2001 年情報論的学習理論ワークショップ
2001 Workshop on Information-Based Induction Sciences (IBIS2001)
Tokyo, Japan, July 30 - August 1, 2001.

Properties of Jeffreys Mixture for Markov Sources

Jun-ichi Takeuchi * Tsutomu Kawabata † Andrew R. Barron ‡

Abstract: We discuss the properties of Jeffreys mixture for general FSMX model (a certain class of Markov sources [11]). First, we show that modified Jeffreys mixture asymptotically achieves the minimax coding regret [7], where we do not put any restriction on data sequences at all. This is extension of results in [13, 15]. Then, we give an approximation formula for the prediction probability of Jeffreys mixture for FSMX models (review of the result in [10, 19]). By this formula, it is revealed that the prediction probability by Jeffreys mixture for the first order Markov chain with alphabet $\{0, 1\}$ is not of the form $(k + \alpha)/(n + \beta)$ (n is data size, k is number of occurrences of '1'). Moreover, we evaluate by simulation the regret of our approximation formula for the first order Markov chain and show that the prediction strategy using our approximation formula gives smaller coding regret than the one using Laplace estimator.

1 Introduction

We discuss the properties of Jeffreys mixture for general FSMX model (FSMX model is a certain class of Markov sources [11]) in the problem of prediction and universal coding.

First, we show that slightly modified Jeffreys mixture asymptotically achieves the minimax regret [7], where we do not put any restriction on data sequences at all. This is an extension of results in [13, 15] and provide evaluation of stochastic complexity defined by Rissanen [6]. The regret is defined as the difference of the loss incurred and the loss of an ideal coding or prediction strategy for each sequence. Xie & Barron treat the all sequences with finite alphabet in multinominal Bernoulli model and show that the modified Jeffreys mixtures, one of which is shown to be asymptotically minimax in terms of redundancy (expected regret) [14], achieve the minimax regret asymptotically [15]. Takeuchi & Barron [9] show that the similar mixtures are minimax for (i.i.d.) exponential families and certain near exponential families that permit dependence, but their bounds are valid only for the restricted set of sequences so that the MLE locate in a certain compact set interior to the parameter space (an exception is one-dimensional exponential family). Our result is a generalization of [15] to Markov models and that of [9] to the set of all sequences. (Strictly speaking, the first order Markov chain with alphabet size 2 is treated in [13]). Concerning Markov models, Atteson [1] obtains pointwise bound on redundancy of the Jeffreys mixture. Also, Gotoh et. al. [3] give an upper bound on the regret, which holds almost surely.

As well known, the Jeffreys mixture for Bernoulli model induces Laplace estimator. While Laplace estimator is in very simple form, Jeffreys mixture for FSMX model is not, even when the model is first order Markov chain. Hence, we give an approximation formula for the prediction probability of Jeffreys mixture for the FSMX models (review of the result in [10, 19]). This is a certain extension of the approximation formulas of the Bayes estimator for (i.i.d.) exponential families, which Takeuchi showed [8]. We can see the behavior of Jeffreys mixture by this formula. In particular, it is revealed that the prediction probability by Jeffreys mixture for the first order Markov chain with alphabet $\{0, 1\}$ is not of the form $(k + \alpha)/(n + \beta)$ (n is data seize, k is number of occurrences of '1'). Moreover, we evaluate by simulation the regret of our approximation formula for the first order Markov chain and show that the prediction strategy using our approximation formula gives smaller coding regret than the one using Laplace estimator.

2 Preliminaries

We review the definition of FSMX model [11]. Let alphabet $\mathcal{X}$ be a set $\{0, 1, ..., k\}$. Define $\mathcal{X}' \stackrel{\text{def}}{=} \mathcal{X} \setminus \{0\}$. Let T be a subset of $\mathcal{X}^* \stackrel{\text{def}}{=} \{\lambda\} \cup \mathcal{X} \cup \mathcal{X}^2 \cup ...$, where λ denotes a null sequence. Assume that for all $s \in T$, any postfix of s belongs to T (e.g., the postfixes of x_1x_2 are x_1x_2, x_2 and λ). Such set T is called a context tree. For a context tree T, define ∂T as

$$\partial T \stackrel{\text{def}}{=} \{xs : x \in \mathcal{X}, s \in T\} \cup \{\lambda\} \setminus T.$$

*RWC Theoretical Foundation NEC Laboratory c/o Internet System Res. Labs., NEC Corp. 4-1-1 Miyazaki, Miyamae, Kawasaki, Kanagawa 216-8555, Japan.

