}^{-1}$. Hence, in particular, if R_η^k is an arbitrary k-dimensional rectangle, its probability induced by $P(x^n|\theta^{k*})$ satisfies

$$P(\eta^k \in R_\eta^k) = \sum_{\eta^k \in R_\eta^k} P(\eta^k|\theta^{k*}) \to \frac{\sqrt{\det I(\theta^{k*})}}{(2\pi)^{k/2}} \int_{\eta^k \in R_\eta^k} e^{-\frac{1}{2}\|\eta^k\|^2_{I(\theta^{k*})}} d\eta^k. \tag{71}$$

□

Remark 1: Next we consider Condition 2, iii). As an example, the Dirichlet distribution is the conjugate prior for the multinomial distribution class and is obviously three times continuously differentiable. This prior is also useful for the Markov model [21]. □

Remark 2: We consider Condition 2, v) and vi). For many practical classes used for source coding, e.g., a finite ergodic Markov source, the iterated logarithm law of the maximum-likelihood estimator is satisfied, which leads to the strong consistency of the maximum-likelihood estimator, Condition 2, v) [8]. Moreover, this leads to Condition 2, vi), in practical cases, see Examples 4 and 5. □

B. Examples for Model Class

Next, we show examples of the parametric model classes satisfying Condition 2. The model classes of these examples are useful for the source coding.

Example 4 (Multinomial i.i.d. Source): Consider the multinomial i.i.d. source on $\mathcal{X} = \{0, 1, 2, \ldots, \beta\}$. Let θ_i be the probability of the symbol i, $i \in \mathcal{X}$. The vector $\theta^k = (\theta_0, \theta_1, \ldots, \theta_{k-1})^T$ specifies the probabilistic model, where $k = \beta$. Letting θ^k be continuous parameter on

$$\Theta^k = \{(\theta_0, \theta_2, \ldots, \theta_{k-1})^T | 0 < \theta_0, \theta_2, \ldots, \theta_{k-1}, \theta_k < 1\}$$

where $\theta_k = 1 - \sum_{i=0}^{k-1} \theta_i$, this model class is a parametric model class.

We assume that θ^{k*} exists in the interior of Θ^k. The information matrix $I^*(\theta^k)$ is given by

$$I^*(\theta^k) = -\frac{\partial^2 \sum_{i=0}^{k} \theta_i^* \log \theta_i}{\partial \theta^k (\partial \theta^k)^T} \tag{72}$$

where the θ_i^*'s are given by $\theta_i^* = E^*\{\frac{N_i}{n}\}$ for $n \in \{1, 2, \ldots\}$ and N_i's are random variables representing the appearance numbers of symbols $0, 1, \ldots \beta$ in X^n, respectively. That is, the (i, i)th element of $I^*(\theta^k)$ is given by

$$\frac{\theta_i^*}{(\theta_i)^2} + \frac{\theta_k^*}{\left(1 - \sum_{j=1}^{k-1} \theta_j\right)^2} \tag{73}$$

and the (i, j)th element of $I^*(\theta^k)$ is given by

$$\frac{\theta_k^*}{\left(1 - \sum_{j=1}^{k-1} \theta_j\right)^2}, \tag{74}$$

where $\theta^{k*} = (\theta_0^*, \theta_1^*, \ldots, \theta_{k-1}^*)^T$. Therefore, $\det I^*(\theta^k)$ is obviously differentiable. When $\theta^k = \theta^{k*}$, $I^*(\theta^{k*})$ reduces to $I(\theta^{k*})$. Since the determinant of $I(\theta^k)$ is given by

$$\det I(\theta^k) = \frac{1}{\theta_0 \theta_1 \cdots \theta_k} \tag{75}$$

$\det I(\theta^k)$ is minimized when $\theta_0 = \theta_1 = \cdots = \theta_k$ and its minimum value is given by

$$\min \det I(\theta^k) = (k+1)^{k+1} > 0. \tag{76}$$

Since $0 < \theta_j < 1$ for $j \in \{0, 1, \ldots, k\}$, we have $\det I(\theta^k) < \infty$ and $\| \frac{\partial \det I(\theta^k)}{\partial \theta^k} \| < \infty$. Therefore, Condition 2, i) is satisfied. Since

$$\lim_{\theta_j \to 0} \sqrt{\det I(\theta^k)} \to \infty, \qquad \text{for } \forall j \in \{0, 1, \ldots, k\}$$

we cannot assume $\sqrt{\det I(\theta^k)} < C'$ for some constant C' for the multinomial distribution class.

The likelihood function $P(x^n|\theta^k)$ is given by

$$P(x^n|\theta^k) = \prod_{i=0}^{k} (\theta_i)^{n_i} \tag{77}$$

and this function has a unique maximum in Θ^k for $x^n \in \mathcal{X}^n$, where n_i is the appearance number of the symbol i in x^n. The maximum-likelihood estimator is given by $\hat{\theta}_i = n_i/n$, $i = 0, 1, \ldots, k$. It was shown in [8] that this model class satisfies the strong law of large numbers, that is, $\lim_{n\to\infty} \frac{n_i}{n} \to \theta_i^* > 0$, a.s. That is, $\hat{\theta}^k \to \theta^{k*}$, a.s. If we assume the Dirichlet prior density on Θ^k, then $\tilde{\theta}_i$ is given by

$$\tilde{\theta}_i = \frac{n_i + \gamma_i - 1}{n + \sum_j \gamma_j - k - 1}$$

where γ_i is a parameter of the Dirichlet prior density and Condition 2, vi) is satisfied. Therefore, we have $\tilde{\theta}^k \to \hat{\theta}^k \to \theta^{k*}$, a.s. in this case. If $\sup_{\theta^k \in \Theta^k} f(\theta^k) < \infty$, then

$$\frac{1}{n} \log P(x^n|\theta^k) + \frac{1}{n} f(\theta^k) \to \frac{1}{n} \log P(x^n|\theta^k)$$

uniformly for $\theta^k \in \Theta^k$ and we have $\tilde{\theta}^k \to \theta^{k*}$, a.s. Then, Condition 2, v), is satisfied in these cases. Moreover, Condition 2, iv) is obviously satisfied.

Next we consider Condition 2, vi). $\frac{1}{n} \frac{\partial^2 \log P(x^n|\theta)}{\partial \theta^k (\partial \theta^k)^T}$ is given by

$$\frac{1}{n} \frac{\partial^2 \log P(x^n|\theta)}{\partial \theta^k (\partial \theta^k)^T} = \frac{\partial^2 \sum_{i=0}^{k} \frac{n_i}{n} \log \theta_i}{\partial \theta^k (\partial \theta^k)^T}. \tag{78}$$

That is, the (i, i)th element of $\frac{1}{n} \frac{\partial^2 \log P(x^n|\theta)}{\partial \theta^k (\partial \theta^k)^T}$ is

$$-\frac{n_i}{n(\theta_i)^2} - \frac{n_k}{n\left(1 - \sum_{j=0}^{k-1} \theta_j\right)^2} \tag{79}$$

and the (i, j)th element of $\frac{1}{n} \frac{\partial^2 \log P(x^n|\theta)}{\partial \theta^k (\partial \theta^k)^T}$ is

$$-\frac{n_k}{n\left(1 - \sum_{j=0}^{k-1} \theta_j\right)^2} \tag{80}$$

when $i \neq j$. Since $n_i/n \to \theta_i^*$, a.s., we have

$$\frac{1}{n} \log P(x^n|\theta^k) \to \sum_{i=0}^{k} \theta_i^* \log \theta_i \quad \text{a.s.} \tag{81}$$

and

$$\frac{1}{n} \frac{\partial^2 \log P(x^n|\theta^k)}{\partial \theta^k (\partial \theta^k)^T} \to \frac{\partial^2 \sum_{i=0}^{k} \theta_i^* \log \theta_i}{\partial \theta^k (\partial \theta^k)^T} \quad \text{a.s.} \tag{82}$$

Therefore, there exists $\delta > 0$ so that $-\frac{1}{n} \log P(x^n|\theta) \to H^*(\theta^k)$, a.s., and

$$-\frac{1}{n} \frac{\partial^2 \log P(x^n|\theta)}{\partial \theta^k (\partial \theta^k)^T} \to I^*(\theta^k) \text{ a.s.}$$

uniformly for $\theta^k \in B_\delta(\theta^{k*})$. Then Condition 2, vi) is satisfied.

Therefore, this model class satisfies Condition 2 for the multinomial distribution class. When θ^{k*} exists in the interior of Θ^k, we can also see that Condition 3 is satisfied from [8]. □

Example 5 (Finite Ergodic Markov Source): Consider the finite ergodic Markov source on $\mathcal{X} = \{0, 1, 2, \ldots, \beta\}$. Let $\theta_{i,j} = p_i^{(s_j)}$ be the probability of the symbol i at the jth state s_j, and q_{s_j}, the stationary probability of the states s_j, where $i \in \mathcal{X}$ and $j = 0, 1, \ldots, S$. We assume that the set of states is known but the optimal parameter $\theta_{i,j}^*$ is unknown. We may regard $\theta^k = (\theta_{0,0}, \ldots, \theta_{\beta-1,S})^T$ as a continuous parameter, where $k = \beta(S+1)$.

Let $N_{i,j}$ be random variables representing the appearance numbers of the symbols $0, 1, \ldots, \beta$ conditioned on the states $s_j, j = 0, \ldots, S$, in the random variable X^n. We define

$$p^*_{i,j} = \lim_{n\to\infty} \tfrac{1}{n} E^* N_{i,j}.$$

Then the optimal parameter is given by

$$\theta^{k*} = (\theta^*_{0,0}, \ldots, \theta^*_{\beta-1,S})^T = (p^*_{0,0}, \ldots, p^*_{\beta-1,S})^T.$$

The stationary probabilities on the states calculated by $p^*_{i,j}$ are denoted by $q^*_{s_0} \cdots q^*_{s_S}$.

