PAPER Special Section on Information Theory and Its Applications

A Generalization of B.S. Clarke and A.R. Barron's Asymptotics of Bayes Codes for FSMX Sources

Masayuki GOTOH†, Toshiyasu MATSUSHIMA†, *and* Shigeichi HIRASAWA†, *Members*

SUMMARY We shall generalize B.S. Clarke and A.R. Barron's analysis of the Bayes method for the FSMX sources. The FSMX source considered here is specified by the set of all states and its parameter value. At first, we show the asymptotic codelengths of individual sequences of the Bayes codes for the FSMX sources. Secondly, we show the asymptotic expected codelengths. The Bayesian posterior density and the maximum likelihood estimator satisfy asymptotic normality for the finite ergodic Markov source, and this is the key of our analysis.

key words: *Bayes code, source coding, universal coding, universal modeling*

1. Introduction

Bayes coding [2], [7], whose codelength is also called stochastic complexity [9] is Bayes optimal solution based on Bayes decision theory [7]. This is the method which uses mixture probability of all models in model class for coding function.

The properties of the Bayes code have been studied from various viewpoints [2], [3], [6], [7], [9], [10], [13]. B.S. Clarke and A.R. Barron analyzed asymptotic mean codelength of the Bayes code [2] and showed that Jeffreys prior is asymptotically least favorable under entropy risk [3] for i.i.d. parametric sources. Recently, the efficient algorithms which calculate the mixture probability of the data sequence similar to the context tree weighting (CTW) method have been reported for the FSMX model class [6], [8]. Since the FSMX sources are not i.i.d. sources, the analysis of codelength for the FSMX sources is important.

In this paper, we generalize the part of Clarke and Barron's analysis of the Bayes code for the (ergodic) FSMX sources. The FSMX model class is one of the partial nested model class. At first, we show the asymptotic codelengths of the Bayes codes for individual sequences for the FSMX sources. J. Rissanen have advocated the stochastic complexity and the universal modeling and he stated in [10] such that instead of measuring a code's performance by mean codelength as in universal coding, the central question of interest in universal modeling is the codelength achievable for individual sequences. Then an analysis in this paper of the Bayes codelengths for the individual sequences will be important for this concept. Secondly, we shall prove that Clarke and Barron's asymptotics of the Bayes code (Expected codelength) are generalized for the FSMX sources which are useful for the source coding in practice. The key of the analysis is that the posterior probability of the parameter satisfies asymptotic normality. Clarke and Barron have described in [2] that their asymptotics can be intuitively stated from the asymptotic normality. Generally speaking, the asymptotic normalcy of posterior distribution holds for other than i.i.d. sources [1], [4]. We directly use the asymptotic normality for the analysis and show the similar asymptotics are satisfied for the FSMX sources.

Manuscript received January 16, 1998.
Manuscript revised April 15, 1998.

†The authors are with the Department of the Industrial and Management Systems Engineering, School of Science and Engineering, Waseda University, Tokyo, 169–8555 Japan.

2. Preliminaries

2.1 The FSMX Sources

Let $\mathcal{X} = \{0, 1, 2, \cdots, \beta\}$ be the discrete source alphabet. We denote the data sequence with length n emitted from the source by $x^n = x_1 x_2 \cdots x_n$, where $\forall i,\ x_i \in \mathcal{X}$. $\mathcal{X}^n$ is the set of all x^n. We also denote the infinite data sequence by x^∞.

An FSMX source model m is specified by the set of the states and is represented by a $(\beta+1)$-ary complete tree $T(m)$ called a context tree*. Each arc in tree corresponds to a symbol $x \in \mathcal{X}$. A path from a leaf to the root in the tree represents a context or state in the FSMX model. The state of an FSMX model at t is determined by the source sequence x^{t-1}. We denote this mapping from x^{t-1} to a state of the FSMX model m by $\varphi_m(x^{t-1})$. Let $s(m)$ be a state of model m, and $S(m)$, the set of all states in m. $S(m)$ also represents the set of all leaf nodes in the context tree $T(m)$.

An FSMX model determines the parameter space. Then, we denote the k_m-dimensional parameter and the parameter space by θ^{k_m} and Θ^{k_m} respectively.

We assume $\Theta^{k_m} = (0,1)^{k_m}$, where $(0,1)^{k_m}$ represents interior of the k_m-dimensional rectangle with sides $[0,1]$. Let $\theta^{k_m}_{i,j}$ be the probability of the symbol i at the j-th state $s_j(m)$, and $q_{s_j(m)}$, the stationary probability of the state $s_j(m)$ calculated by θ^{k_m}, where $i \in \mathcal{X}$

*Since the FSMX model is one of the Markov model, the probability structure is determined by the set of the states and the transition probabilities.

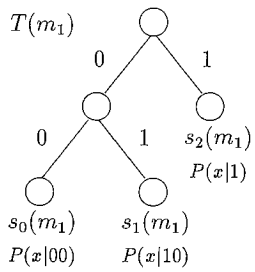


Fig. 1 An example: $T(m_1)$.

and $j = 0, 1, \cdots, |S(m)| - 1$. We may regard $\theta^{k_m} = (\theta^{k_m}_{0,0}, \cdots, \theta^{k_m}_{\beta-1,|S(m)|-1})^T$ as $\beta|S(m)|$-dimensional continuous parameter, that is $k_m = \beta|S(m)|$.

The probabilistic structure of an FSMX model m is represented by the conditional probabilities of x_t given x^{t-1} denoted by $P(x_t|x^{t-1}, m, \theta^{k_m})$, where θ^{k_m} is $\beta|S(m)|$-dimensional parameter vector representing the probabilities of each symbol at each state of m. The notation $P(\cdot)$ stands for the probability.

The probability of the next symbol x_t given by x^{t-1} is

$$P(x_t|x^{t-1}, m, \theta^{k_m}) = P(x_t|\varphi_m(x^{t-1}), m, \theta^{k_m}). \quad (1)$$

In Fig. 1, we show the context tree of an FSMX model m_1 for binary alphabet, $\beta = 1$.

Since the context $x^{t-1} = \cdots 10$ is represented by the path from the leaf node s_1 to the root, the state determined by the context corresponds to the leaf s_1. The $S(m_1)$ is given by $\{s_0(m_1), s_1(m_1), s_2(m_1)\}$. If $x^5 = 10010$, then the state at $t = 2$ is $s_2(m_1) = \varphi_{m_1}(1)$, the state at $t = 3$ is $s_1(m_1) = \varphi_{m_1}(10)$, the state at $t = 4$ is $s_0(m_1) = \varphi_{m_1}(100)$ and so on. The parameter $\theta^{k_{m_1}}$ of the FSMX model m_1 in Fig. 1 is given by $(\theta^{k_{m_1}}_{0,0}, \theta^{k_{m_1}}_{0,1}, \theta^{k_{m_1}}_{0,2})^T = (P(0|00), P(0|10), P(0|1))^T$ and $k^{m_1} = 3$.

The set of m is denoted by $\mathcal{M}$. $\mathcal{M}$ is a finite countable set. We assume that the data sequence is derived from the true model m^* with the true parameter $\theta^{k^*_{m^*}}$. We also assume that the true model m^* exists in $\mathcal{M}$ and m^* has (finite) k^{m^*}−dimensional parameter $\theta^{k_{m^*}}$. The true stationary probability $q^*_{j(m^*)}$ on $S(m^*)$ is specified by $\theta^{k^*_{m^*}}$. An FSMX model m specifies a parametric model class. Let $\mathcal{H}^{k_m}$ be this model class of m: $\mathcal{H}^{k_m} = \{P(\cdot|m, \theta^{k_m})|\theta^{k_m} \in \Theta^{k_m}\}$. Then the FSMX model class $\mathcal{H}$ is defined by

$$\mathcal{H} = \cup_m \mathcal{H}^{k_m}, \quad (2)$$

where, the nested structure

$$\mathcal{H}^{k_{m1}} \subset \mathcal{H}^{k_{m2}} \subset \mathcal{H}^{k_{m3}} \subset \cdots, \quad (3)$$

is partially satisfied for $m1, m2, \cdots \in \mathcal{M}$. That is, the model set $\mathcal{M}$ has partial order.