†Dept. of Information & Communications Engineering, University of Electro-Communications. 1-5-1 Choufugaoka, Chofushi, Tokyo 182-8585, Japan.

‡Dept. of Statistics, Yale University. P.O. Box 208290, New Haven, CT 06520, USA.

It can be shown that ∂T is a complete postfix set of $\mathcal{X}$, i.e. no element of ∂T is a postfix of another element and their length satisfies Kraft inequality with equality. For example, let $T_e = \{\lambda, 0\}$ ($\mathcal{X} = \{0, 1\}$), then we have $\partial T_e = \{1, 10, 00\}$, which is a complete postfix set. For $s \in \mathcal{X}^*$, we let $\tau(s)$ denote an element of ∂T which matches a postfix of s. Let $d \overset{\text{def}}{=} \max_{s \in \partial T} |s|$ ($|s|$ is length of s). When $|s| \geq d$, $\tau(s)$ exists and is unique. For example, in the case of T_e in the above, we have $\tau(11) = 1$, $\tau(10) = 10$, $\tau(100) = 00$, $\tau(101) = 1$, $\tau(000) = 00$ and $\tau(001) = 1$.

Let η_s^x denote the probability that x is generated at the context s. We let η_s denote a k-dimensional vector $(\eta_s^1, \eta_s^2, ..., \eta_s^k)$ and η a $k|\partial T|$-dimensional vector $(\eta_{s_1}^1, , ..., \eta_{s_1}^k, \eta_{s_2}^1, , ..., \eta_{s_2}^k, , ..., \eta_{s_{|\partial T|}}^1, , ..., \eta_{s_{|\partial T|}}^k)$. Here, we let $\eta_s^0 \overset{\text{def}}{=} 1 - \sum_{x \in \mathcal{X}'} \eta_s^x$. Define the parameter space as $H_s \overset{\text{def}}{=} \{\eta_s : \forall x \in \mathcal{X}', \ \eta_s^x \geq 0 \wedge \sum_{x \in \mathcal{X}'} \eta_s^x \leq 1\}$ and $H = H(T) \overset{\text{def}}{=} \prod_{s \in \partial T} H_s$.

We let x_m^n denote a sequence $x_m x_{m+1} ... x_n$ ($m \leq n$) and x^n a sequence x_1^n. We assume that we have an initial sequence $x^0{}_{-d+1}$ in advance. We denote the initial context $\tau(x^0{}_{-d+1})$ by s_0. Let n_s^x denote the number of occurrences of x at the context s in the sequence x^n and define $n_s \overset{\text{def}}{=} \sum_{x \in \mathcal{X}} n_s^x$. We denote the probability mass function for the sequence x^n by $p(x^n | \eta, x^0_{-d+1})$. We define a class of Markov sources as

$$S(T) \overset{\text{def}}{=} \{p(\cdot|\eta, \cdot) : \eta \in H(T)\}.$$

Then, $S(T)$ is referred to as a tree model [11, 4]. Further, we assume that $\tau(sx)$ for any $s \in \partial T$ and any $x \in \mathcal{X}$ is determined, even if $|sx| < d$, i.e. τ defines a state transition function. When this condition is satisfied, a tree model is referred to as an FSMX model [11, 4]. Note that $S(T_e)$ is an example of FSMX model. In this paper we treats an FSMX model only.

We usually omit x^0_{-d+1} from $p(x^n|\eta, x^0_{-d+1})$ and simply denote as $p(x^n|\eta)$. Then, we have

$$\log p(x^n|\eta) = \sum_{s \in \partial T, \, x \in \mathcal{X}} n_s^x \log \eta_s^x, \tag{1}$$

where we let 'log' denote natural logarithm. Let $\hat{\eta} = \hat{\eta}(x^n)$ denote MLE (maximum likelihood estimate) of η given x^n, then we have $\hat{\eta}_s^x = n_s^x / n_s$.