The information matrix $I^*(\theta^k)$ is given by

$$I^*(\theta^k) = -\frac{\partial^2 \sum_{j=0}^{S} \sum_{i=0}^{\beta} q^*_{s_j} \theta^*_{i,j} \log \theta_{i,j}}{\partial \theta^k (\partial \theta^k)^T} \tag{83}$$

where

$$\theta_{\beta,j} = 1 - \sum_{i=0}^{\beta-1} \theta_{i,j}$$

and

$$\theta^*_{\beta,j} = 1 - \sum_{i=0}^{\beta-1} \theta^*_{i,j}.$$

So $\det I^*(\theta^k)$ is obviously differentiable. When $\theta^k = \theta^{k*}$, $I^*(\theta^{k*}) = I(\theta^{k*})$. The determinant of $I(\theta^k)$ is given by

$$\det I(\theta^k) = \prod_{j=0}^{S} (q_{s_j})^{\beta} \frac{1}{\theta_{0,j}\theta_{1,j}\cdots\theta_{\beta,j}}. \tag{84}$$

Here q_{s_j} depends on θ^k, and it is difficult to derive the general formula of $\det I(\theta^k)$ using only θ^k. However, the structure of the finite ergodic Markov source is given, q_{s_j} can be written as θ^k, and we can see $C_0 < \det I(\theta^k) < \infty$. For example, for the binary first-order Markov chain such that $\mathcal{X} = \{0, 1\}$, $S = 1$, and s_i means that the current symbol is $i \in \{0, 1\}$, q_{s_0} and q_{s_1} are given by $\frac{\theta_{0,1}}{\theta_{1,0}+\theta_{0,1}}$, and $\frac{\theta_{1,0}}{\theta_{1,0}+\theta_{0,1}}$, respectively. The $\det I(\theta^k)$ is therefore given by

$$\det I(\theta^k) = \frac{1}{(\theta_{1,0}+\theta_{0,1})^2 \theta_{0,0}\theta_{1,1}}. \tag{85}$$

When $\theta_{0,0} = \theta_{1,1} = 1/2$, the above function is minimized and its minimum value is 4. Letting $\theta_{0,0} \to 0$ or $\theta_{1,1} \to 0$ or $\theta_{1,0} + \theta_{0,1} \to 0$, $\det I(\theta^k) \to \infty$. Thus, $\det I(\theta^k)$ is not upper-bounded, but is lower-bounded by some positive constant C_0. Therefore, this model class also satisfies Condition 2, ii).

We assume the initial state is known again. We define $n_0 \cdots n_S$ and $n_{0,j}, n_{1,j}, \ldots, n_{\beta,j}$ similar to Example 2. That is, $n = \sum_j n_j$ and $n_j = \sum_i n_{i,j}$. The likelihood function $P(x^n|\theta^k)$ is given by

$$P(x^n|\theta^k) = \prod_{j=0}^{S} \prod_{i=0}^{\beta} (\theta_{i,j})^{n_{i,j}} \tag{86}$$

and this function has a unique maximum in Θ^k. It is known that the strong law of large number is satisfied [8], that is,

$$\frac{n_{i,j}}{n_j} \to \theta^*_{i,j} \quad \text{a.s.} \tag{87}$$

which is equivalent to Condition 2, v). If Condition 2, i) is satisfied, then Condition 2, iv) is also satisfied.

Moreover, since $\log P(x^n|\theta^k)$ is given by

$$\log P(x^n|\theta^k) = \sum_{i,j} n_{i,j} \log \theta_{i,j} \tag{88}$$

and (87) is satisfied, we see that Condition 2, vi) is satisfied by a discussion similar to that for Example 4.

Therefore, the finite ergodic Markov source satisfies Condition 2. We also see that Condition 3 is satisfied from [8]. □

We do not retain the assumption of i.i.d. property in Condition 2. Generally speaking, asymptotic normality holds for other than i.i.d. property. These conditions are general and practical, especially for the discrete distributions used in source coding.

C. Essential Lemma for Analysis (Asymptotic Normality of Posterior Density)

Before analyzing of the code lengths, we state the asymptotic normality of the posterior distribution. Rissanen discussed the code length of the maximum-likelihood code on the asymptotic normality of the maximum-likelihood estimator [30]. See also [6]. The key to the analysis in this paper is the asymptotic normality of the posterior distribution. We can prove the almost sure convergence of the posterior density under Condition 2 from similar discussion in [2, Propositions 5.13 and 5.14, pp. 285–297]. Then, we have the following important lemma.

Lemma 2 (Asymptotic Normality): Under Condition 2, the Bayesian posterior densities of the parameter satisfy asymptotic normality in almost sure. That is, the posterior distribution of $\xi^k = \sqrt{n}(\theta^k - \tilde{\theta}^k)$ converges almost surely to a normal distribution with mean zero and covariance matrix $\{I^*(\tilde{\theta}^k)\}^{-1}$. In particular, if R^k_ξ is arbitrary k-dimensional rectangle, its probability mass induced by $f(\theta^k|x^n)$ satisfies

$$\begin{aligned} P(\xi^k \in R^k_\xi|x^n) &= \int_{\xi^k \in R^k_\xi} f_\xi(\xi^k|x^n)\, d\xi^k \\ &\to \frac{\sqrt{\det I^*(\tilde{\theta}^k)}}{(2\pi)^{k/2}} \int_{\xi^k \in R^k_\xi} e^{-\frac{1}{2}\|\xi^k\|^2_{I^*(\tilde{\theta}^k)}}\, d\xi^k \quad \text{a.s.} \end{aligned} \tag{89}$$

where $f_\xi(\xi^k|x^n)$ is the posterior density of ξ^k which is given by

$$f_\xi(\xi^k|x^n) = \frac{1}{\sqrt{n}^k} f(\theta^k|x^n). \tag{90}$$

Moreover, the posterior density $f_\xi(\xi^k|x^n)$ satisfies

$$f_\xi(\xi^k|x^n) \to \frac{\sqrt{\det I^*(\tilde{\theta}^k)}}{(2\pi)^{k/2}} e^{-\frac{1}{2}\|\xi^k\|^2_{I^*(\tilde{\theta}^k)}} \quad \text{a.s.} \tag{91}$$

uniformly for $\xi^k \in \Omega^k$, where Ω^k is arbitrary k-dimensional rectangle satisfying $|\Omega^k| < \infty$.

Proof: See Appendix A. □

The asymptotic normality is essential for the proof of Theorem 5.

D. Main Results for the Parametric Model Class

For the difference of code lengths, we show the following result first.

Lemma 3: We define

$$\overline{\xi}^k = \sqrt{n}(\overline{\theta}^k - \tilde{\theta}^k) \quad \text{and} \quad \overline{\Xi}^k{}_n = \{\overline{\xi}^k | \overline{\theta}^k \in \overline{\Theta}^k{}_n\}.$$

Then the relation between (6) and (7) is given by

$$L^{\overline{\theta}^k}_{\mathrm{MDL}}(x^n) = L^{\theta^k}_{\mathrm{Bayes}}(x^n) - \max_{\overline{\xi}^k \in \overline{\Xi}^k{}_n} \left\{ \log \frac{f_\xi(\overline{\xi}^k|x^n)}{\sqrt{\det I\left(\tilde{\theta}^k + \frac{\overline{\xi}^k}{\sqrt{n}}\right)}} \right\}. \quad (92)$$

Proof: From (6) and (7), we have

$$L^{\overline{\theta}^k}_{\mathrm{MDL}}(x^n) = L^{\theta^k}_{\mathrm{Bayes}}(x^n) - \max_{\overline{\theta}^k \in \overline{\Theta}^k{}_n} \left\{ \log \frac{P(x^n|\overline{\theta}^k) f(\overline{\theta}^k)}{\int_{\theta^k} P(x^n|\theta^k) f(\theta^k)\, d\theta^k} \cdot \frac{1}{\sqrt{n}^k \sqrt{\det I(\overline{\theta}^k)}} \right\}. \quad (93)$$

In the above equation

$$f(\overline{\theta}^k|x^n) = \frac{P(x^n|\overline{\theta}^k) f(\overline{\theta}^k)}{\int_{\theta^k} P(x^n|\theta^k) f(\theta^k)\, d\theta^k} \quad (94)$$

is the posterior density at the point $\theta^k = \overline{\theta}^k$. From $\overline{\theta}^k = \tilde{\theta}^k + \frac{1}{\sqrt{n}}\overline{\xi}^k$ and (90), the proof is completed. □

Next, under Condition 2, we obtain the asymptotic difference between $L^{\overline{\theta}^k}_{\mathrm{MDL}}(x^n)$ and $L^{\theta^k}_{\mathrm{Bayes}}(x^n)$ using Lemma 2 and Lemma 3.

Theorem 5: Under Condition 2, the relation between (6) and (7) is asymptotically given by

$$L^{\overline{\theta}^k}_{\mathrm{MDL}}(x^n) = L^{\theta^k}_{\mathrm{Bayes}}(x^n) + O^+(1) \quad \text{a.s.} \quad (95)$$

where $O^+(1)$ is the term such that $0 < C_1 < O^+(1) < C_2 < \infty$ for sufficient large n. That is, there exist $C_1 > 0$ and $C_2 > 0$ such that

$$C_1 < L^{\overline{\theta}^k}_{\mathrm{MDL}}(x^n) - L^{\theta^k}_{\mathrm{Bayes}}(x^n) < C_2 \quad \text{a.s.} \quad (96)$$

when $n \to \infty$.

Proof: See Appendix B. □

We can interpret Theorem 5 as follows: The second term on the RHS in (92) can be interpreted as the posterior probability of the cell of the quantized parameter. The posterior distribution of the parameter is asymptotically normal whose variance–covariance matrix is $(1/n)\{I^*(\tilde{\theta}^k)\}$. On the other hand, the quantized width of the parameter is also proportional to the standard deviation of the posterior probability density toward the quantizing axis. The more the standard deviation of the posterior density decreases as the sample size increases, the smaller the quantizing widths in relation to this standard deviation. Thus, this posterior probability of the quantization cell does not converge to 1, that is, the true quantization cell does not exist from the beginning although the true parameter exists. For this reason, the difference of the code lengths does not converge to 0.