Remark the fact such that if $\mathcal{H}^{k_{m^*}}$ does not equal to $\mathcal{H}$, then there exists $m \neq m^*$, θ^{k_m}, that satisfies $P(x^n|m, \theta^{k_m}) = P(x^n|m^*, \theta^{k^*_{m^*}})$. Then, we redefine the true model m^* as follows:

$$m^* = \arg_m \min \left\{ k_m \middle| \exists \theta^{k_m}, \forall x^n, P(x^n|m, \theta^{k_m}) = P^*(x^n) \right\}, \quad (4)$$

where $P^*(x^n)$ is true distribution from which the data sequence x^n is derived. Of course, $P^*(\cdot) = P(\cdot|m^*, \theta^{k^*_{m^*}})$. We also denote the model class by $\{P(\cdot|m, \theta^{k_m})|m \in \mathcal{M}, \theta^{k_m} \in \Theta^{k_m}\}$.

We assume the initial state is known. Let $n^{(m)}_0, \cdots n^{(m)}_{|S(m)|-1}$ be appearance numbers of the states $s_0(m)$, $\cdots s_{|S(m)|-1}(m)$, and $n^{(m)}_{0,j}, n^{(m)}_{1,j}, \cdots, n^{(m)}_{\beta,j}$, appearance numbers of the symbol $0, 1, \cdots, \beta$ conditioned by the state $s_j(m)$ in data sequence x^n. That is, $n = \sum_j n^{(m)}_j$ and $n^{(m)}_j = \sum_i n^{(m)}_{i,j}$. Then the likelihood function $P(x^n|m, \theta^{k_m})$ is given by

$$P(x^n|m, \theta^{k_m}) = \prod_{j=0}^{|S(m)|-1} \prod_{i=0}^{\beta} (\theta^{k_m}_{i,j})^{n^{(m)}_{i,j}}, \quad (5)$$

where $\theta^{k_m}_{\beta,j} = 1 - \sum_{i=0}^{\beta-1} \theta^{k_m}_{i,j}$, and this function has an unique maximum. The elements of the maximum likelihood estimator of θ^{k_m}, $\hat{\theta}^{k_m}$, are given by

$$\hat{\theta}^{k_m}_{i,j} = \frac{n^{(m)}_{i,j}}{n^{(m)}_j}. \quad (6)$$

We define the information matrix $I(\theta^{k_m}|m)$ such as

$$I(\theta^{k_m}|m) = -\lim_{n\to\infty} \frac{1}{n} E^* \frac{\partial^2 \log P(x^n|m, \theta^{k_m})}{\partial\theta^{k_m}(\partial\theta^{k_m})^T}, \quad (7)$$

where E^* represents the expectation by $P(\cdot|m^*, \theta^{k^*_{m^*}})$. That is, $I(\theta^{k_m}|m)$ of the FSMX source is given by

$$I(\theta^{k_m}|m) = -\frac{\partial^2 \sum_{i,j} q^*_{s_j(m)} \theta^{k^*_m}_{i,j} \log \theta^{k_m}_{i,j}}{\partial\theta^{k_m}(\partial\theta^{k_m})^T}. \quad (8)$$

If $m = m^*$ and $\theta^{k_m} = \theta^{k^*_{m^*}}$, then $I(\theta^{k^*_{m^*}}|m^*)$ reduces to the Fisher information matrix and $\det I(\theta^{k^*_{m^*}}|m^*)$ is given by

$$\det I(\theta^{k^*_{m^*}}|m^*) = \prod_{j=0}^{|S(m^*)|-1} (q^*_{s_j(m^*)})^{\beta} \frac{1}{\theta^{k^*_{m^*}}_{0,j} \theta^{k^*_{m^*}}_{1,j} \cdots \theta^{k^*_{m^*}}_{\beta,j}}, \quad (9)$$

respectively.

The FSMX model is one of the finite ergodic Markov model class, and the following property has been known.

Lemma 1 (The iterated logarithm law [5]): For the true model m^*, we have

$$\hat{\theta}^{k_{m^*}} = \theta^{k^*_{m^*}} + O\left(\left(\frac{\log\log n}{n}\right)^{1/2}\right), \quad a.s. \quad (10)$$

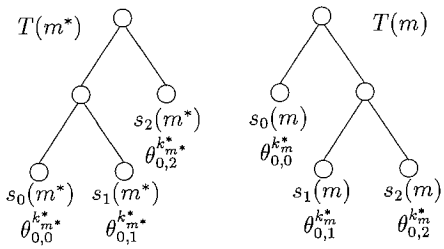


Fig. 2 An example of $\theta^{k_m^*}$.

$$\frac{n_j^{(m^*)}}{n} = q^*_{s_j(m^*)} + O\left(\left(\frac{\log\log n}{n}\right)^{1/2}\right), \quad a.s. \tag{11}$$

where $a.s.$ represents *almost sure convergence*†. □

Next, we consider the case such that the parameter θ^{k_m} of $m \neq m^*$ is estimated from x^n. We define $\theta^{k_m^*}_{i,j}$ for $m \neq m^*$ as follows: If a leaf of a state $s_j(m)$ of $T(m)$ corresponds to a leaf of $s_l(m^*)$ of $T(m^*)$ or is a descendant leaf of an intermediate node of $T(m)$ corresponding $s_l(m^*)$ of $T(m^*)$, then we define

$$\theta^{k_m^*}_{i,j} = \theta^{k_{m^*}^*}_{i,l}. \tag{12}$$

Else if a leaf of a state $s_j(m)$ of $T(m)$ corresponds to an intermediate node of $T(m^*)$, we define

$$\theta^{k_m^*}_{i,j} = \sum_{s_l(m^*)\in\mathcal{S}(j,m,m^*)} \theta^{k_{m^*}^*}_{i,l} \frac{q^*_{s_l(m^*)}}{q^*_{\mathcal{S}(j,m,m^*)}}, \tag{13}$$

where $\mathcal{S}(i,m,m^*)$ is the set of all descendant leaf of the intermediate node of $T(m^*)$ corresponding to $s_j(m)$ and $q^*_{\mathcal{S}(j,m,m^*)} = \sum_{s_l(m^*)\in\mathcal{S}(j,m,m^*)} q^*_{s_l(m^*)}$. An example of $\theta^{k_m^*}$ for $m \neq m^*$ is shown in Fig. 2. From above definition, we have $\theta^{k_m^*}_{0,0} = \frac{q^*_{s_0(m^*)}}{q^*_{s_0(m^*)}+q^*_{s_1(m^*)}}\theta^{k_{m^*}^*}_{0,0} + \frac{q^*_{s_1(m^*)}}{q^*_{s_0(m^*)}+q^*_{s_1(m^*)}}\theta^{k_{m^*}^*}_{0,1}$ and $\theta^{k_m^*}_{0,1} = \theta^{k_m^*}_{0,2} = \theta^{k_{m^*}^*}_{0,2}$ for m in Fig. 2. Then, from the property of the multinomial distribution, we can regard $\theta^{k_m^*}$ as the true parameter from which x^n is derived on condition that m is fixed. From Lemma 1, we can see that the following corollary is clearly satisfied.

Corollary 1: For $\forall m \in \mathcal{M}$, we have

$$\hat{\theta}^{k_m} = \theta^{k_m^*} + O\left(\left(\frac{\log\log n}{n}\right)^{1/2}\right), \quad a.s. \tag{14}$$

$$\frac{n_{i,j}^{(m)}}{n} = q^*_{s_j(m)} + O\left(\left(\frac{\log\log n}{n}\right)^{1/2}\right), \quad a.s. \tag{15}$$

□

2.2 The Bayes Code for FSMX Model Class

In this paper, we discuss ideal codelength $-\log q(x^n)$. We call $q(x^n)$ coding function. We suppose that the logarithm base is e and the measure of the codelength is nats through the paper.

Let $P(m)$ and $f(\theta^{k_m}|m)$ be prior probability of model m and prior density of parameter θ^{k_m} respectively. Through the paper, $f(\cdot)$ represents the probability density function. $P(x^n|m)$ and $f(\theta^{k_m}|x^n,m)$ are given by

$$P(x^n|m) = \int_{\theta^{k_m}} P(x^n|m,\theta^{k_m}) f(\theta^{k_m}|m) d\theta^{k_m}, \tag{16}$$

$$f(\theta^{k_m}|x^n,m) = \frac{P(x^n|m,\theta^{k_m}) f(\theta^{k_m}|m)}{P(x^n|m)}, \tag{17}$$

respectively. Finally, we define the Bayes code for the FSMX model class.