Let W_n be a subset of $\mathcal{X}^{n+d}$. Let $\mathcal{P}(W_n)$ denote the set of all probability mass functions on W_n. The maximum regret of $q \in \mathcal{P}(W_n)$ with respect to a family of probability mass function $S = \{p(\cdot|\eta) : \eta \in H\}$ and W_n (denoted by $\bar{r}(q, W_n)$) is defined as

$$\bar{r}(q, W_n) \overset{\text{def}}{=} \sup_{x^n_{-d+1} \in W_n} (\log \frac{1}{q(x^n | x^0_{-d+1})} - \log \frac{1}{p(x^n|\hat{\eta})}).$$

The minimax regret with respect to a family of probability mass function S and a set of the sequences W_n (denoted by $\bar{r}(W_n)$) is defined as

$$\bar{r}(W_n) \overset{\text{def}}{=} \inf_{q \in \mathcal{P}(W_n)} \bar{r}(q, W_n).$$

The regret $\bar{r}(q, W_n)$ is the difference between the code length based on q and the minimum of the codelength $\log(1/p(x^n|\eta))$ achieved by distributions in the family. Also, $\log(1/q(x^n|x^n_{-d+1})) - \log(1/p(x^n|\eta))$ is the sum of the incremental regret of prediction $\log(1/q(x_{i+1}|x^i_{-d+1}))$ $- \log(1/p(x_{i+1}|x^i, \eta))$.

The maximin regret for set W_n (denoted by $\underline{r}_n(W_n)$) is defined as

$$\underline{r}_n(W_n) \overset{\text{def}}{=} \sup_{q \in \mathcal{P}(W_n)} \inf_{r \in \mathcal{P}(\mathcal{X}^{n+d})} E_q \log \frac{p(x^n|\hat{\eta})}{r(x^n|x^0_{-d+1})}.$$

It is known that $\bar{r}_n(W_n) = \underline{r}_n(W_n)$ holds [7, 15].

We define Jeffreys prior as $w_J(\eta) \overset{\text{def}}{=} \sqrt{\det J(\eta)} / C_J$, where J is Fisher information matrix with respect to η and $C_J \overset{\text{def}}{=} \int_H \sqrt{\det J(\eta)} d\eta$.

Now we introduce Fisher information and empirical Fisher information. Empirical Fisher information is the Hessian of $-(1/n) \log p(x^n|\eta)$. We denote its component with respect to η_s^x and η_t^y, by $\hat{J}_{(s,x)(t,y)}(x^n, \eta)$. Then, we can derive from (1),

$$\hat{J}_{(s,x)(t,y)}(x^n, \eta) = \delta_{st} \hat{p}_s (\frac{\delta_{xy} \hat{p}_s^x}{(\eta_s^x)^2} + \frac{\hat{p_s^0}}{(\eta_s^0)^2}), \tag{2}$$

where we let $\hat{p}_s \overset{\text{def}}{=} n_s / n$ and $\hat{p}_s^x \overset{\text{def}}{=} n_s^x / n_s$. Also δ_{xy} and δ_{st} denote Kronecker's delta. Fisher information is denoted as

$$\begin{aligned} J_{(s,x)(t,y)}(\eta) &= \lim_{n \to \infty} E_\eta \hat{J}_{(s,x)(t,y)}(x^n, \eta) \\ &= \delta_{st} p_c(s|\eta) (\frac{\delta_{xy}}{\eta_s^x} + \frac{1}{\eta_s^0}), \end{aligned} \tag{3}$$

where $p_c(s|\eta)$ is the stationary probability of the state s determined by $p(\cdot|\eta)$, E_η denote the expectation with respect to $p(\cdot|\eta)$.

Let $D_{(\alpha)}(\eta_s) \overset{\text{def}}{=} \prod_{x \in \mathcal{X}} (\eta_s^x)^{-(1-\alpha)}$ (Dirichlet function), then we have

$$w_J(\eta) = \frac{\prod_{s \in \partial T} p_c(s|\eta)^{k/2} D_{(1/2)}(\eta_s)}{C_J}.$$

We define another prior density as

$$w_{(\alpha)}(\eta) \overset{\text{def}}{=} \frac{\prod_{s \in \partial T} D_{(\alpha)}(\eta_s)}{(C_{(\alpha)})^{|\partial T|}},$$

where $C_{(\alpha)} \overset{\text{def}}{=} \int D_{(\alpha)}(\eta_s) d\eta_s$. Note that $w_{(\alpha)}(\eta) / w_J(\eta)$ $\to \infty$ holds as η approaches the boundaries of H, if $0 < \alpha < 1/2$ holds.

3 Results

3.1 Minimax Regret

We define a modified Jeffreys prior as

$$w_n \overset{\text{def}}{=} (1 - n^{-b}) w_J + n^{-b} w_{(\alpha)},$$

where $0 < \alpha < 1/2$ is assumed. We let m_n denote the mixture with respect to w_n. We can show the following.