V. Analysis for the Hierarchical Model Class

We analyze the difference of the code lengths between the MDL code and the Bayes code for the hierarchical model class. For the hierarchical model class, we denote

$$P(x^n|m) = \int_{\theta^{k_m}} P(x^n|m, \theta^{k_m}) f(\theta^{k_m}|m)\, d\theta^{k_m}. \quad (97)$$

The data sequence x^n is emitted from the true distribution $P^*(x^n)$ and we do not assume that $P^*(\cdot)$ exists in $\mathcal{H}$. Although the hierarchical model class defined in this paper may not have a nested structure, it is a trivial case and does not lead to a contradiction of the results.

The optimal parameter $\theta^{k^*_m}$ of a model m is defined by

$$\theta^{k^*_m} = \arg_{\theta^{k_m}} \min H^*(m, \theta^{k_m}), \quad (98)$$

where $H^*(m, \theta^{k_m})$ is given by

$$H^*(m, \theta^{k_m}) = \lim_{n\to\infty} \frac{1}{n} E^*\{-\log P(X^n|m, \theta^{k_m})\}. \quad (99)$$

We denote the maximum-likelihood estimator and the maximum posterior estimator of the model m given x^n by $\hat{\theta}^{k_m}$ and $\tilde{\theta}^{k_m}$, respectively. And we define the information matrix $I^*(\theta^{k_m}|m)$ as follows:

$$I^*(\theta^{k_m}|m) = \lim_{n\to\infty} E^* \left\{ -\frac{\partial^2 \log P(X^n|m, \theta^{k_m})}{\partial\theta^{k_m} (\partial\theta^{k_m})^T} \right\}. \quad (100)$$

$I^*(\theta^{k_m}|m) = I(\theta^{k_m}|m)$ for $\forall \theta^{k_m} \in \Theta^{k_m}$ is not generally satisfied. But, $I^*(\theta^{k^*_m}|m) = I(\theta^{k^*_m}|m)$ holds for most of the practical model classes for source coding.

Note that there may exist m_1, m_2, $\theta^{k_{m_1}}$, and $\theta^{k_{m_2}}$ where $m_1 \neq m_2$, satisfying $H^*(m_1, \theta^{k_{m_1}}) = H^*(m_2, \theta^{k_{m_2}})$. The optimal model m^* is defined as follows:

$$m^* = \arg_m \min \left\{ k_m \,\middle|\, H^*(m, \theta^{k^*_m}) = H^* \right\} \quad (101)$$

where H^* is given by

$$H^* = \min_{P(\cdot|m,\theta^{k_m}) \in \mathcal{H}} H^*(m, \theta^{k_m}). \quad (102)$$

Finally, we define the ball

$$B_\delta(\theta^{k^*_m}|m) = \left\{ \theta^{k_m} \in \Theta^{k_m} \,\middle|\, \|\theta^{k_m} - \theta^{k^*_m}\| < \delta \right\}.$$

We assume the following conditions for the model class and Bayesian inference. These conditions are stronger than those assumed in analysis for the parametric model class, that is, we assume the iterated logarithm law in this section.

Condition 4:

i) (Existence of $\theta^{k^*_m}$) For $m \in \mathcal{M}$, the function $-H^*(m, \theta^{k_m})$ is a unimodal function on Θ^{k_m} with maximum point in the interior of Θ^{k_m}. That is, the optimal parameter of model m, $\theta^{k^*_m}$, exists uniquely in the interior of Θ^{k_m} for $\forall m \in \mathcal{M}$.

ii) (Smoothness of class) For $m \in \mathcal{M}$, the Fisher information matrix $I(\theta^{k_m}|m)$ satisfies

$$0 < C_0 < \det I(\theta^{k_m}|m) < \infty, \qquad \text{for } \theta^{k_m} \in \Theta^{k_m}$$

where C_0 is some positive constant. When $\theta^{k_m} = \theta^{k_m^*}$, $I(\theta^{k_m^*}|m) = I^*(\theta^{k_m^*}|m)$. Moreover, $\det I(\theta^{k_m}|m)$ and $\det I^*(\theta^{k_m}|m)$ satisfy

$$\left\| \frac{\partial \det I(\theta^{k_m}|m)}{\partial \theta^{k_m}} \right\| < \infty \tag{103}$$

$$\left\| \frac{\partial \det I^*(\theta^{k_m}|m)}{\partial \theta^{k_m}} \right\| < \infty \tag{104}$$

for $\theta^{k_m} \in \Theta^{k_m}$.

iii) (Smoothness of prior) For $m \in \mathcal{M}$ and $\forall\, \theta^{k_m} \in \Theta^{k_m}$, $f(\theta^{k_m}|m) > 0$ and $f(\theta^{k_m}|m)$ is three times continuously differentiable for θ^{k_m}. That is, we have

$$\left| \sum_{i=0}^{k_m-1} \sum_{j=0}^{k_m-1} \sum_{l=0}^{k_m-1} \frac{\partial f(\theta^{k_m}|m)}{\partial \theta_i^{k_m} \partial \theta_j^{k_m} \partial \theta_l^{k_m}} \right| < c_f(\theta^{k_m}) \tag{105}$$

where $c_f(\theta^{k_m})$ is some finite constant depending only on θ^{k_m}.

iv) (Existence of estimators) The likelihood function $P(x^n|m, \theta^{k_m})$ with respect to θ^{k_m} given x^n is unimodal or monotonic on Θ^{k_m} for $x^n \in \mathcal{X}^n$. The posterior density function $f(\theta^{k_m}|x^n, m)$ has a unique maximum in Θ^{k_m}. That is, both the maximum-likelihood estimator $\hat{\theta}^{k_m}$ and the maximum posterior estimator $\tilde{\theta}^{k_m}$ uniquely exist for $\forall\, x^n \in \mathcal{X}^n$ and $\forall\, m \in \mathcal{M}$.

v) (Iterated logarithm law: I) For $m \in \mathcal{M}$

$$\tilde{\theta}^{k_m} = \theta^{k_m^*} + O\left(\frac{(\log\log n)^{1/2}}{\sqrt{n}} \right) \quad \text{a.s.} \tag{106}$$

vi) (Iterated logarithm law: II) There exists $\delta > 0$ for $m \in \mathcal{M}$ so that

$$\frac{1}{n} \log P(x^n|m, \theta^{k_m}) = H^*(m, \theta^{k_m}) + O\left(\frac{(\log\log n)^{1/2}}{\sqrt{n}} \right) \quad \text{a.s.} \tag{107}$$

$$\frac{1}{n} \frac{\partial \log P(x^n|m, \theta^{k_m})}{\partial \theta^{k_m}} = \frac{\partial H^*(m, \theta^{k_m})}{\partial \theta^{k_m}} + O\left(\frac{(\log\log n)^{1/2}}{\sqrt{n}} \right) \quad \text{a.s.} \tag{108}$$

$$-\frac{1}{n} \frac{\partial^2 \log P(x^n|m, \theta^{k_m})}{\partial \theta^{k_m} (\partial \theta^{k_m})^T} = I^*(\theta^{k_m}|m) + O\left(\frac{(\log\log n)^{1/2}}{\sqrt{n}} \right) \quad \text{a.s.} \tag{109}$$

uniformly for $\theta^{k_m} \in B_\delta(\theta^{k_m^*}|m)$. □

From Condition 4, ii), the distribution of x_t conditioned by x^{t-1} satisfies the following inequality for $\forall\, \theta^{k_m} \in \Theta^{k_m}$ and $\forall\, m \in \mathcal{M}$:

$$\forall\, t \geq 1, \quad \left| \sum_{i=1}^{k_m} \sum_{j=1}^{k_m} \sum_{l=1}^{k_m} \frac{\partial \log P(x_t|x^{t-1}, m, \theta^{k_m})}{\partial \theta_i^{k_m} \partial \theta_j^{k_m} \partial \theta_l^{k_m}} \right| < c_p'(\theta^{k_m}). \tag{110}$$

Here, $c_p'(\theta^{k_m}) > 0$ is a finite constant depending only on θ^{k_m} and $P(x_1|x^0, m, \theta^{k_m}) = P(x_1|m, \theta^{k_m})$.

Example 6 (Finite Ergodic Markov Source): Consider a finite ergodic Markov source on $\mathcal{X} = \{0, 1, 2, \ldots, \beta\}$. Let the set of states of the Markov model be previously unknown. That is, we do not know the length of memory.

Let m_1 be a model specifying a set of states

$$\{s_0(m_1), s_1(m_1), \ldots, s_{S_{m_1}}(m_1)\}$$

where S_{m_1} is the number of states of the model m_1. Similarly, we can construct $m_2, m_3, \ldots$. For example, we consider the binary alphabet $\mathcal{X} = \{0, 1\}$. Let m_1 be the simple Markov model with two states. Let m_2 be the second-order Markov model with four states. Similarly, we define $m_3, m_4, \ldots$, as the third-order Markov model, the fourth-order Markov model, $\ldots$, respectively. Then we construct $\mathcal{M}$ by $\{m_1, m_2, \ldots, m_M\}$, where M is the number of models and m_M is the Mth-order Markov model.

Each model m has a parameter $\theta_{i,j}^{k_m}$ with elements $\theta_{i,j}^{k_m} = p_i^{(s_j(m))}$. Here, $p_i^{(s_j(m))}$ is the probability of the symbol i at the jth state $s_j(m)$ of a model m, and $q_{s_j}(m)$, the stationary probability of the states $s_j(m)$, where $i \in \mathcal{X}$ and $j = 0, 1, \ldots, S_m$ and S_m is the number of states of the model m. We may regard $\theta^{k_m} = (\theta_{0,0}^{k_m}, \ldots, \theta_{\beta-1, S_m}^{k_m})^T$ as a continuous parameter, where $k_m = \beta(S_m + 1)$.

We assume the initial state is known. Let $N_{i,j}(m)$ be random variables representing appearance numbers of each symbols $0, 1, \ldots, \beta$ conditional on the states $s_j(m)$, $j = 0, \ldots, S_m$, of the random variable X^n. We define

$$p_{i,j}^*(m) = \lim_{n\to\infty} \tfrac{1}{n} E^* N_{i,j}(m).$$

Then, the optimal parameter $\theta^{k_m^*}$ is given by

$$\theta^{k_m^*} = (p_{0,0}^*(m), \ldots, p_{\beta-1, S_m}^*(m))^T.$$

The stationary probabilities on the states of the model m calculated by $p_{i,j}^*(m)$ are denoted by $q_{s_0(m)}^*(m) \cdots q_{s_{S_m}(m)}^*(m)$.