Definition 1 (The Bayes code)**:** The codelength of the Bayes code $L^{m,\theta^{k_m}}_{Bayes}(x^n)$ is given by

$$\begin{aligned} &L^{m,\theta^{k_m}}_{Bayes}(x^n) \\ &= -\log\sum_m \int_{\theta^{k_m}} P(x^n|m,\theta^{k_m}) f(\theta^{k_m}|m) P(m) d\theta^{k_m} \\ &= -\log\sum_m P(x^n|m)P(m). \end{aligned} \tag{18}$$

Here, the summation and the integral are calculated through $\mathcal{M}$ and Θ^{k_m} respectively. □

3. Main Results

3.1 Condition

We assume the following condition.

Condition 1:

i) $\Theta^{k_m} = (0,1)^{k_m}$ for $\forall m \in \mathcal{M}$ and $\theta^{k_{m^*}^*} \in \Theta^{k_{m^*}}$.

ii) For $\forall m \in \mathcal{M}$ and $\forall\theta^{k_m} \in \Theta^{k_m}$, $f(\theta^{k_m}|m) > 0$. And $P(m^*) > 0$.

iii) For $\forall m \in \mathcal{M}$, $f(\theta^{k_m}|m)$ is three times continuously differentiable for θ^{k_m} in Θ^{k_m}.

Remark 1: Here, we consider about Condition 1, iii). For example, the Dirichlet distribution which is the conjugate prior for the multinomial distribution class is obviously three times continuously differentiable. This prior is useful for an FSMX model class [8].

3.2 Asymptotic Normality of Posterior Density

At first, in order to show the main theorem, we quote

†"$X_n \to X$, $a.s.$" means $P^*\{|X_n - X| > \epsilon$, *infinitely often*$\} = 0$ for $\forall\epsilon > 0$. This is equivalent to $P^*\{X_n \to X\} = 1$. And the description such as "$|X_n - X| \leq \epsilon$, $a.s.$ when $n \to \infty$" also means $P^*\{|X_n - X| > \epsilon$, *infinitely often*$\} = 0$.

the following lemma from [1], pp.285–297. The following lemma shows the asymptotic normality of the posterior density $f(\theta^{k_m}|x^n,m)$ on Θ^{k_m} when we regard θ^{k_m} as the random variable.

Lemma 2 ([1],[4]): Fix an infinite sequence x^∞. And let $\tilde{\theta}^{k_m}$ be a strict local maximum of $L_n(\theta^{k_m}|m) = \log f(\theta^{k_m}|x^n,m)$ satisfying

$$L_n'(\tilde{\theta}^{k_m}|m) = \mathbf{0}, \tag{19}$$

and implying positive definiteness of

$$\Sigma_n = -\left(L_n''(\tilde{\theta}^{k_m}|m)\right)^{-1}, \tag{20}$$

where $L_n'(\tilde{\theta}^{k_m}|m)$ and $L_n''(\tilde{\theta}^{k_m}|m)$ are defined by

$$L_n'(\tilde{\theta}^{k_m}|m) = \left.\frac{\partial L_n(\theta^{k_m}|m)}{\partial \theta^{k_m}}\right|_{\theta^{k_m}=\tilde{\theta}^{k_m}}, \tag{21}$$

$$L_n''(\tilde{\theta}^{k_m}|m) = \left.\frac{\partial^2 L_n(\theta^{k_m}|m)}{\partial \theta^{k_m}(\partial\theta^{k_m})^T}\right|_{\theta^{k_m}=\tilde{\theta}^{k_m}}, \tag{22}$$

respectively. Defining $B_\delta(\tilde{\theta}^{k_m}) = \left\{\theta^{k_m} \in \Theta^{k_m} \middle| \|\theta^{k_m} - \tilde{\theta}^{k_m}\| < \delta\right\}$, the following three basic conditions are necessary and sufficient for the asymptotic normality of the posterior distribution.

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

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

$$I - A(\epsilon) \leqq L_n''(\theta^{k_m}|m)\left\{L_n''(\tilde{\theta}^{k_m}|m)\right\}^{-1} \leqq I + A(\epsilon), \tag{23}$$

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

(c.3) "*Concentration*" For any $\delta > 0$, there exists N and c, $d > 0$ such that, for any $n > N$ and $\theta^{k_m} \notin B_\delta(\tilde{\theta}^{k_m})$,

$$L_n(\theta^{k_m}|m) - L_n(\tilde{\theta}^{k_m}|m) < -c\left\{(\theta^{k_m} - \tilde{\theta}^{k_m})^T\Sigma_n^{-1}(\theta^{k_m} - \tilde{\theta}^{k_m})\right\}^d. \tag{24}$$

The conditions (c.1), (c.2), and (c.3) imply that[†]

$$f(\tilde{\theta}^{k_m}|x^n,m)\,(\det\Sigma_n)^{1/2} = (2\pi)^{-k_m/2} + o(1). \tag{27}$$

□

In above lemma, (c.1), (c.2), and (c.3) are the conditions for a sequence of $f(\theta^{k_m}|x^n,m)$, $n = 1,2,\cdots$ for the fixed x^∞. In fact, x^n is emitted from $P^*(x^n)$ and we should discuss the convergence based on the probability theory. Therefore, we shall show that the conditions of above lemma are *almost surely* satisfied for the FSMX sources in the following[††]:

Since $\frac{1}{n}L_n(\theta^{k_m}) = \frac{1}{n}\log f(\theta^{k_m}|x^n,m)$ is given by

$$\frac{1}{n}L_n(\theta^{k_m}|m) = \sum_{i,j}\frac{n_{i,j}^{(m)}}{n}\log\theta_{i,j}^{k_m} + \frac{1}{n}\log f(\theta^{k_m}|m) - \frac{1}{n}\log P(x^n|m), \tag{28}$$

where $f(\theta^{k_m}|m)$ does not depend on n. Therefore

$$\frac{1}{n}L_n(\theta^{k_m}|m) \to \sum_{i,j}\frac{n_{i,j}^{(m)}}{n_j^{(m)}}\frac{n_j^{(m)}}{n}\log\theta_{i,j}^{k_m} - \frac{1}{n}\log P(x^n|m), \tag{29}$$

and $\log P(x^n|m)$ does not depend on θ^{k_m}, $\tilde{\theta}^{k_m}$ maximizing $\frac{1}{n}L_n(\theta^{k_m}|m)$ converges to $\hat{\theta}^{k_m}$.

From Condition 1, iii),

$$\frac{\partial^2 L_n(\theta^{k_m}|m)}{\partial\theta^{k_m}(\partial\theta^{k_m})^T} = \frac{\partial^2\sum_{i,j}n_{i,j}^{(m)}\log\theta_{i,j}^{k_m}}{\partial\theta^{k_m}(\partial\theta^{k_m})^T} + \frac{\partial^2\log f(\theta^{k_m}|m)}{\partial\theta^{k_m}(\partial\theta^{k_m})^T}, \tag{30}$$

is differentiable for θ^{k_m}. Since $\frac{n_{i,j}^{(m)}}{n} \to q^*_{s_j(m)}\theta_{i,j}^{k_m^*}$, $a.s.$ is satisfied from Corollary 1, we have

$$\frac{1}{n}\frac{\partial^2\sum_{i,j}n_{i,j}^{(m)}\log\theta_{i,j}^{k_m}}{\partial\theta^{k_m}(\partial\theta^{k_m})^T} = \frac{\partial^2\sum_{i,j}\frac{n_{i,j}^{(m)}}{n}\log\theta_{i,j}^{k_m}}{\partial\theta^{k_m}(\partial\theta^{k_m})^T} \to \frac{\partial^2\sum_{i,j}q^*_{s_j(m)}\theta_{i,j}^{k_m^*}\log\theta_{i,j}^{k_m}}{\partial\theta^{k_m}(\partial\theta^{k_m})^T}, \quad a.s. \tag{31}$$

Then, because $\frac{\partial^2\log f(\theta^{k_m}|m)}{\partial\theta^{k_m}(\partial\theta^{k_m})^T}$ is differentiable from Condition 1, iii), and does not depend on n, $\frac{1}{n}\frac{\partial^2 L_n(\theta^{k_m}|m)}{\partial\theta^{k_m}(\partial\theta^{k_m})^T}$

[†]Moreover, on (c.1), (c.2), and (c.3), when we regard θ^{k_m} as the random variable drawn from $f(\theta^{k_m}|x^n,m)$, the density of $\phi_n^{k_m} = \Sigma_n^{-1/2}(\theta^{k_m} - \tilde{\theta}^{k_m})$, $f_\phi(\phi_n^k|x^n,m)$, satisfies

$$\int_R f_\phi(\phi_n^{k_m}|x^n,m)d\phi_n^{k_m} \to \int_R (2\pi)^{-k_m/2}\exp\left\{-\frac{1}{2}(\phi^{k_m})^T\phi^{k_m}\right\}d\phi^{k_m}, \tag{25}$$

where R is an arbitrary rectangle and $f_\phi(\phi_n^{k_m}|x^n,m)$ is given by

$$f_\phi(\phi_n^{k_m}|x^n,m) = (\det\Sigma_n)^{1/2}\,f(\theta^{k_m}|x^n,m). \tag{26}$$

This means the asymptotic normality of the posterior density.