Theorem 1 *There exists $b > 0$ such that the following holds.*

$$\bar{r}(m_n, \mathcal{X}^n) = \frac{|\partial T|k}{2} \log \frac{n}{2\pi} + \log C_J + o(1) \qquad (4)$$

where $o(1)$ converges to 0 as n goes to infinity.

Outline of the proof is given in Section 5.

Remark 1: For the lower bound, it is possible to show directly that the lower bound on the maximin regret matches the above upper bound. Further, we can show that the Jeffreys mixture without modification is asymptotically maximin [18]. Also, Rissanen's result [6] contains equivalent lower bound (we cannot apply it for the upper bound).

Remark 2: This is a generalization of the result about the first order Markov chains with alphabet size 2 in [13]. The proof is not its straightforward extension.

Remark 3: The similar upper bound on minimax redundancy (expected regret) for Markov chains is obtained in [1], but it is not uniform but pointwise. In [6, 9, 18] also, upper bounds of the same form on regret are obtained for more general models, but they hold under the restriction on the sequences that MLE is located in a compact set included in the interior of the parameter space (an exception is one-dimensional exponential family in [9]). Under that condition, we can show that Jeffreys mixture is minimax, however when MLE goes near the boundaries, we need the help from the second term of w_n, which has higher density near the boundaries than Jeffreys prior.

Let $p(x_{n+1}|x^n, \eta) \stackrel{\text{def}}{=} p(x^{n+1}|\eta)/p(x^n|\eta)$. The prediction probability based on Bayes mixture with prior w is given by $p_w(x_{n+1}|x^n) = \int p(x_{n+1}|x^n, \eta) w(\eta|x^n) d\eta$ $= \int \eta^{x_{n+1}}_{s(x^n_{-d+1})} w(\eta|a^n) d\eta$, where $w(\eta|x^n)$ denotes the posterior density of η. For the prior w_n we must use it as $m_n(x_t|x^{t-1})$ $(t \leq n)$. This implies that we have to know the length of the sequence in advance, in order to use m_n for the prediction, while the Laplace estimator doesn't depend on the total lenght of the sequennce. That is, Theorem 1 means that m_n, which is obtained by slightly modifying the Jeffreys mixture, is the best prediction strategy, when we can know the lenght of the sequence in advance.

3.2 Approximation Formula

As mentioned above, the Jeffreys mixture is nearly best strategy. However, it is hard to calculate in general. We give a way to approximate it. For that purpose, we introduce the another parameter θ than η. Note that $p(x^n|\eta)$ is rewritten as follows:

$$p(x^n|\eta) = \prod_{s \in \partial T} \exp(n_s(\sum_{x \in \mathcal{X}'} \theta^x_s \hat{\eta}^x_s - \psi(\theta_s))),$$

where we let $\theta^x_s = \log(\eta^x_s/\eta^0_s)$ and $\psi(\theta_s) = -\log \eta^0_s = \log(1 + \sum_{x \in \mathcal{X}'} \exp \theta^x_s)$. Note that $(\partial/\partial\theta^x_s)\psi(\theta_s) = \eta^x_s$ holds. We let $\Theta_s \stackrel{\text{def}}{=} \{\theta_s(\eta_s) : \eta_s \in H^\circ_s\}$, then $\Theta_s = \Re^{|\mathcal{X}'|}$ holds. Let $\Theta(T) \stackrel{\text{def}}{=} \prod_{s \in \partial T} \Theta_s = \Re^{|\partial T| \cdot |\mathcal{X}'|}$. It is known that the map $\eta_s \mapsto \theta_s$ on H°_s is one to one and analytic (see [2]).

The following theorem gives an efficient method to approximate the Jeffreys mixture, where we let $\tilde{\eta}^x_s \stackrel{\text{def}}{=} \int p(x|x^n, \eta) w_J(\eta|x^n) d\eta$ and $\bar{\eta}^x_s \stackrel{\text{def}}{=} (n^x_s + 0.5)/(n_s + (k+1)/2)$.

Theorem 2 *Let K be a compact set included in the interior of H and n_0 be an arbitrary natural number. Then,*

$$\begin{aligned} \tilde{\eta}^x_s \quad &- \quad \bar{\eta}^x_s - \frac{1}{n_s + (k+1)/2} \left. \frac{\partial \log(w_J(\eta)/w_{(1/2)}(\eta))}{\partial \theta^x_s} \right|_{\eta = \hat{\eta}} \\ &= \quad O(\frac{\sqrt{\log n}}{n\sqrt{n}}) \end{aligned}$$

holds, uniformly for all sequences $x_{-d+1}...x_1x_2...$ such that $\hat{\eta} \in K$ holds for all $n \geq n_0$.