From the iterated logarithm law [8], [10], we have

$$\frac{n_{i,j}(m)}{n} = p_{i,j}^*(m) + O\left(\frac{(\log\log n)^{1/2}}{\sqrt{n}} \right) \tag{111}$$

for $m \in \mathcal{M}$. Using this fact, we can show that this model class satisfies Condition 4 for $\forall\, m \in \mathcal{M}$ by a discussion similar to that of Example 5. □

Theorem 6: For the same prior probability, the relation between (12) and (14) is

$$L_{\mathrm{MDL}}^{m, \theta^{k_m}}(x^n) = L_{\mathrm{Bayes}}^{m, \theta^{k_m}}(x^n) - \log P(\tilde{m}|x^n). \tag{112}$$

Here, $\tilde{m}$ represents the model which maximizes the posterior probability $P(m|x^n)$, which is given by

$$P(m|x^n) = \frac{\int_{\theta^{k_m}} P(x^n|m, \theta^{k_m}) f(\theta^{k_m}|m) P(m)\, d\theta^{k_m}}{\sum_m \int_{\theta^{k_m}} P(x^n|m, \theta^{k_m}) f(\theta^{k_m}|m) P(m)\, d\theta^{k_m}}. \tag{113}$$

Proof: This can be proved by a discussion similar to that of Theorem 1. □

This theorem shows that the code length of the Bayes code is smaller than that of the MDL code by $-\log P(\tilde{m}|x^n)$ using the same prior, and the Bayes code is effective on finite data sequences. Next we consider the order of the term $-\log P(\tilde{m}|x^n)$.

The following lemma showing the order of $-\log P(\tilde{m}|x^n)$ is very important for the analysis in this section. From this lemma, we can analyze the difference of the code lengths by regarding

$$\left\{P(x^n|m)=\int P(x^n|m,\theta^{k_m})f(\theta^{k_m}|m)\,d\theta^{k_m}|m\in\mathcal{M}\right\}$$

as a discrete model class.

Lemma 4: Under Condition 4, we have[8]

$$P(m^*|x^n)\to 1 \quad \text{a.s.} \tag{114}$$

Proof: See Appendix C. □

Then, we have the following theorem.

Theorem 7: Under Condition 4, the relation between (12) and (14) is given by

$$L_{\mathrm{MDL}}^{m,\theta^{k_m}}(x^n)=L_{\mathrm{Bayes}}^{m,\theta^{k_m}}(x^n)+o^+(1) \quad \text{a.s.} \tag{115}$$

where $o^+(1)$ is positive and $\lim_{n\to\infty} o^+(1)=1$.

Proof: From Lemma 4, the model $\tilde{m}$ which maximizes the posterior probability $P(m|x^n)$ asymptotically corresponds to the optimal model m^*. Then

$$-\log P(\tilde{m}|x^n)=o^+(1). \tag{116}$$

From this equation and Theorem 6, the theorem is proved. □

This theorem shows that the difference of the code lengths between the code using mixture of all models and the code based on model selection vanishes. Next, we consider the difference of the code lengths between (10) and (14).

Theorem 8: Under Condition 4, the relation between (10) and (14) is given by

$$L_{\mathrm{MDL}}^{m,\overline{\theta}^{k_m}}(x^n)=L_{\mathrm{Bayes}}^{m,\theta^{k_m}}(x^n)+O^+(1) \quad \text{a.s.} \tag{117}$$

where $O^+(1)$ is the term such as $0<C_1<O^+(1)<C_2<\infty$.

Proof: Define $L_{\mathrm{MDL}}^{\overline{\theta}^{k_m}}(x^n|m)$ as

$$L_{\mathrm{MDL}}^{\overline{\theta}^{k_m}}(x^n|m)=\min_{\overline{\theta}^{k_m}}\left\{-\log P(x^n|m,\overline{\theta}^{k_m})-\log\frac{f(\overline{\theta}^{k_m}|m)}{\sqrt{n}^{k_m}\sqrt{\det I(\overline{\theta}^{k_m}|m)}}\right\}. \tag{118}$$

[8]This result shows the strong consistency of the model selection by maximization of the posterior probability, whose asymptotic formula, the Bayesian information criterion (BIC), was proposed by Schwarz [31].

From Lemma 3, under Condition 4, we have

$$L_{\mathrm{MDL}}^{\overline{\theta}^{k_m}}(x^n|m)=-\log P(x^n|m)-\max_{\overline{\xi}^k}\left\{\log\frac{f_\xi(\overline{\xi}^k|x^n,m)}{\sqrt{\det I\left(\hat{\theta}^k+\frac{\overline{\xi}^k}{\sqrt{n}}|m\right)}}\right\} \tag{119}$$

for $m\in\mathcal{M}$ where $f_\xi(\xi^k|x^n,m)$ is the posterior density of $\xi^k=\sqrt{n}(\theta^k-\hat{\theta}^k)$ given by

$$f_\xi(\xi^k|x^n,m)=\frac{1}{\sqrt{n}^k}f(\theta^k|x^n,m). \tag{120}$$

From Theorem 5, the following asymptotic in equation is satisfied:

$$0<C_1<-\max_{\overline{\xi}^k}\left\{\log\frac{f_\xi(\overline{\xi}^k|x^n,m)}{\sqrt{\det I\left(\hat{\theta}^k+\frac{\overline{\xi}^k}{\sqrt{n}}|m\right)}}\right\}<C_2<\infty \quad \text{a.s.} \tag{121}$$

when $n\to\infty$, where C_1 and C_2 are positive constants.

From (119) and (121), we have

$$\begin{aligned}L_{\mathrm{MDL}}^{m,\overline{\theta}^{k_m}}(x^n)&=\min_{m\in\mathcal{M}}\left\{L_{\mathrm{MDL}}^{\overline{\theta}^{k_m}}(x^n|m)-\log P(m)\right\}\\&=L_{\mathrm{MDL}}^{m,\theta^{k_m}}(x^n)+O^+(1) \quad \text{a.s.}\end{aligned} \tag{122}$$

From (122) and Theorem 7, the proof is completed. □

For the hierarchical model class, it is clear that parameter quantization is not effective for source coding. In the above theorem, the difference of both code lengths is given by constant order.

VI. Discussion

Although it had been shown that the Bayes code is more effective than the MDL code from the viewpoint of code length with the same prior distribution for finite value of n [29], we analyzed the difference quantitatively in this paper.

Since the difference of both code lengths is not larger than $O(1)$, the difference of the mean code lengths per symbol (compression rate) is not larger than $O(1/n)$ and converges to 0. The nonpredictive MDL principle which is discussed in this paper has two points of operation, i.e., the operation of parameter quantization and that of selection of a model or a representative point of the quantized parameter, where the former has a stronger influence on the difference of the code lengths than the latter.

If the prior distribution of the Bayes code may be different from that of the MDL code, then it is possible to find a case in which the code length of the MDL code is smaller than that of the Bayes code for some information source. The coding which has larger prior probability for the optimal model is effective. However, we cannot practically establish the prior distribution with a large probability for the optimal model when the optimal probability model is unknown. Therefore, the results which were discussed under the condition that the same prior

was assumed are practical and important. In practice, since we can recognize the prior distribution of the MDL code beforehand, it is surely possible to construct the Bayes' code who's code length is equal or smaller than that of the MDL code.

From the above, the selection of a probabilistic model is not always effective for all various purposes. We have reconfirmed that the effectiveness of the MDL principle occurs within the framework of statistical model selection or universal modeling [30].

When the MDL principle is applied to the model selection problem, $L_{\mathrm{MDL}}^{m,\hat{\theta}^{km}}(x^n)$ should be applied rather than $L_{\mathrm{MDL}}^{m,\overline{\theta}^{km}}(x^n)$. This is because

$$L_{\mathrm{MDL}}^{m,\theta^{km}}(x^n) \le L_{\mathrm{MDL}}^{m,\overline{\theta}^{km}}(x^n), \qquad \text{for } \forall\, x^n \in \mathcal{X}^n$$

that is, $L_{\mathrm{MDL}}^{m,\overline{\theta}^{km}}(x^n)$ are not true MDL. On the other hand, $-\log P(x^n|m)$ is Bayes optimum code length when m is fixed. Minimization of $-\log P(x^n|m) - \log P(m)$ is equivalent to maximization of the posterior probability $P(m|x^n)$. From Lemma 4, this criterion for model selection has strong consistency for the hierarchical model class.

VII. Conclusion

In this paper, we have analyzed the difference of the code lengths between the MDL code and Bayes code. From the results, the effectiveness of the Bayes code against the MDL code with parameter quantization has been shown from the stand point of code length. Future work includes discussion of the properties of other types of MDL codes [30], or the relation between the MDL criterion, the Bayesian model selection, and conventional information criteria in model selection [1], [15], [19], [22], [31], [32] from new viewpoints.

Appendix A
The Proof of Lemma 2

For the asymptotic normality of the posterior distribution, the following necessary and sufficient condition shown in [2, pp. 285–297] and [7] is useful.