[††]Of course, almost sure satisfaction of (c.1) to (c.3) implicitly depends not only on the prior but also the probability measure on $\mathcal{X}^n$.

converges to $I(\theta^{k_m}|m)$ almost surely from (8). We have therefore

$$L_n''(\theta^{k_m}|m)\left\{L_n''(\tilde{\theta}^{k_m}|m)\right\}^{-1} \to I(\theta^{k_m}|m)\left\{I(\tilde{\theta}^{k_m}|m)\right\}^{-1}, \quad a.s. \tag{32}$$

On the other hand, from the continuity of $I(\theta^{k_m}|m)$, there exists $\delta > 0$ such that

$$I - A(\epsilon) \leqq I(\theta^{k_m}|m)\left\{I(\tilde{\theta}^{k_m}|m)\right\}^{-1} \leqq I + A(\epsilon), \tag{33}$$

for $\forall\theta^{k_m} \in B_\delta(\tilde{\theta}^{k_m})$ if $\tilde{\theta}^{k_m}$ is fixed. Then, from $\tilde{\theta}^{k_m} \to \hat{\theta}^{k_m} \to \theta^{k_m^*}$, $a.s.$, (c.2) is almost surely satisfied when $n \to \infty$.

Moreover, since $\lim_{n\to\infty} n_{i,j}^{(m)} \to \infty$, $a.s.$, the smallest eigenvalue of $\frac{\partial^2 L_n(\theta^{k_m}|m)}{\partial\theta^{k_m}(\partial\theta^{k_m})^T}$ tends to ∞ almost surely. This means that (c.1) is also almost surely satisfied.

On the other hand, $L_n(\theta^{k_m}|m) - L_n(\tilde{\theta}^{k_m}|m)$ is given by

$$L_n(\theta^{k_m}|m) - L_n(\tilde{\theta}^{k_m}|m) = \sum_{i,j} n_{i,j}^{(m)} \log\frac{\theta_{i,j}^{k_m}}{\tilde{\theta}_{i,j}^{k_m}} + \log\frac{f(\theta^{k_m}|m)}{f(\tilde{\theta}^{k_m}|m)}, \tag{34}$$

where $\tilde{\theta}^{k_m} = (\tilde{\theta}_{0,0}^{k_m}, \tilde{\theta}_{1,0}, \cdots, \tilde{\theta}_{\beta-1,|S(m)|-1}^{k_m})^T$. From Taylor expansion, we have

$$L_n(\theta^{k_m}|m) - L_n(\tilde{\theta}^{k_m}|m) = (\theta^{k_m} - \tilde{\theta}^{k_m})^T \frac{\partial L_n(\theta^{k_m}|m)}{\partial\theta^{k_m}}\bigg|_{\theta^{k_m}=\theta^{k_m}+} \tag{35}$$

where $\theta^{k_m}+$ is a point on the line segment between θ^{k_m} and $\tilde{\theta}^{k_m}$. Here, $\frac{1}{n}\frac{\partial L_n(\theta^{k_m}|m)}{\partial\theta^{k_m}}\Big|_{\theta^{k_m}=\theta^{k_m}+}$ satisfies

$$\begin{aligned}
&\frac{1}{n}\frac{\partial L_n(\theta^{k_m}|m)}{\partial\theta^{k_m}}\bigg|_{\theta^{k_m}=\theta^{k_m}+} \\
&= \frac{1}{n}\frac{\partial\sum_{i,j} n_{i,j}^{(m)}\log\theta_{i,j}^{k_m}}{\partial\theta^{k_m}} + \frac{1}{n}\frac{\partial\log f(\theta^{k_m}|m)}{\partial\theta^{k_m}}\bigg|_{\theta^{k_m}=\theta^{k_m}+} \\
&\to \frac{\partial\sum_{i,j} q_{s_j(m)}^* \theta_{i,j}^{k_m^*}\log\theta_{i,j}^{k_m}}{\partial\theta^{k_m}}\bigg|_{\theta^{k_m}=\theta^{k_m}+}, \quad a.s.
\end{aligned} \tag{36}$$

since $\frac{n_{i,j}^{(m)}}{n} \to q_{s_j(m)}^* \theta_{i,j}^{k_m^*}$, $a.s.$ is satisfied and $f(\theta^{k_m}|m)$ is three times continuously differentiable. The last term of (36) is reduced to 0 only when $\theta^{k_m}+ = \theta^{k_m^*}$. On the other hand, $\theta^{k_m^*} \in B_\delta(\tilde{\theta}^{k_m})$ is almost surely satisfied when $n \to \infty$. Therefore, there exists $c_\delta > 0$ such that $c_\delta < \left\|\frac{1}{n}\frac{\partial L_n(\theta^{k_m}|m)}{\partial\theta^{k_m}}\Big|_{\theta^{k_m}=\theta^{k_m}+}\right\|$, $a.s.$ when $n \to \infty$ for $\forall\theta^{k_m} \notin B_\delta(\tilde{\theta}^{k_m})$. We have therefore

$$\begin{aligned}
&L_n(\theta^{k_m}|m) - L_n(\tilde{\theta}^{k_m}|m) \\
&\leq \left\|\theta^{k_m} - \tilde{\theta}^{k_m}\right\| \left\|L_n'(\theta^{k_m}+|m)\right\| \cos\psi \\
&< -nc_\delta\left\|\theta^{k_m} - \tilde{\theta}^{k_m}\right\|, \quad a.s.
\end{aligned} \tag{37}$$

when $n \to \infty$[†]. Here, ψ is an angle between $\theta^{k_m} - \tilde{\theta}^{k_m}$ and $L_n'(\theta^{k_m}+|m)$.

On the other hand, there exists $c_{\bar{\psi}}$ such that $\bar{\psi} < c_{\bar{\psi}}$, $a.s.$ when $n \to \infty$, where $\bar{\psi}$ is the largest eigenvalue of $\frac{1}{n}\Sigma_n^{-1}$ because $\frac{1}{n}\Sigma_n^{-1} \to I(\tilde{\theta}^{k_m}|m)$, $a.s.$ and $\tilde{\theta}^{k_m} \to \hat{\theta}^{k_m} \to \theta^{k_m^*}$, $a.s.$ Therefore, there exists $c_{\delta,\psi}$ such that

$$\begin{aligned}
&L_n(\theta^{k_m}|m) - L_n(\tilde{\theta}^{k_m}|m) \\
&< -c_{\delta,\psi}\left\{(\theta^{k_m} - \tilde{\theta}^{k_m})^T\Sigma_n^{-1}(\theta^{k_m} - \tilde{\theta}^{k_m})\right\}^{1/2}, \\
&\qquad a.s.
\end{aligned} \tag{38}$$

when $n \to \infty$, and hence (c.3) of Lemma 2 is satisfied almost surely.

Since (c.1), (c.2), and (c.3) are satisfied and $\lim_{n\to\infty}\frac{1}{n}\Sigma_n^{-1} = I(\tilde{\theta}^{k_m}|m)$, $a.s.$, we have the following lemma from Lemma 2.