We omit the proof. See [10, 19].

Remark: Actually, Theorem 2 can be extended to the mixture with general prior density under appropriate conditions. Then, if we replace w_J with Dirichlet prior with $\alpha = 1/2$ $(w_{(1/2)})$, the third term of the left hand side of (5) vanishes. In that case, Theorem 2 yields $\tilde{\eta} \sim \bar{\eta}$ and coincides with the well known fact that $\tilde{\eta} = \bar{\eta}$ holds when the prior is $w_{(1/2)}$.

From Theorem 2, we can obtain more explicit approximation formula. Note that

$$\log \frac{w_J(\eta)}{w_{(1/2)}(\eta)} = \frac{k}{2} \sum_{t \in \partial T} \log p_c(t|\eta) + C_1$$

and

$$\frac{\partial}{\partial \theta^x_s} = \sum_{y \in \mathcal{X}'} \frac{\partial \eta^y_s}{\partial \theta^x_s} \frac{\partial}{\partial \eta^y_s} = \sum_{y \in \mathcal{X}'} \eta^y_s (\delta_{xy} - \eta^x_s) \frac{\partial}{\partial \eta^y_s}.$$

Hence, we have

$$\begin{aligned} \tilde{\eta}^x_s \quad &- \quad \bar{\eta}^x_s - \sum_{y \in \mathcal{X}',\, t \in \partial T} \frac{k \hat{\eta}^y_s (\delta_{xy} - \hat{\eta}^x_s)}{2n_s + k + 1} \left. \frac{\partial \log p_c(t|\eta)}{\partial \eta^y_s} \right|_{\eta = \hat{\eta}} \\ &= \quad O(\frac{\sqrt{\log n}}{n\sqrt{n}}). \end{aligned}$$

It is known that the Jeffreys mixture for the i.i.d. case is asymptotically maximin in terms of regret and induces Laplace estimator, which is used in CONTEXT[5] and CTW method[12]. We compare our approximation formula for w_J with the prediction strategy with Laplace estimator.

Let $\mathcal{X} = \{0, 1\}$ and $\partial T = \{0, 1\}$. Suppose that x_n equals 0. By Theorem 2, the approximation of the Jeffreys mixture for this case is given by

$$\tilde{\eta}^1_0 \sim \frac{n^1_0 + 0.5}{n_0 + 1} + \frac{1}{n_0 + 1} \Big(\frac{1 - \hat{\eta}^1_0}{2} - \frac{\hat{\eta}^1_0 (1 - \hat{\eta}^1_0)}{\hat{\eta}^1_0 + \hat{\eta}^0_1} \Big), \qquad (5)$$

Note that this depends on not only $\tilde{\eta}_0^1$ but $\hat{\eta}_1^0$ and that the difference between $\tilde{\eta}_0^1$ and the Laplace estimator is of order $\Omega(1/n_0)$ (negation of $o(1/n_0)$).

4 Simulation

We evaluated the regret of the strategy with $\tilde{\eta}$ (denoted as q_A) and the strategy with Laplace estimator (denoted as q_L). Actually, we evaluated a quantity $\tilde{r}(q, x_{-d+1}^n) \stackrel{\mathrm{def}}{=} \log(p(x^n|\hat{\eta})/q(x^n|x_{-d+1}^0)) - \log(n/2\pi)$. Since value of $\tilde{r}(q, x_{-d+1}^n)$ depends on x_{-d+1}^n, we generated a number of x_{-d+1}^n using pseudo random number with respect to $p(x^n|\eta)$. Table 1 shows a part of our results, where the first row and the first column indicate the values of η_0^1 and η_1^0 respectively, which we generate x_{-d+1}^n with respect to. Each cell except them indicates average value of $\tilde{r}(q, x_{-d+1}^n)$. The upper one is the average of $\tilde{r}(q_A, x^n)$ and the lower one is that of $\tilde{r}(q_L, x^n)$, where the number of trials is 100. Note that $p(x^n|\eta)$ is i.i.d., when $\eta_0^1 = 1 - \eta_1^0$ holds. One might expect that q_L would perform better than q_A for these cases, but it didn't.