Lemma 5 [2], [7]: Fix a sequence x^∞. Let $\tilde{\theta}^k$ be a strict local maximum of $L_n(\theta^k) = \log f(\theta^k|x^n)$ satisfying

$$L_n'(\tilde{\theta}^k) = \left.\frac{\partial L_n(\theta^k)}{\partial \theta^k}\right|_{\theta^k=\tilde{\theta}^k} = \mathbf{0} \tag{123}$$

and implying positive definiteness of

$$\Sigma_n = -\left(L_n''(\tilde{\theta}^k)\right)^{-1} = -\left(\left.\frac{\partial^2 L_n(\theta^k)}{\partial\theta^k(\partial\theta^k)^T}\right|_{\theta^k=\tilde{\theta}^k}\right)^{-1}. \tag{124}$$

Defining

$$B_\delta(\tilde{\theta}^k) = \{\theta^k \in \Theta^k \,|\, \|\theta^k - \tilde{\theta}^k\| < \delta\}$$

the following three basic conditions are necessary and sufficient for the asymptotic normality of the posterior distribution.

c.1) "Steepness:" $\lim_{n\to\infty} \overline{\sigma}_n^2 \to 0$, where $\overline{\sigma}_n^2$ is the largest eigenvalue of Σ_n.

c.2) "Smoothness:" For any $\epsilon > 0$, there exists N and $\delta > 0$ such that, for any $n > N$ and $\theta^k \in B_\delta(\tilde{\theta}^k)$, $L_n''(\theta^k)$ exists and satisfies

$$I - A(\epsilon) \le L_n''(\theta^k)\left\{L_n''(\tilde{\theta}^k)\right\}^{-1} \le I + A(\epsilon) \tag{125}$$

where I is the $k\times k$ identity matrix and $A(\epsilon)$ is a $k\times k$ symmetric positive–semidefinite matrix whose largest eigenvalue tends to 0 as $\epsilon \to 0$.

c.3) "Concentration:" For $\forall\,\delta > 0$, when $n \to \infty$

$$\int_{\theta^k \in B_\delta(\tilde{\theta}^k)} f(\theta^k|x^n)\, d\theta^k \to 1. \tag{126}$$

Conditions c.1) and c.2) imply that

$$\lim_{n\to\infty} f(\tilde{\theta}^k|x^n)\,(\det\Sigma_n)^{1/2} \le (2\pi)^{-k/2} \tag{127}$$

with equality if and only if c.3) holds.

Moreover, given c.1), c.2), and c.3), $\phi^k = \Sigma_n^{-1/2}(\theta^k - \tilde{\theta}^k)$ converges in distribution to k-dimensional standard normal distribution

$$f(\phi^k) = (2\pi)^{-k/2}\exp\left\{-\frac{1}{2}(\phi^k)^T\phi^k\right\}$$

where $\Sigma_n^{-1/2}$ is $k\times k$-matrix satisfying $\Sigma_n^{-1/2}\Sigma_n^{-1/2} = \Sigma_n^{-1}$. □

At first, we show that c.1)–c.3) of Lemma 5 hold almost surely under Condition 2 and the posterior distribution converges almost surely to the normal distribution from Lemma 5.

Since $f(\theta^k)$ is three times continuously differentiable from Condition 2, iii), we have

$$\frac{1}{n}L_n''(\theta^k) = \frac{1}{n}\frac{\partial^2 \log P(x^n|\theta^k)}{\partial\theta^k(\partial\theta^k)^T} + \frac{1}{n}\frac{\partial^2 \log f(\theta^k)}{\partial\theta^k(\partial\theta^k)^T}. \tag{128}$$

Since $\frac{\partial^2 \log f(\theta^k)}{\partial\theta^k(\partial\theta^k)^T}$ does not depend on n and Condition 2, vi), we have

$$-\frac{1}{n}L_n''(\theta^k) \to I^*(\theta^k) \quad \text{a.s.} \tag{129}$$

uniformly for $\forall\,\theta^k \in B_\delta(\theta^{k*})$. We have, therefore,

$$-\frac{1}{n}\,(\Sigma_n)^{-1} \to I^*(\tilde{\theta}^k) \quad \text{a.s.} \tag{130}$$

where Σ_n is given by

$$\Sigma_n = -\left(L_n''(\tilde{\theta}^k)\right)^{-1} = -\left(\left.\frac{\partial^2 L_n(\theta^k)}{\partial\theta^k(\partial\theta^k)^T}\right|_{\theta^k=\tilde{\theta}^k}\right)^{-1}. \tag{131}$$

From $\tilde{\theta}^k \to \theta^{k*}$, a.s. and Condition 2, ii), there exist the positive constants $C_1 > 0$ and $C_2 > 0$ such as

$$C_1 < \det I^*(\tilde{\theta}^k) < C_2 \quad \text{a.s.}$$

when $n \to \infty$. Then, since $\Sigma_n \to -\frac{1}{n}(I^*(\tilde{\theta}^k))^{-1}$, a.s., the largest eigenvalue of Σ_n tends to 0 almost surely, and c.1) is almost surely satisfied.

Since $\tilde{\theta}^k \to \theta^{k*}$, $L_n''(\theta^k)\{L_n''(\tilde{\theta}^k)\}^{-1}$ satisfies

$$L_n''(\theta^k)\{L_n''(\tilde{\theta}^k)\}^{-1} \to I^*(\theta^k)\{I^*(\tilde{\theta}^k)\}^{-1} \quad \text{a.s.} \tag{132}$$

uniformly for $\forall\,\theta^k \in B_\delta(\tilde{\theta}^k)$ form (129). From Condition 2, ii), c.2) is almost surely satisfied.

Therefore, from

$$\xi^k = \sqrt{n}(\theta^k - \tilde{\theta}^k) \quad \text{and} \quad f_\xi(\xi^k|x^n) = \frac{f(\theta^k|x^n)}{\sqrt{n}^k}$$

we have

$$\lim_{n\to\infty} f_\xi(\tilde{\xi}^k|x^n) \le \frac{\sqrt{\det I^*(\tilde{\theta}^k)}}{(2\pi)^{k/2}} \quad \text{a.s.} \tag{133}$$

from Lemma 5 and (130). Here $\tilde{\xi}^k = \mathbf{0}$. If

$$\int_{\theta^k \in B_\delta(\tilde{\theta}^k)} f(\theta^k|x^n)\, d\theta^k \to 1 \quad \text{a.s.} \tag{134}$$

holds, then we have

$$\lim_{n\to\infty} f(\tilde{\theta}^k|x^n)\,(\det \Sigma_n)^{1/2} = (2\pi)^{-k/2} \quad \text{a.s.} \tag{135}$$

from Lemma 5. Then we show (134) at last.

We have

$$\frac{1}{n}\log\frac{f(\theta^k|x^n)}{f(\theta^{k*}|x^n)} = \frac{1}{n}\log\frac{P(x^n|\theta^k)}{P(x^n|\theta^{k*})} + \frac{1}{n}\log\frac{f(\theta^k)}{f(\theta^{k*})} \to -H^*(\theta^k) + H^*(\theta^{k*}) \quad \text{a.s.} \tag{136}$$

as $n\to\infty$ uniformly for $\forall\,\theta^k \in B_\delta(\theta^{k*})$ for $\forall\,\delta > 0$ from Condition 2, iii) and vi). From Condition 2, i), $H^*(\theta^{k*}) \le H^*(\theta^k)$ for $\forall\,\theta^k \in \Theta^k$. On the other hand, since $\tilde{\theta}^k \to \theta^{k*}$, almost surely , $P(x^n|\theta^k)$ given x^n is a unimodal function with respect to θ^k almost surely when $n\to\infty$ from Condition 2, i) and iv). This is because $P(x^n|\theta^k)$ given x^n is a unimodal or monotone function with respect to θ^k and $\tilde{\theta}^k$ exists in the interior of Θ^k almost surely when $n\to\infty$. From (136) and unimodality of $P(x^n|\theta^k)$, there exists $C_\delta > 0$ such that

$$\frac{1}{n}\log\frac{f(\theta^k|x^n)}{f(\theta^{k*}|x^n)} < -C_\delta \quad \text{a.s.} \tag{137}$$

uniformly for $\forall\,\theta^k \notin B_\delta(\theta^{k*})$ when $n\to\infty$.

Therefore,

$$\frac{f(\theta^k|x^n)}{f(\theta^{k*}|x^n)} < \exp\{-nC_\delta\}, \ \text{a.s.} \tag{138}$$

holds uniformly for $\forall\,\theta^k \notin B_\delta(\theta^{k*})$ when $n\to\infty$.

On the other hand, from (133) and $f(\theta^k|x^n) = \sqrt{n}^k f_\xi(\xi^k|x^n)$, there exists $C^* > 0$ such that

$$f(\theta^{k*}|x^n) < C^*\sqrt{n}^k \quad \text{a.s.} \tag{139}$$

holds when $n\to\infty$.

From (138) and (139), we have

$$f(\theta^k|x^n) < C^*\sqrt{n}^k \exp\{-nC_\delta\} \to 0 \quad \text{a.s.} \tag{140}$$

holds uniformly for $\forall\,\theta^k \notin B_\delta(\theta^{k*})$ when $n\to\infty$. We have, therefore,

$$\int_{\theta^k \notin B_\delta(\theta^{k*})} f(\theta^k|x^n)\, d\theta^k \to 0 \quad \text{a.s.} \tag{141}$$

because $|\Theta^k| < \infty$. This means

$$\int_{\theta^k \in B_\delta(\theta^{k*})} f(\theta^k|x^n)\, d\theta^k \to 1 \quad \text{a.s.} \tag{142}$$

From $\tilde{\theta}^k \to \theta^{k*}$, a.s., we have $B_{\delta'}(\theta^{k*}) \subset B_{\delta''}(\tilde{\theta}^k)$, a.s., when $n\to\infty$ for $0 < \forall\,\delta' < \forall\,\delta''$. This means that c.3) is almost surely satisfied.

Since c.1)–c.3) are almost surely satisfied, we have

$$\lim_{n\to\infty} \frac{f(\tilde{\theta}^k|x^n)}{\sqrt{n}^k} = \frac{\sqrt{\det I^*(\tilde{\theta}^k)}}{(2\pi)^{k/2}} \quad \text{a.s.} \tag{143}$$

from Lemma 5. And the posterior distribution of $\xi^k = \sqrt{n}(\theta^k - \tilde{\theta}^k)$ converges almost surely to the normal distribution with mean zero and covariance matrix $\{I^*(\tilde{\theta}^k)\}^{-1}$. The first half of the theorem is proved.

Next, we show the uniform convergence of the posterior density for $\forall\,\xi^k \in \Omega^k$. From Taylor expansion, we have

$$\begin{aligned} f(\theta^k|x^n) &= f(\tilde{\theta}^k|x^n)\exp\{L_n(\theta^k) - L_n(\tilde{\theta}^k)\} \\ &= f(\tilde{\theta}^k|x^n)\exp\left\{-\tfrac{1}{2}(\theta^k - \tilde{\theta}^k)^T(I + R_n)\Sigma_n^{-1}(\theta^k - \tilde{\theta}^k)\right\} \end{aligned} \tag{144}$$

where R_n is given by

$$R_n = L_n''(\theta^{k+})\left\{\mathrm{L}_n''(\tilde{\theta}^k)\right\}^{-1} - I \tag{145}$$

for some θ^{k+} lying between θ^k and $\tilde{\theta}^k$.