Lemma 3: For $\forall m \in \mathcal{M}$, we have

$$f(\tilde{\theta}^{k_m}|x^n, m) = \left(\frac{n}{2\pi}\right)^{k_m/2}\sqrt{\det I(\tilde{\theta}^{k_m}|m)} + o(n^{k_m/2}), \quad a.s. \tag{39}$$

□

In [1] and the above lemma, the asymptotic normality around $\tilde{\theta}^{k_m}$ has been proved. Moreover we can show that the similar asymptotic is satisfied for $\hat{\theta}^{k_m}$ instead of $\tilde{\theta}^{k_m}$ from the same discussion as Lemma 3 and [1]. The result using $\hat{\theta}^{k_m}$ is useful for the analysis of the expected code length because the property of the asymptotic normality of the maximum likelihood estimator $\hat{\theta}^{k_m}$ is well known and can be used.

Lemma 4: For $\forall m \in \mathcal{M}$, we have

$$f(\hat{\theta}^{k_m}|x^n, m) = \left(\frac{n}{2\pi}\right)^{k_m/2}\sqrt{\det I(\hat{\theta}^{k_m}|m)} + o(n^{k_m/2}), \quad a.s. \tag{40}$$

Proof: See Appendix A. □

3.3 Codelengths for Individual Sequences

From Lemma 4, we have the following theorem.

Theorem 1: On Condition 1, for $\forall m \in \mathcal{M}$ we have

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

[†]Similar discussion is appeared in [1], pp.293–294.

$$+\log\frac{\sqrt{\det I(\hat{\theta}^{k_m}|m)}}{f(\hat{\theta}^{k_m}|m)}+o(1),\ a.s. \tag{41}$$

Proof: From the Bayes rule, we have

$$-\log P(x^n|m)=-\log\frac{P(x^n|m,\hat{\theta}^{k_m})f(\hat{\theta}^{k_m}|m)}{f(\hat{\theta}^{k_m}|x^n,m)}. \tag{42}$$

From (40), we have

$$\log f(\hat{\theta}^{k_m}|x^n,m)=\frac{k_m}{2}\log\frac{n}{2\pi}+\log\sqrt{\det I(\hat{\theta}^{k_m}|m)}+\log\Big(1+o(1)\Big),\quad a.s. \tag{43}$$

Similarly, we have the first term of r.h.s. of (41) from (39). □

Theorem 1 also shows the codelength of the Bayes code for an FSMX model, $|\mathcal{M}|=1$. When an FSMX model m is assumed although another FSMX model is true, its codelength is given by (41).

On the other hand, we have the following lemma.

Lemma 5: On Condition 1, for $\forall m\neq m^*$, $m\in\mathcal{M}$, we have

$$\frac{P(x^n|m)}{P(x^n|m^*)}=o^+(1),\quad a.s. \tag{44}$$

where $o^+(1)$ is the term such as $\lim_{n\to\infty}o^+(1)=+0$, $a.s.$

Proof: See Appendix B. □

This lemma shows the strong consistency of the model selection by maximization of the posterior probability, whose asymptotic formula for exponential family, Bayesian information criterion (BIC), was proposed by Schwarz [11].

From Theorem 1 and Lemma 5, we have the following theorem.

Theorem 2: On Condition 1, we have

$$\begin{aligned}L^{m,\theta^{k_m}}_{Bayes}(x^n)&=-\log P(x^n|m^*,\tilde{\theta}^{k_{m^*}})-\log P(m^*)\\&\quad+\frac{k_{m^*}}{2}\log\frac{n}{2\pi}+\log\frac{\sqrt{\det I(\tilde{\theta}^{k_{m^*}}|m^*)}}{f(\tilde{\theta}^{k_{m^*}}|m^*)}+o(1),\quad a.s.\\&=-\log P(x^n|m^*,\hat{\theta}^{k_{m^*}})-\log P(m^*)\\&\quad+\frac{k_{m^*}}{2}\log\frac{n}{2\pi}+\log\frac{\sqrt{\det I(\hat{\theta}^{k_{m^*}}|m^*)}}{f(\hat{\theta}^{k_{m^*}}|m^*)}+o(1),\quad a.s.\end{aligned} \tag{45}$$

Proof: From Lemma 5 and $\sum_m P(m)=1$, we have

$$\begin{aligned}&-\log\sum_m\frac{P(x^n|m)P(m)}{P(x^n|m^*)P(m^*)}\\&=-\log\left\{1+\sum_{m\neq m^*}\frac{P(x^n|m)P(m)}{P(x^n|m^*)P(m^*)}\right\}\\&=-\log\{1+o^+(1)\},\quad a.s.\end{aligned} \tag{46}$$

We have therefore

$$\begin{aligned}&L^{m,\theta^{k_m}}_{Bayes}(x^n)\\&=-\log P(x^n|m^*)-\log P(m^*)-o^+(1),\quad a.s.\end{aligned} \tag{47}$$

Applying Theorem 1, the proof is completed. □

3.4 Expected Codelength

We derive the asymptotic expected codelength of the Bayes code. At first, we show the following lemma.

Lemma 6: On Condition 1, for $m^*\in\mathcal{M}$, we have

$$E^*\log\frac{P(x^n|m^*,\hat{\theta}^{k_{m^*}})}{P(x^n|m^*,\theta^{k^*_{m^*}})}=\frac{k_{m^*}}{2}+o(1), \tag{48}$$

where E^* represents the expectation by $P(\cdot|m^*,\theta^{k^*_{m^*}})$.

Proof: See Appendix C. □

From this lemma, we have the following theorem.

Theorem 3: On the condition 1, we have

$$\begin{aligned}&E^*L^{m,\theta^{k_m}}_{Bayes}(x^n)\\&=-E^*\log P(x^n|m^*,\theta^{k^*_{m^*}})-\log P(m^*)\\&\quad+\frac{k_{m^*}}{2}\log\frac{n}{2\pi e}+\log\frac{\sqrt{\det I(\theta^{k^*_{m^*}}|m^*)}}{f(\theta^{k^*_{m^*}}|m^*)}+o(1),\end{aligned} \tag{49}$$

where $I(\theta^{k^*_{m^*}}|m^*)$ corresponds to the Fisher information matrix at $\theta^{k^*_{m^*}}$.

Proof: In Theorem 2, the asymptotic equation, (45), is based on almost surely convergence. Then, we can apply the bounded convergence theorem. Therefore, we have (49) from Lemma 4 and continuity of $I(\theta^{k_m}|m)$ and $f(\theta^{k_m}|m)$. □

4. Discussion

We have analyzed the asymptotic codelengths of the Bayes codes for the individual sequences in 3.3. This codelengths are also called stochastic complexity, which are taken to represent the information in the sequences on a given model class. From Theorem 1, we can see that the asymptotic codelength of the Bayes code for the parametric model class coincides with that of the maximum likelihood code [10].

In 3.4, we have shown the asymptotic expected codelength of the Bayes code. In [2], Clarke and Barron discussed the expected codelength of the Bayes code for the i.i.d. parametric model class. We have generalized this result for the FSMX model class which is not i.i.d. source and belongs to the partial nested model classes.

5. Conclusion

We have discussed the asymptotic codelengths of the Bayes code for the FSMX model class. The analysis of the Bayes risk and the minimax risk [3] will be future work. The results in this paper suggests that the minimax redundancy may be achieved by the Jeffreys prior for the parameters θ^{k_m} and the uniform prior for FSMX models m.

Acknowledgements

The authors would like to acknowledge all of the member of Hirasawa Lab. and Matsushima Lab. for their helpful suggestions and discussions to this work. This research was supported in part by the Ministry of Education under Grant-Aids 07558168 for Scientific Research and Waseda University under Grant 97A-158 for Special Research Projects.