We have, therefore,

$$\frac{f(\theta^k|x^n)}{f(\tilde{\theta}^k|x^n)} = \exp\left\{-\frac{1}{2}(\xi^k)^T(I + R_n)(\Sigma_n)^{-1}(\xi^k)\right\}. \tag{146}$$

For any rectangle Ω^k, there exists $V_\Omega > 0$ such as

$$\sup_{\xi^k\in\Omega^k} \|\xi^k\| < V_\Omega.$$

Therefore, we have $\theta^k = \tilde{\theta}^k + \frac{\xi^k}{\sqrt{n}} \to \tilde{\theta}^k$ uniformly for $\forall\,\xi^k \in \Omega^k$. This implys

$$(1-\epsilon)I \le I + R_n \le (1+\epsilon)I \quad \text{a.s.} \tag{147}$$

when $n\to\infty$ uniformly for $\forall\,\xi^k \in \Omega^k$ for $\forall\,\epsilon > 0$, since

$$L_n''(\theta^{k+})\{L_n''(\tilde{\theta}^k)\}^{-1} \to I^*(\theta^{k+})\{I^*(\tilde{\theta}^k)\}^{-1} \quad \text{a.s.}$$

and $\|\theta^k - \tilde{\theta}^k\| \to 0$, a.s., uniformly for $\forall\,\xi^k \in \Omega^k$.

From (146) and (147), we have

$$\frac{f(\theta^k|x^n)}{f(\tilde{\theta}^k|x^n)} = \exp\left\{-\tfrac{1}{2}(\xi^k)^T I^*(\tilde{\theta}^k)(\xi^k)\right\}\{1 + o(1)\} \quad \text{a.s.} \tag{148}$$

uniformly for $\forall\,\xi^k \in \Omega^k$. Since $\exp\{-\frac{1}{2}(\xi^k)^T I^*(\tilde{\theta}^k)(\xi^k)\}$ is upper-bounded for $\forall\,\xi^k \in \Omega^k$, we have

$$\frac{f(\theta^k|x^n)}{f(\tilde{\theta}^k|x^n)} = \exp\left\{-\tfrac{1}{2}(\xi^k)^T I^*(\tilde{\theta}^k)(\xi^k)\right\} + o(1) \quad \text{a.s.} \tag{149}$$

uniformly for $\forall\,\xi^k \in \Omega^k$. From (143), the proof is completed.□

Appendix B
The Proof of Theorem 5

From Lemma 2

$$f_\xi(\xi^k|x^n) \to \frac{\sqrt{\det I^*(\tilde{\theta}^k)}}{(2\pi)^{k/2}}\, e^{-\frac{\|\xi^k\|^2_{I^*(\tilde{\theta}^k)}}{2}} \quad \text{a.s.} \tag{150}$$

uniformly for $\forall\,\xi^k \in \Omega^k$. On the other hand, $\tilde{\theta}^k$ gives the strict maximum of $f(\theta^k|x^n)$. Then, from (150), the posterior density $f_\xi(\xi^k|x^n)$ satisfies

$$\max_{\overline{\xi}^k\in\overline{\Xi}^k} f_\xi(\overline{\xi}^k|x^n) \le \frac{\sqrt{\det I^*(\tilde{\theta}^k)}}{(2\pi)^{k/2}} + \epsilon_1 \quad \text{a.s.} \tag{151}$$

when $n \to \infty$ for $\forall \epsilon_1 > 0$, where $\overline{\xi}^k$ is given by

$$\overline{\xi}^k = \sqrt{n}\left(\overline{\theta}^k - \tilde{\theta}^k\right). \tag{152}$$

Let Ω^k be a rectangle whose volume is sufficiently large such that $\overline{\xi}^k \in \Omega^k$. Since

$$0 < \sqrt{C_0} < \sqrt{\det I\left(\tilde{\theta}^k + \frac{1}{\sqrt{n}}\overline{\xi}^k\right)}$$

from Condition 2, ii), and $f_\xi(\xi^k|x^n)$ is almost surely unimodal when $n \to \infty$ from Condition 2, iv) and v), we have

$$\begin{aligned}
&\max_{\overline{\xi}^k \in \overline{\Xi}^k} \frac{f_\xi(\overline{\xi}^k|x^n)}{\sqrt{\det I\left(\tilde{\theta}^k + \frac{1}{\sqrt{n}}\overline{\xi}^k\right)}} \\
&\quad = \max_{\overline{\xi}^k \in (\Omega^k \cap \overline{\Xi}^k)} \frac{f_\xi(\overline{\xi}^k|x^n)}{\sqrt{\det I\left(\tilde{\theta}^k + \frac{1}{\sqrt{n}}\overline{\xi}^k\right)}} \quad \text{a.s.} \\
&\quad \le \frac{\sqrt{\det I^*(\tilde{\theta}^k)}\Big/(2\pi)^{k/2}}{\min_{\overline{\xi}^k \in (\Omega^k \cap \overline{\Xi}^k)} \sqrt{\det I\left(\tilde{\theta}^k + \frac{1}{\sqrt{n}}\overline{\xi}^k\right)}} + \epsilon_2 \quad \text{a.s.}
\end{aligned} \tag{153}$$

when $n \to \infty$ for $\forall \epsilon_2 > 0$.

On the other hand, the derivatives of $\sqrt{\det I^*(\theta^k)}$ and $\sqrt{\det I(\theta^k)}$ with respect to θ^k exist from Condition 2, ii). Moreover, $\tilde{\theta}^k \to \hat{\theta}^k \to \theta^{k*}$, a.s. from Condition 2, v), and $I^*(\theta^{k*}) = I(\theta^{k*})$ from Condition 2, ii). Therefore, we have

$$\sqrt{\det I^*(\tilde{\theta}^k)} \to \sqrt{\det I(\theta^{k*})} \quad \text{a.s.} \tag{154}$$

$$\sqrt{\det I\left(\tilde{\theta}^k + \frac{1}{\sqrt{n}}\overline{\xi}^k\right)} \to \sqrt{\det I(\theta^{k*})} \quad \text{a.s.} \tag{155}$$

uniformly for $\xi^k \in \Omega^k$ when $n \to \infty$. Equation (155) is derived from $\|\frac{1}{\sqrt{n}}\xi^k\| \to 0$ uniformly for $\overline{\xi}^k \in \Omega^k$ because $|\Omega^k| < \infty$. From (153), (154), and (155)

$$\max_{\overline{\xi}^k \in \overline{\Xi}^k} \log \frac{f_\xi(\overline{\xi}^k|x^n)}{\sqrt{\det I\left(\tilde{\theta}^k + \frac{\overline{\xi}^k}{\sqrt{n}}\right)}} \le \frac{k}{2}\log\frac{1}{2\pi} + \epsilon_3 \quad \text{a.s.} \tag{156}$$

is satisfied when $n \to \infty$ for $\forall \epsilon_3 > 0$. Therefore, from Lemma 3, it has been proved that there exists $\exists C_1 > 0$ such as

$$C_1 < L^{\overline{\theta}^k}_{\mathrm{MDL}}(x^n) - L^{\theta^k}_{\mathrm{Bayes}}(x^n) \quad \text{a.s.} \tag{157}$$

when $n \to \infty$, where $0 < C_1 < \frac{k}{2}\log 2\pi$.

Next, we shall show

$$L^{\overline{\theta}^k}_{\mathrm{MDL}}(x^n) - L^{\theta^k}_{\mathrm{Bayes}}(x^n) < C_2 \quad \text{a.s.} \tag{158}$$

for sufficient large n. From Condition 2, ii), we have

$$\begin{aligned}
&\frac{\max_{\overline{\xi}^k \in (\Omega^k \cap \overline{\Xi}^k)} f_\xi(\overline{\xi}^k|x^n)}{\sqrt{\det I(\tilde{\theta}^k) + \frac{K_{\Omega^k}}{\sqrt{n}}}} \\
&\qquad \le \max_{\overline{\xi}^k \in (\Omega^k \cap \overline{\Xi}^k)} \frac{f_\xi(\overline{\xi}^k|x^n)}{\sqrt{\det I\left(\tilde{\theta}^k + \frac{\overline{\xi}^k}{\sqrt{n}}\right)}} \quad \text{a.s.}
\end{aligned} \tag{159}$$

when $n \to \infty$ for $\exists K_{\Omega^k} > 0$, where K_{Ω^k} depends on Ω^k.

From (155), there clearly exists $C' > 0$ such as

$$C' < \max_{\overline{\xi}^k \in (\Omega^k \cap \overline{\Xi}^k)} f_\xi(\overline{\xi}^k|x^n) \le \max_{\overline{\xi}^k \in \overline{\Xi}^k} f_\xi(\overline{\xi}^k|x^n) \quad \text{a.s.} \tag{160}$$

when $n \to \infty$.

From (155), (159), and (160) there exists $C_2 > 0$ such as

$$-\max_{\overline{\xi}^k \in \overline{\Xi}^k} \log \frac{f_\xi(\overline{\xi}^k|x^n)}{\sqrt{\det I\left(\tilde{\theta}^k + \frac{\overline{\xi}^k}{\sqrt{n}}\right)}} < C_2 \quad \text{a.s.} \tag{161}$$

is satisfied when $n \to \infty$, which leads to (158).[9]

Since both of (157) and (158) are proved, the proof is completed. □

Appendix C
The Proof of Lemma 4

At first, we show

$$\left|\frac{P(x^n|m)}{P(x^n|m^*)}\right| \to 0 \quad \text{a.s.} \tag{162}$$

for $\forall m \neq m^*$, $m \in \mathcal{M}$ under Condition 4.

From Lemma 2 and

$$\log P(x^n|m) = \log \frac{P(x^n|m, \theta^{k_m}) f(\theta^{k_m}|m)}{f(\theta^{k_m}|x^n, m)} \tag{163}$$

we have

$$\begin{aligned}
\log P(x^n|m) = \log P(x^n|m, \tilde{\theta}^{k_m}) - \frac{k_m}{2}\log\frac{n}{2\pi} \\
- \frac{\sqrt{\det I^*(\tilde{\theta}^{k_m}|m)}}{f(\tilde{\theta}^{k_m}|m)} + o(1); \quad \text{a.s.}
\end{aligned} \tag{164}$$

for $\forall m \in \mathcal{M}$. We have, therefore,

$$\begin{aligned}
&\log \frac{P(x^n|m)}{P(x^n|m^*)} \\
&\quad = \log \frac{P(x^n|m, \tilde{\theta}^{k_m})}{P(x^n|m^*, \tilde{\theta}^{k_{m^*}})} - \left(\frac{k_m}{2} - \frac{k_{m^*}}{2}\right)\log\frac{n}{2\pi} \\
&\qquad - \frac{f(\tilde{\theta}^{k_{m^*}}|m^*)\sqrt{\det I^*(\tilde{\theta}^{k_m}|m)}}{f(\tilde{\theta}^{k_m}|m)\sqrt{\det I^*(\tilde{\theta}^{k_{m^*}}|m^*)}} + o(1) \quad \text{a.s.}
\end{aligned} \tag{165}$$

for $\forall m \in \mathcal{M}$. On the other hand, from Condition 4, v), $\tilde{\theta}^{k_m}$ is a strong consistent estimator of $\theta^{k^*_m}$, that is, $\tilde{\theta}^{k_m} \to \theta^{k^*_m}$, almost surely for $\forall m \in \mathcal{M}$. We have, therefore, $f(\tilde{\theta}^{k_m}|m) = O(1)$ almost surely and $\det I(\tilde{\theta}^{k_m}|m) = O(1)$ almost surely, where $O(1)$ is the term such that $|O(1)| \le C$ almost surely. when $n \to \infty$ for $\exists C > 0$.

Therefore, from (165), if the equation

$$\log \frac{P(x^n|m, \tilde{\theta}^{k_m})}{P(x^n|m^*, \tilde{\theta}^{k_{m^*}})} - \frac{k_m - k_{m^*}}{2}\log n \to -\infty \quad \text{a.s.} \tag{166}$$

for $\forall m \neq m^*$ is proved, then the proof is completed.

[9]This means that the difference of code lenghts between the MDL code and the Bayes code does not diverge to ∞.

At first, we shall estimate $\log \frac{P(x^n|m,\tilde{\theta}^{k_m})}{P(x^n|m,\theta^{k_m^*})}$. From Taylor expansion, we have

$$\begin{aligned}
&-\log P(x^n|m,\theta^{k_m^*}) \\
&\quad = -\log P(x^n|m,\tilde{\theta}^{k_m}) - \left(\tilde{\theta}^{k_m}-\theta^{k_m^*}\right)^T \\
&\qquad \cdot \left.\frac{\partial \log P(x^n|m,\theta^{k_m})}{\partial \theta^{k_m}}\right|_{\theta^{k_m}=\tilde{\theta}^{k_m}} \\
&\qquad - \frac{1}{2}\left(\tilde{\theta}^{k_m}-\theta^{k_m^*}\right)^T \left.\frac{\partial^2 \log P(x^n|m,\theta^{k_m})}{\partial\theta^{k_m}(\partial\theta^{k_m})^T}\right|_{\theta^{k_m}=\tilde{\theta}^{k_m}} \\
&\qquad \cdot \left(\tilde{\theta}^{k_m}-\theta^{k_m^*}\right) \\
&\qquad + O\left(\left\|\tilde{\theta}^{k_m}-\theta^{k_m^*}\right\|^3 \left|\sum_{i=1}^{k_m}\sum_{j=1}^{k_m}\sum_{l=1}^{k_m}\right.\right. \\
&\qquad\qquad \left.\left.\cdot\left.\frac{\partial \log P(x^n|m,\theta^{k_m})}{\partial\theta_i^{k_m}\partial\theta_j^{k_m}\partial\theta_l^{k_m}}\right|_{\theta^{k_m}=\theta^{k_m}+}\right) \qquad (167)
\end{aligned}$$

for some $\theta^{k_m}+$ lying between $\tilde{\theta}^{k_m}$ and $\theta^{k_m^*}$.

From Condition 4, v) and vi), we have

$$\begin{aligned}
&-(\tilde{\theta}^{k_m}-\theta^{k_m^*})^T \left.\frac{\partial^2 \log P(x^n|m,\theta^{k_m})}{\partial\theta^{k_m}(\partial\theta^{k_m})^T}\right|_{\theta^{k_m}=\tilde{\theta}^{k_m}} (\tilde{\theta}^{k_m}-\theta^{k_m^*}) \\
&\quad = \sqrt{n}(\tilde{\theta}^{k_m}-\theta^{k_m^*})^T I^*(\tilde{\theta}^{k_m}|m)\sqrt{n}(\tilde{\theta}^{k_m}-\theta^{k_m^*}) \\
&\qquad + O\left(\frac{(\log\log n)^{3/2}}{\sqrt{n}}\right) \quad \text{a.s.} \qquad (168)
\end{aligned}$$

On the other hand, from

$$\left\|\tilde{\theta}^{k_m}-\theta^{k_m^*}\right\| = O\left(\frac{(\log\log n)^{1/2}}{\sqrt{n}}\right) \quad \text{a.s.} \qquad (169)$$

and (110), we have

$$\begin{aligned}
&\left\|\tilde{\theta}^{k_m}-\theta^{k_m^*}\right\|^3 \left|\sum_{i=1}^{k_m}\sum_{j=1}^{k_m}\sum_{l=1}^{k_m}\frac{\partial \log P(x^n|m,\theta^{k_m})}{\partial\theta_i^{k_m}\partial\theta_j^{k_m}\partial\theta_l^{k_m}}\right|_{\theta^{k_m}=\theta^{k_m}+} \\
&\qquad = O\left(\frac{(\log\log n)^{3/2}}{\sqrt{n}}\right) \quad \text{a.s.} \qquad (170)
\end{aligned}$$

because $\theta^{k_m}+ \to \theta^{k_m^*}$, almost surely. We have, therefore,

$$\begin{aligned}
&-\log P(x^n|m,\theta^{k_m^*}) \\
&\quad = -\log P(x^n|m,\tilde{\theta}^{k_m}) \\
&\qquad - \left(\tilde{\theta}^{k_m}-\theta^{k_m^*}\right)^T \left.\frac{\partial \log P(x^n|m,\theta^{k_m})}{\partial\theta^{k_m}}\right|_{\theta^{k_m}=\tilde{\theta}^{k_m}} \\
&\qquad + \frac{1}{2}\sqrt{n}(\tilde{\theta}^{k_m}-\theta^{k_m^*})^T I^*(\tilde{\theta}^{k_m}|m)\sqrt{n}(\tilde{\theta}^{k_m}-\theta^{k_m^*}) \\
&\qquad + O\left(\frac{(\log\log n)^{3/2}}{\sqrt{n}}\right) \quad \text{a.s.} \qquad (171)
\end{aligned}$$

Next, we show the order of $\frac{1}{n}\left.\frac{\partial \log P(x^n|m,\theta^{k_m})}{\partial\theta^{k_m}}\right|_{\theta^{k_m}=\tilde{\theta}^{k_m}}$. From Taylor expansion, we have

$$\begin{aligned}
&\frac{1}{n}\left.\frac{\partial \log P(x^n|m,\theta^{k_m})}{\partial\theta^{k_m}}\right|_{\theta^{k_m}=\tilde{\theta}^{k_m}} \\
&\quad = \frac{1}{n}\left.\frac{\partial \log P(x^n|m,\theta^{k_m})}{\partial\theta^{k_m}}\right|_{\theta^{k_m}=\theta^{k_m^*}} \\
&\qquad + \frac{1}{n}(\tilde{\theta}^{k_m}-\theta^{k_m^*})^T \left.\frac{\partial^2 \log P(x^n|m,\theta^{k_m})}{\partial\theta^{k_m}(\partial\theta^{k_m})^T}\right|_{\theta^{k_m}=\theta^{k_m}++} \qquad (172)
\end{aligned}$$

for some $\theta^{k_m}++$ lying between $\tilde{\theta}^{k_m}$ and $\theta^{k_m^*}$. From Condition 4, vi), we have

$$\begin{aligned}
&\frac{1}{n}\left.\frac{\partial \log P(x^n|m,\theta^{k_m})}{\partial\theta^{k_m}}\right|_{\theta^{k_m}=\theta^{k_m^*}} \\
&\quad = \left.\frac{\partial H^*(m,\theta^{k_m})}{\partial\theta^{k_m}}\right|_{\theta^{k_m}=\theta^{k_m^*}} + O\left(\frac{(\log\log n)^{1/2}}{\sqrt{n}}\right) \quad \text{a.s.} \qquad (173)
\end{aligned}$$

where $\left.\frac{\partial H^*(m,\theta^{k_m})}{\partial\theta^{k_m}}\right|_{\theta^{k_m}=\theta^{k_m^*}} = \mathbf{0}$.

On the other hand, from identical discussion with (168), we have

$$\begin{aligned}
&\frac{1}{n}(\tilde{\theta}^{k_m}-\theta^{k_m^*})^T \left.\frac{\partial^2 \log P(x^n|m,\theta^{k_m})}{\partial\theta^{k_m}(\partial\theta^{k_m})^T}\right|_{\theta^{k_m}=\theta^{k_m}++} \\
&\quad = (\tilde{\theta}^{k_m}-\theta^{k_m^*})^T I^*(\theta^{k_m}++|m) + O\left(\frac{\log\log n}{n}\right) \quad \text{a.s.} \qquad (174)
\end{aligned}$$

from

$$\|\tilde{\theta}^{k_m}-\theta^{k_m^*}\| = O\left(\frac{(\log\log n)^{1/2}}{\sqrt{n}}\right) \quad \text{a.s.}$$

Accordingly, from (172)–(174), we have

$$\begin{aligned}
&\left(\tilde{\theta}^{k_m}-\theta^{k_m^*}\right)^T \left.\frac{\partial \log P(x^n|m,\theta^{k_m})}{\partial\theta^{k_m}}\right|_{\theta^{k_m}=\tilde{\theta}^{k_m}} \\
&\qquad = O(\log\log n) \quad \text{a.s.} \qquad (175)
\end{aligned}$$

because

$$(\tilde{\theta}^{k_m}-\theta^{k_m^*})^T I^*(\theta^{k_m}++|m) = O\left(\frac{(\log\log n)^{1/2}}{\sqrt{n}}\right) \quad \text{a.s.}$$

which is given by Condition 4, ii) and v), and $\theta^{k_m}++ \to \theta^{k_m^*}$ almost surely.