References

[1] J.M. Bernardo and A.F.M. Smith, "Bayesian Theory," John Wiley & Sons, 1994.
[2] B.S. Clarke and A.R. Barron, "Information—Theoretic asymptotics of bayes methods," IEEE Trans. Inf. Theory, vol.36, no.3, pp.453–471, 1990.
[3] B.S. Clarke and A.R. Barron, "Jeffreys' Prior is asymptotically least favorable under entropy risk," J. Statistical Planning and Inference, 41, pp.37–60, 1994.
[4] C.-F. Chen, "On asymptotic normality of limiting density function with bayesian implications," J.R. Statist. Soc. B, vol.47, no.3, pp.540–546, 1985.
[5] W. Feller, "An Introduction to Probability and Its Applications," vol.I and II, John Wiley & Sons, New York, 1957, 1966.
[6] T. Kawabata, "Bayes codes and context tree weighting method," IEICE Technical Report, IT93-121, 1994.
[7] T. Matsushima, H. Inazumi, and S. Hirasawa, "A class of distortionless codes designed by bayes decision theory," IEEE Trans. Inf. Theory, vol.37, no.5, pp.1288–1293, 1991.
[8] T. Matsushima and S. Hirasawa, "A bayes coding algorithm for Markov models," IEICE Technical Report, IT95-1, 1995.
[9] J. Rissanen, "Stochastic complexity," J. Roy, Statist., Soc. B, vol.49, pp.223–265, 1987.
[10] J. Rissanen, "Fisher information and stochastic complexity," IEEE Trans. Inf. Theory, vol.42, no.1, pp.40–47, 1996.
[11] C. Schwarz, "Estimating the dimension of a model," Ann. Statist., vol.6, pp.461–464, 1978.
[12] J. Suzuki, "Some notes on universal noiseless coding," IEICE Trans. Fundamentals, vol.E78-A, no.12, 1995.
[13] J. Takeuchi, "Characterization of the bayes estimator and the MDL estimator for exponential families," IEEE Trans. Inf. Theory, vol.43, no.4, pp.1165–1174, 1996.

Appendix A: The Proof of Lemma 4

This proof is similar to the discussion in [1], pp.285–297.

We define $L_n(\theta^{k_m}|m) = \log f(\theta^{k_m}|x^n, m)$, $B_\delta(\hat{\theta}^{k_m}) = \left\{\theta^{k_m} \in \Theta^{k_m} \middle| \|\theta^{k_m} - \hat{\theta}^{k_m}\| < \delta\right\}$, and

$$\hat{\Sigma}_n = -\left(L_n''(\hat{\theta}^{k_m}|m)\right)^{-1} \tag{A·1}$$

Then, for $\forall\theta^{k_m}$, a Taylor expansion establishes that

$$\begin{aligned} &f(\theta^{k_m}|x^n, m) \\ &= f(\hat{\theta}^{k_m}|x^n, m)\exp\left\{(\theta^{k_m} - \hat{\theta}^{k_m})^T L'(\hat{\theta}^{k_m}|m)\right\} \\ &\exp\left\{-\frac{1}{2}(\theta^{k_m} - \hat{\theta}^{k_m})^T(I + R_n)L''(\hat{\theta}^{k_m}|m)(\theta^{k_m} - \hat{\theta}^{k_m})\right\}, \end{aligned} \tag{A·2}$$

where R_n is given by

$$R_n = L_n''(\theta^{k_m+}|m)\{L_n''(\hat{\theta}^{k_m}|m)\}^{-1} - I, \tag{A·3}$$

for some θ^{k_m+} lying between θ^{k_m} and $\hat{\theta}^{k_m}$. Here, from the definition of $\hat{\theta}^{k_m}$, $L_n'(\hat{\theta}^{k_m}|m)$ is given by

$$L_n'(\hat{\theta}^{k_m}|m) = \left.\frac{\partial \log f(\theta^{k_m}|m)}{\partial\theta^{k_m}}\right|_{\theta^{k_m}=\hat{\theta}^{k_m}}, \tag{A·4}$$

On the other hand, from

$$\frac{1}{n}L_n''(\theta^{k_m}|m) \to I(\theta^{k_m}|m), \ a.s. \tag{A·5}$$

we have

$$\begin{aligned} &L_n''(\theta^{k_m+}|m)\{L_n''(\hat{\theta}^{k_m}|m)\}^{-1} \\ &\quad \to I(\theta^{k_m+}|m)\{I(\hat{\theta}^{k_m}|m)\}^{-1}, \quad a.s. \end{aligned} \tag{A·6}$$

From the smoothness of $I(\theta^{k_m}|m)$ and $\hat{\theta}^{k_m} \to \theta^{k_m^*}$, $a.s.$, there exists $\delta > 0$ such that

$$\begin{aligned} I - A(\epsilon) &\leqq L_n''(\theta^{k_m+}|m)\{L_n''(\hat{\theta}^{k_m}|m)\}^{-1} \\ &\leqq I + A(\epsilon), \quad a.s., \end{aligned} \tag{A·7}$$

is satisfied for $\forall\theta^{k_m} \in B_\delta(\hat{\theta}^{k_m})$ and $\forall\epsilon > 0$ when $n \to \infty$. It follows that

$$P_n(\delta) = \int_{B_\delta(\hat{\theta}^{k_m})} f(\theta^{k_m}|x^n, m)d\theta^{k_m}, \tag{A·8}$$

is almost surely bounded above by

$$\begin{aligned} &f(\hat{\theta}^{k_m}|x^n, m)\exp\{\bar{c}_f\delta\}|\det\hat{\Sigma}_n|^{1/2}|I - A(\epsilon)|^{-1/2} \\ &\cdot\int_{|z|<s_n}\exp\left\{-\frac{1}{2}z^Tz\right\}dz, \end{aligned} \tag{A·9}$$

and below by

$$\begin{aligned} &f(\hat{\theta}^{k_m}|x^n, m)\exp\{-\bar{c}_f\delta\}|\det\hat{\Sigma}_n|^{1/2}|I + A(\epsilon)|^{-1/2} \\ &\cdot\int_{|z|<t_n}\exp\left\{-\frac{1}{2}z^Tz\right\}dz, \end{aligned} \tag{A·10}$$

when $n \to \infty$, where $\bar{c}_f$ is the largest absolute value of elements in $\frac{\partial \log f(\theta^{k_m}|m)}{\partial\theta^{k_m}}$, $s_n = \delta(1 - \overline{a(\epsilon)})^{1/2}/\underline{l_n}^{1/2}$ and $t_n = \delta(1 - \underline{a(\epsilon)})^{1/2}/\overline{l_n}^{1/2}$ with $\overline{a(\epsilon)}$ ($\underline{a(\epsilon)}$) and $\overline{l_n}$ ($\underline{l_n}$) the largest (smallest) eigenvalues of $A(\epsilon)$ and $\hat{\Sigma}_n$, respectively.

From

$$\frac{1}{n}\left(\hat{\Sigma}_n\right)^{-1} \to I(\theta^{k_m^*}|m), \quad a.s. \tag{A·11}$$

we have $\overline{l_n} \to 0$, $a.s.$ and $\underline{l_n} \to 0$, $a.s.$, which lead to $s_n \to \infty$, $a.s.$ and $t_n \to \infty$, $a.s.$. Then, when $n \to \infty$,

$$\begin{aligned}
&|I - A(\epsilon)|^{1/2} \exp\{-\bar{c}_f \delta\} \lim_{n\to\infty} P_n(\delta) \\
&\leqq \lim_{n\to\infty} f_\xi(\hat{\xi}^{k_m}|x^n, m) \\
&\quad \cdot \left\{ -\det\left(\frac{1}{n} L''(\hat{\theta}^{k_m}|m)\right)\right\}^{-1/2} (2\pi)^{k_m/2} \\
&\leqq |I + A(\epsilon)|^{1/2} \exp\{\bar{c}_f \delta\} \lim_{n\to\infty} P_n(\delta),
\end{aligned} \tag{A·12}$$

is satisfied almost surely. Here, ξ^{k_m} is given by $\xi^{k_m} = \sqrt{n}(\theta^{k_m} - \hat{\theta}^{k_m})$.

If $\lim_{n\to\infty} P_n(\delta) = 1$, $a.s.$ for $\forall \epsilon > 0$ is satisfied, then

$$\lim_{n\to\infty} f_\xi(\hat{\xi}^{k_m}|x^n, m) \to \frac{\{\det I(\hat{\theta}^{k_m}|m)\}^{1/2}}{(2\pi)^{k_m/2}}, \quad a.s. \tag{A·13}$$

is clearly satisfied. Then, at last, we shall show $\lim_{n\to\infty} P_n(\delta) = 1$, $a.s.$ for $\forall \epsilon > 0$.