From Condition 4, ii) and v), we have $I^*(\tilde{\theta}^{k_m}|m) \to I(\theta^{k_m^*}|m)$ almost surely. We have, therefore,

$$\log \frac{P(x^n|m,\tilde{\theta}^{k_m})}{P(x^n|m,\theta^{k_m^*})} = O(\log\log n) \quad \text{a.s.} \qquad (176)$$

from (171) and (175). Then we have

$$\begin{aligned}
&\log \frac{P(x^n|m,\tilde{\theta}^{k_m})}{P(x^n|m^*,\tilde{\theta}^{k_{m^*}})} - \frac{k_m-k_{m^*}}{2}\log n \\
&\quad = \log \frac{P(x^n|m,\theta^{k_m^*})}{P(x^n|m^*,\theta^{k_{m^*}^*})} - \frac{k_m-k_{m^*}}{2}\log n + O(\log\log n) \text{ a.s.} \qquad (177)
\end{aligned}$$

At first, we consider the case $k_{m^*} > k_m$. In this case, $H^*(m,\theta^{k_m}) > H^*(m^*,\theta^{k_{m^*}^*})$ is satisfied. Then we obtain

$$\log \frac{P(x^n|m,\theta^{k_m})}{P(x^n|m,\theta^{k_{m^*}^*})} = -Cn + o(n) \quad \text{a.s.} \qquad (178)$$

for $\forall \theta^{k_m} \in \Theta^{k_m}$ from Condition 4, vi), where C is some positive constant.

We have, therefore,

$$\begin{aligned}
&\log \frac{P(x^n|m,\tilde{\theta}^{k_m})}{P(x^n|m^*,\tilde{\theta}^{k_{m^*}})} - \frac{k_m-k_{m^*}}{2}\log n \\
&\quad = -Cn - \frac{k_m-k_{m^*}}{2}\log n + o(n) \to -\infty, \text{ a.s.} \qquad (179)
\end{aligned}$$

for $\forall m \in \mathcal{M}$ when $k_{m^*} > k_m$.

When $k_{m^*} < k_m$, the equation

$$\frac{1}{n}\log P(x^n|m, \theta^{k_m^*}) = \frac{1}{n}\log P(x^n|m^*, \theta^{k_{m^*}^*}) + O\left(\frac{(\log\log n)^{1/2}}{\sqrt{n}}\right) \quad \text{a.s.} \tag{180}$$

is satisfied from Condition 4, vi) and the definition of m^*. We have, therefore,

$$\log\frac{P(x^n|m, \tilde{\theta}^{k_m})}{P(x^n|m^*, \tilde{\theta}^{k_{m^*}})} - \frac{k_m - k_{m^*}}{2}\log n = \frac{k_{m^*} - k_m}{2}\log n + O(\log n \log n) \to -\infty \quad \text{a.s.} \tag{181}$$

for $\forall m \in \mathcal{M}$ when $k_{m^*} < k_m$ from (177). Then (166) is satisfied for $\forall k_m \neq k_{m^*}$, and (162) holds.

From Bayes theorem

$$P(m|x^n) \propto P(x^n|m)P(m) \tag{182}$$

and (162), we have

$$\left|\frac{P(m|x^n)}{P(m^*|x^n)}\right| \to 1 \quad \text{a.s.} \tag{183}$$

for $\forall m \neq m^*$ because $|\mathcal{M}|$ is finite. This means (114). □

ACKNOWLEDGMENT

The authors wish to thank one of the anonymous referees for the valuable comments that led to the improvement of the first version of this paper. The authors would also like to acknowledge M. Nakazawa, Dr. M. Kobayashi, and all the members of Hirasawa Laboratory and Matsushima Laboratory for their helpful suggestions and discussions to this work.

REFERENCES

[1] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automat. Contr.*, vol. AC-19, no. 6, pp. 716–722, 1974.
[2] J. M. Bernardo and A. F. M. Smith, *Bayesian Theory*. New York: Wiley, 1994.
[3] R. E. Blahut, *Principles and Practice of Information Theory*. Reading, MA:: Addison-Weley, 1987.
[4] B. S. Clarke and A. R. Barron, "Information-theoretic asymptotics of Bayes methods," *IEEE Trans. Inform. Theory*, vol. 36, pp. 453–471, May 1990.
[5] ——, "Jeffreys' prior is asymptotically least favorable under entropy risk," *JSPI*, vol. 41, pp. 37–60, 1994.
[6] B. S. Clarke, "Asymptotic normality of the posterior in relative entropy," *IEEE Trans. Inform. Theory*, vol. 45, pp. 165–176, Jan. 1999.
[7] C.-F. Chen, "On asymptotic normality of limiting density function with Bayesian implications," *J. Roy. Statist. Soc. B*, vol. 47, no. 3, pp. 540–546, 1985.
[8] W. Feller, *An Introduction to Probability and Its Applications*. New York: Wiley, 1957, 1966, vol. I–II.
[9] T. S. Ferguson, *Mathematical Statistics, A Decision Theoretic Approach*. New York and London: Academic, 1967.
[10] M. Gotoh, T. Matsushima, and S. Hirasawa, "A generalization of B.S. Clarke and A. R. Barron's asymptotics of Bayes codes for FSMX sources," *IEICE Trans. Fundamentals*, vol. E81-A, no. 10, pp. 2123–2132, 1998.
[11] ——, "Almost sure and mean convergence of extended stochastic complexity," *IEICE Trans. Fundamentals*, vol. E82-A, no. 10, pp. 2129–2137, 1999.
[12] J. A. Hartigan, *Bayes Theory*. Berlin, Germany: Springer-Verlag, 1983.
[13] P. Hall and E. J. Hannan, "On stochastic complexity and nonparametric density estimation," *Biometrika*, vol. 75, no. 4, pp. 705–714, 1988.
[14] T. S. Han and K. Kobayashi, *Mathematics of Information and Coding* (in Japanese): Iwanami Syoten, 1994.
[15] E. J. Hannan and B. G. Quinn, "The determination of the order of an autoregression," *J. Roy. Statist. Soc. B*, vol. 41, no. 2, pp. 190–195, 1979.
[16] C. C. Heyde and I. M. Johnstone, "On asymptotic posterior normality for stochastic process," *J. Roy. Statist. Soc. B*, vol. 41, no. 2, pp. 184–189, 1979.
[17] S. Itoh, "Application of MDL principle to pattern classification problems" (in Japanese), *J. JSAI*, vol. 7, no. 4, pp. 608–614, 1992.
[18] T. Kawabata, "Bayes codes and context tree weighting method," (in Japanese), IEICE, Tech. Rep. IT93-121, 1994.
[19] A. S. Martini and F. Spezzaferri, "A predictive model selection criterion," *J. Roy. Stasist. Soc. B*, vol. 46, no. 2, pp. 296–303, 1984.
[20] T. Matsushima, H. Inazumi, and S. Hirasawa, "A class of distortionless codes designed by Bayes decision theory," *IEEE Trans. Inform. heory*, vol. 37, pp. 1288–1293, Sept. 1991.
[21] T. Matsushima and S. Hirasawa, "A Bayes coding algorithm for Markov models," IEICE , Tech. Rep. IT95-1, 1995.
[22] D. S. Poskitt, "Precision, complexity and Bayesian model determination," *J. Roy. Statist. Soc. B*, vol. 49, no. 2, pp. 199–208, 1987.
[23] G. Qian, G. Gabor, and R. P. Gupta, "On stochastic complexity estimation: A decision-theoretic approach," *IEEE Trans. Inform. Theory*, vol. 40, pp. 1181–1191, July 1994.
[24] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 46, pp. 465–471, 1978.
[25] ——, "Universal modeling and coding," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 12–23, Jan. 1981.
[26] ——, "A universal prior for integers and estimation by minimum description length," *Ann. Statist.*, vol. 11, no. 2, pp. 416–431, 1983.
[27] ——, "Universal coding, information, prediction, and estimation," *IEEE Trans. Inform. Theory*, vol. IT-30, July 1984.
[28] ——, "Stochastic complexity and modeling," *Ann. Statist.*, vol. 14, no. 3, pp. 1080–1100, 1986.
[29] ——, "Stochastic complexity," *J. Roy. Statist., Soc. B.*, vol. 49, pp. 223–265, 1987.
[30] ——, "Fisher information and stochastic complexity," *IEEE Trans. Inform. Theory*, vol. 42, pp. 40–47, Jan. 1996.
[31] C. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, pp. 461–464, 1978.
[32] R. Shibata, "Consistency of model selection and parameter estimation," *Appl. Probab. Trust*, pp. 127–141, 1986.
[33] J. Suzuki, "Some notes on universal noiseless coding," *IEICE Trans. Fundamentals*, vol. E78-A, no. 12, 1995.
[34] J. Takeuchi, "On the convergence rate of the MDL estimator with respect to the KL-divergence" (in Japanese), in *Proc. 16th Symp. Information Theory and Its Applications*, 1993.
[35] ——, "Characterization of the Bayes estimator and the MDL estimator for exponential families," *IEEE Trans. Inform. Theory*, vol. 43, pp. 1165–1174, July 1996.
[36] H. Tsuchiya, S. Itoh, and T. Hashimoto, "An algorithm for designing a pattern classifier by using MDL criterion," *IEICE Trans. Fundamentals*, vol. E79-A, no. 6, pp. 910–920, 1996.
[37] C. S. Wallace and P. R. Freeman, "Estimation and inference by compact coding," *J. Roy. Statist. Soc. B*, vol. 49, no. 3, pp. 240–265, 1987.
[38] F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens, "The context-tree weighting method: Basic properties," *IEEE Trans. Inform. Theory*, vol. 41, pp. 653–664, May 1995.
[39] K. Yamanishi and T. S. Han, "Introduction to MDL from viewpoints of information theory" (in Japanese), *J. JSAI*, vol. 7, no. 3, pp. 427–434, 1992.