From a simple Taylor expansion, we have

$$\begin{aligned}
&L_n(\theta^{k_m}|m) - L_n(\hat{\theta}^{k_m}|m) \\
&= (\theta^{k_m} - \hat{\theta}^{k_m})^T \left.\frac{\partial L_n(\theta^{k_m}|m)}{\partial \theta^{k_m}}\right|_{\theta^{k_m} = \theta^{k_m +}}.
\end{aligned} \tag{A·14}$$

where $\theta^{k_m +}$ is a point lying between θ^{k_m} and $\hat{\theta}^{k_m}$. From the same discussion as (36)~(37), there exists a constant $c_\delta > 0$ such that $c_\delta < \left\|\frac{1}{n} L_n'(\theta^{k_m +}|m)\right\|$, $a.s.$, for $\theta^{k_m} \in B_\delta(\hat{\theta}^{k_m})$ when $n \to \infty$, and we have

$$\begin{aligned}
&L_n(\theta^{k_m}|m) - L_n(\hat{\theta}^{k_m}|m) \\
&= \left(\theta^{k_m} - \hat{\theta}^{k_m}\right)^T L_n'(\theta^{k_m +}|m) \\
&\leqq -n c_\delta \left\|\theta^{k_m} - \hat{\theta}^{k_m}\right\| \\
&\leqq -c_{\delta,\phi} \left\{(\theta^{k_m} - \hat{\theta}^{k_m})^T L''(\hat{\theta}^{k_m}|m)(\theta^{k_m} - \hat{\theta}^{k_m})\right\}^{1/2}
\end{aligned} \quad a.s. \tag{A·15}$$

for $\exists c_{\delta,\phi} > 0$ and $\theta^{k_m} \in B_\delta(\hat{\theta}^{k_m})$. Using this inequation, we have

$$\begin{aligned}
&\int_{\Theta^{k_m} - B_\delta(\hat{\theta}^k)} f(\theta^{k_m}|x^n, m) d\theta^{k_m} \\
&\leqq f(\hat{\theta}^{k_m}|x^n, m) \left(\det L''(\hat{\theta}^{k_m}|m)\right)^{-1/2} \\
&\quad \cdot \int_{|z| > \delta/\overline{l_n}^{1/2}} \exp\{-c_{\delta,\phi}(z^T z)^{1/2}\} dz
\end{aligned} \tag{A·16}$$

From (A·12) and $\overline{l_n} \to 0$, $a.s.$, we have

$$\int_{\Theta^{k_m} - B_\delta(\hat{\theta}^k)} f(\theta^{k_m}|x^n, m) d\theta^{k_m} \to 0, \quad a.s. \tag{A·17}$$

We have therefore $\lim_{n\to\infty} P_n(\delta) = 1$ $a.s.$, and (A·13) is shown. Then the proof is completed.

Appendix B: The Proof of Lemma 5

$$\begin{aligned}
&\log \frac{P(x^n|m)}{P(x^n|m^*)} \\
&= \log \frac{P(x^n|m, \hat{\theta}^{k_m})}{P(x^n|m^*, \hat{\theta}^{k_{m^*}})} - \left(\frac{k_m}{2} - \frac{k_{m^*}}{2}\right) \log \frac{n}{2\pi} \\
&\quad - \frac{f(\hat{\theta}^{k_{m^*}}|m^*)\sqrt{\det I(\hat{\theta}^{k_m}|m)}}{f(\hat{\theta}^{k_m}|m)\sqrt{\det I(\hat{\theta}^{k_{m^*}}|m^*)}} + o(1), \quad a.s.
\end{aligned} \tag{A·18}$$

for $\forall m \in \mathcal{M}$.

The equations $f(\hat{\theta}^{k_m}|m) = O(1)$, $a.s.$ and $\det I(\hat{\theta}^{k_m}|m) = O(1)$, $a.s.$ are satisfied, since $f(\theta^{k_m}|m)$ and $I(\theta^{k_m}|m)$ are differentiable and $\hat{\theta}^{k_m} \to \theta^{k_m^*}$, $a.s.$ Therefore from (A·18), if the equation

$$\log \frac{P(x^n|m, \hat{\theta}^{k_m})}{P(x^n|m^*, \hat{\theta}^{k_{m^*}})} - \frac{k_m - k_{m^*}}{2} \log n \to -\infty, \quad a.s. \tag{A·19}$$

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

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

$$-\log P(x^n|m, \theta^{k_m^*}) = -\sum_{j=0}^{|S(m)|-1} \sum_{i=0}^{\beta} n_{i,j}^{(m)} \log \theta_{i,j}^{k_m^*}, \tag{A·20}$$

and

$$\frac{n_{i,j}^{(m)}}{n} \to q_{s_j(m)}^* \theta_{i,j}^{k_m^*} + O\left(\left(\frac{\log\log n}{n}\right)^{1/2}\right), \quad a.s. \tag{A·21}$$

$-\log P(x^n|m, \theta^{k_m^*}) = O(n)$, $a.s.$ is clearly satisfied. We have therefore

$$\begin{aligned}
&-\log P(x^n|m, \theta^{k_m^*}) \\
&= -\log P(x^n|m, \hat{\theta}^{k_m}) \\
&\quad - \frac{1}{2}\left(\hat{\theta}^{k_m} - \theta^{k_m^*}\right)^T \left.\frac{\partial^2 \log P(x^n|m, \theta^{k_m})}{\partial \theta^{k_m} (\partial \theta^{k_m})^T}\right|_{\theta^{k_m} = \hat{\theta}^{k_m}} \\
&\quad \cdot \left(\hat{\theta}^{k_m} - \theta^{k_m^*}\right) + O\left(n \left\|\hat{\theta}^{k_m} - \theta^{k_m^*}\right\|^3\right), \quad a.s.
\end{aligned} \tag{A·22}$$

from Taylor expansion.

Since

$$\begin{aligned}
&-\frac{1}{n} \left.\frac{\partial^2 \log P(x^n|m, \theta^{k_m})}{\partial \theta^{k_m} (\partial \theta^{k_m})^T}\right|_{\theta^{k_m} = \hat{\theta}^{k_m}} \\
&= -\frac{1}{n} \left.\frac{\partial^2 \sum_{j=0}^{|S(m)|-1} \sum_{i=0}^{\beta} n_{i,j}^{(m)} \log \theta_{i,j}^{k_m}}{\partial \theta^{k_m} (\partial \theta^{k_m})^T}\right|_{\theta^{k_m} = \hat{\theta}^{k_m}}
\end{aligned}$$

$$\to -\left.\frac{\partial^2 \sum_{j=0}^{|S(m)|-1}\sum_{i=0}^{\beta}\theta_{i,j}^{k_m^*} q_{s_j(m)}^* \log\theta_{i,j}^{k_m}}{\partial\theta^{k_m T}(\partial\theta^{k_m})^T}\right|_{\theta^{k_m}=\hat{\theta}^{k_m}}$$
$$+O\left(\left(\frac{\log\log n}{n}\right)^{1/2}\right), \quad a.s.$$
$$= I(\hat{\theta}^{k_m}|m)+O\left(\left(\frac{\log\log n}{n}\right)^{1/2}\right), \quad a.s. \quad (A\cdot 23)$$

and Corollary 1 are satisfied, we have

$$-\frac{1}{2}\left(\hat{\theta}^{k_m}-\theta^{k_m^*}\right)^T \left.\frac{\partial^2 \log P(x^n|m,\theta^{k_m})}{\partial\theta^{k_m}(\partial\theta^{k_m})^T}\right|_{\theta^{k_m}=\hat{\theta}^{k_m}} \cdot\left(\hat{\theta}^{k_m}-\theta^{k_m^*}\right)$$
$$=-\frac{n}{2}\left(\hat{\theta}^{k_m}-\theta^{k_m^*}\right)^T I(\hat{\theta}^{k_m}|m)\left(\hat{\theta}^{k_m}-\theta^{k_m^*}\right)$$
$$+O\left(\frac{(\log\log n)^{3/2}}{n^{1/2}}\right), \quad a.s. \quad (A\cdot 24)$$

On the other hand, since

$$\left\|\hat{\theta}^{k_m}-\theta^{k_m^*}\right\| = O\left(\left(\frac{\log\log n}{n}\right)^{1/2}\right), \quad a.s. \quad (A\cdot 25)$$

we have

$$n\left\|\hat{\theta}^{k_m}-\theta^{k_m^*}\right\|^3 = O\left(\frac{(\log\log n)^{3/2}}{n^{1/2}}\right), \quad a.s. \quad (A\cdot 26)$$

From (A· 22), we have

$$\log\frac{P(x^n|m,\hat{\theta}^{k_m})}{P(x^n|m,\theta^{k_m^*})}$$
$$=\frac{n}{2}\left(\hat{\theta}^{k_m}-\theta^{k_m^*}\right)^T I(\hat{\theta}^{k_m}|m)\left(\hat{\theta}^{k_m}-\theta^{k_m^*}\right)$$
$$+O\left(\frac{(\log\log n)^{3/2}}{n^{1/2}}\right), \quad a.s. \quad (A\cdot 27)$$

And because $\sqrt{n}\|\hat{\theta}^{k_m}-\theta^{k_m^*}\| = O((\log\log n)^{1/2})$, $a.s.$ and $I(\hat{\theta}^{k_m}|m)\to I(\theta^{k_m^*}|m)$, $a.s.$, we have

$$\frac{n}{2}\left(\hat{\theta}^{k_m}-\theta^{k_m^*}\right)^T I(\hat{\theta}^{k_m}|m)\left(\hat{\theta}^{k_m}-\theta^{k_m^*}\right)$$
$$=O(\log\log n), \quad a.s. \quad (A\cdot 28)$$

We have therefore

$$\log\frac{P(x^n|m,\hat{\theta}^{k_m})}{P(x^n|m^*,\hat{\theta}^{k_{m^*}})}-\frac{k_m-k_{m^*}}{2}\log n$$
$$=\log\frac{P(x^n|m,\theta^{k_m^*})}{P(x^n|m^*,\theta^{k_{m^*}^*})}-\frac{k_m-k_{m^*}}{2}\log n$$
$$+o(\log n), \quad a.s. \quad (A\cdot 29)$$

At first, we consider the case $P^*(x^n)\notin\mathcal{H}^{k_m}$. We obtain

$$\forall\theta^{k_m}\in\Theta^{k_m}, \quad \log\frac{P(x^n|m,\theta^{k_m})}{P(x^n|m,\theta^{k_{m^*}^*})}=-Cn+o(n), \quad (A\cdot 30)$$

from (13) and (A· 21), where C is some positive constant.

We have therefore

$$\log\frac{P(x^n|m,\hat{\theta}^{k_m})}{P(x^n|m^*,\hat{\theta}^{k_{m^*}})}-\frac{k_m-k_{m^*}}{2}\log n$$
$$=-Cn-\frac{k_m-k_{m^*}}{2}\log n+o(n)$$
$$\to-\infty, \quad a.s. \quad (A\cdot 31)$$

for $\forall m\in\mathcal{M}$, $P^*(x^n)\notin\mathcal{H}^{k_m}$.

When $P^*(x^n)\in\mathcal{H}^{k_m}$, the equation

$$P(x^n|m,\theta^{k_m^*})=P(x^n|m^*,\theta^{k_{m^*}^*}), \quad (A\cdot 32)$$

is satisfied from the definition. We have therefore

$$\log\frac{P(x^n|m,\hat{\theta}^{k_m})}{P(x^n|m^*,\hat{\theta}^{k_{m^*}})}-\frac{k_m-k_{m^*}}{2}\log n$$
$$=\frac{k_{m^*}-k_m}{2}\log n+o(\log n)$$
$$\to-\infty, \quad a.s. \quad (A\cdot 33)$$

for $\forall m\in\mathcal{M}$, $P^*(x^n)\in\mathcal{H}^{k_m}$. Then (A· 19) is satisfied for $\forall k_m\neq k_{m^*}$, and the proof is completed. □

Appendix C: The Proof of Lemma 6

From Corollary 1 and (A· 24), since $I(\theta^{k_{m^*}}|m^*)$ is continuous function, we have

$$\log\frac{P(x^n|m^*,\hat{\theta}^{k_{m^*}})}{P(x^n|m^*,\theta^{k_{m^*}^*})}$$
$$=-\frac{n}{2}\left(\hat{\theta}^{k_{m^*}}-\theta^{k_{m^*}^*}\right)^T I(\theta^{k_{m^*}^*}|m^*)\left(\hat{\theta}^{k_{m^*}}-\theta^{k_{m^*}^*}\right)$$
$$+O\left(\frac{(\log\log n)^{3/2}}{n^{1/2}}\right), \quad a.s. \quad (A\cdot 34)$$

Above asymptotic equation is almost surely satisfied, we can apply the dominated convergence theorem and acquire the expectation:

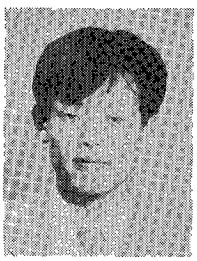
$$E^*\log\frac{P(x^n|m^*,\hat{\theta}^{k_{m^*}})}{P(x^n|m^*,\theta^{k_{m^*}^*})}$$
$$=E^*\frac{n}{2}\left(\hat{\theta}^{k_{m^*}}-\theta^{k_{m^*}^*}\right)^T I(\theta^{k_{m^*}^*}|m^*)\left(\hat{\theta}^{k_{m^*}}-\theta^{k_{m^*}^*}\right)$$
$$+O\left(\frac{(\log\log n)^{3/2}}{n^{1/2}}\right), \quad a.s. \quad (A\cdot 35)$$

Since the distribution of $\xi^{k_{m^*}}=\sqrt{n}\left(\hat{\theta}^{k_{m^*}}-\theta^{k_{m^*}^*}\right)$ converges to $N(0,I^{-1}(\theta^{k_{m^*}^*}|m^*))$ in law from the

property of the FSMX model class [5], we have

$$E^* \frac{1}{2}\sqrt{n}\left(\hat{\theta}^{k_{m^*}} - \theta^{k^*_{m^*}}\right)^T I(\theta^{k^*_{m^*}}|m^*)\sqrt{n}\left(\hat{\theta}^{k_{m^*}} - \theta^{k^*_{m^*}}\right) = \frac{k_{m^*}}{2} + o(1). \quad \text{(A·36)}$$

Then the proof is completed. □



Masayuki Gotoh was born in Tokyo, Japan, on Jan. 1, 1969. He received his B.E. and M.E. degrees from Musashi Institute of Technology, Tokyo, Japan, in 1992 and 1994, respectively. He is now a student of doctoral program and a research associate in Industrial and Management Systems Engineering at Waseda University, Tokyo, Japan. His research interests include intelligent control, machine learning theory, model selection, and Bayesian statistics. He is a member of the Society of Information Theory and Its Applications, the Japan Industrial Management Association, and the Japan Society for Artificial Intelligence.

Toshiyasu Matsushima was born in Tokyo, Japan, on Nov. 26, 1955. He received the B.E. degree, M.E. degree and Dr.E degree in Industrial Engineering and Management from Waseda University, Tokyo, Japan, in 1978, 1980 and 1991, respectively. From 1980 to 1986, he was with the Nippon Electric Corporation, Kanagawa, Japan. From 1986 to 1992, he was a lecture to the Department of Management Information, Yokohama College of Commerce. Since 1993, he has been a associate professor of School of Science and Engineering, Waseda University, Tokyo, Japan. His research interests are information theory and its application, statistics and artificial intelligence. He is a member of the Society of Information Theory and Its Applications, the Japan Society for Quality Control, the Japan Industrial Management Association, the Japan Society for Artificial Intelligence, and IEEE.

Shigeichi Hirasawa was born in Kobe, Japan, on Oct. 2, 1938. He received the B.S. degree in mathematics and the B.E. degree in electrical communication engineering from Waseda University, Tokyo, Japan, 1961 and 1963, respectively, and the Dr.E. degree in electrical communication engineering from Osaka University, Osaka, Japan, in 1975. From 1963 to 1981, he was with the Mitsubishi Electric corporation, Hyogo, Japan. Since 1981, he has been a professor of School of Science and Engineering, Waseda University, Tokyo, Japan. In 1979, he was a Visiting Researcher in the Computer Science Department at the University of California, Los Angels, CA. He was a Visiting Researcher at the Hungarian Academy of Science, Hungary, in 1985, and at the University of Trieste, Italy, in 1986. From 1987 to 1989, he was the Chairman of Technical Group on Information Theory of IEICE. He received the 1993 Achievement Award, and the 1993 Kobayashi-Memorial Achievement Award from IEICE. In 1996, he was the President of the Society of Information Theory and Its Applications (Soc. of ITA). His research interests are information theory and its applications, and information processing systems. He is a member of Soc. of ITA, the Operations a Research Society of Japan, the Information Processing Society of Japan, the Japan Industrial Management Association, IEEE, and Informs.