	0.1	0.3	0.5	0.7	0.9
0.1 q_A	0.99	1.16	1.30	1.35	1.39
q_L	1.60	1.45	1.30	1.18	1.09
0.3		1.12	1.17	1.26	1.30
		1.60	1.56	1.50	1.45
0.5			1.22	1.29	1.35
			1.60	1.58	1.55
0.7				1.34	1.40
				1.60	1.59
0.9					1.44
					1.60

Table 1: Average of regret ($n = 10000$)

Note that $1.296 \le \log C_J \le 1.305$ holds. The regret of q_A is close to this lower bound.

5 Outline of Proof of Theorem 1

The main tool for the proof is the Laplace integration, by which we have the following asymptotics:

$$\frac{\int p(x^n|\eta)w_J(\eta)}{p(x^n|\hat{\eta})} \sim \frac{\sqrt{\det(J(\hat{\eta}))}}{C_J\sqrt{\det(\hat{J}(x^n, \hat{\eta}))}} \frac{(2\pi)^{k|\partial T|/2}}{n^{k|\partial T|/2}}, \quad (6)$$

where $\hat{J}(x^n, \eta)$ is the empirical Fisher information. When the model is an exponential family, $\hat{J}(x^n, \hat{\eta}) = J(\hat{\eta})$ holds. Then for exponential families, our task is to control the the convergence of (6) only, but the FSMX model is not exponential type. However, it is known that FSMX model converges to an exponential family, when the sample size goes to infinity (see [16]). Hence, for the FSMX model, the empirical Fisher information converges to the Fisher information:

$$|\hat{J}(x^n, \hat{\eta}) - J(\hat{\eta})| \to 0 \quad (7)$$

If we restrict the sequence x^n so that MLE $\hat{\eta}(x^n)$ belongs to a compact set K included in the interior of H, then we can prove that the convergence of (6) and (7) is uniform for those sequences, but it is impossible without such restriction. However, we can moderate it, i.e. we can prove the uniform convergence for the sequence belonging to $H^{(2n^{-a})}(T)$. Here, we let

$$H^{(\epsilon)} = H^{(\epsilon)}(T) \stackrel{\mathrm{def}}{=} \{\eta : \forall s \in \partial T, \ \forall x \in \mathcal{X}, \ \eta_s^x > \epsilon\}$$

and a is a certain small positive number. For the sequences which do not belong to $H^{(2n^{-a})}$, we obtain smaller regret than the minimax value with the help from the second term of w_n, $n^{-b}w_{(\alpha)}(\eta)$. For the proof, we use Lemma 4 of [15].

In particular, the problem about (7) makes the proof about the interior region harder (this problem does not exist for multinomial Bernoulli model [15] and one-dimensional exponential family [9]). Hence in this paper, we concentrate on the problem about the ratio of the determinant of empirical Fisher information to that of Fisher information. Comparing (2) with (3), we have only to evaluate the ratio $\hat{p}_s/p_c(s|\hat{\eta})$ ($s \in \partial T$), for which we can show the following Lemma.

Lemma 1 *Let $r = d(|\partial T| - 1)$. There exists a constant $C_1 > 0$, such that the following holds.*

$$\forall n \ge 1, \forall \epsilon : n\epsilon^d > 2, \forall x^n, s_0 : \hat{\eta} \in H^{(2\epsilon)},$$
$$n_s > n\epsilon^d - 1 \quad (8)$$
$$\textit{and} \quad |\log \frac{\hat{p}_s}{p_c(s|\hat{\eta})}| < \frac{C_1}{n\epsilon^{r+d}}. \quad (9)$$

Remark: When the model is the first order Markov chain with alphabet $\{0, 1\}$, the proposition which corresponds to Lemma 1 is easy to show, since the explicit forms of $p_c(s|\eta)$ are very simple.

Let $\epsilon_n = n^{-a}$, where we assume $(1 - ad)/2 > a$ and

$$0 < a < (1/2)\min\{\frac{1}{2r+d}, \frac{1}{2+d}\}.$$

When $\hat{\eta} \in H^{(2\epsilon_n)}$ holds, we have $n\epsilon_n^{r+d} > n^{1-(r+d)/(2r+d)}$ $\to \infty$ as n goes to infinity. Hence, we have $\hat{p}_s/p_c(s|\hat{\eta})$ $\to 1$ uniformly for $x_{-d+1}^n : \hat{\eta} \in H^{(2n^{-a})}$. Hence, Lemma 1 implies that empirical Fisher information converges to Fisher information, uniformly for $x_{-d+1}^n : \hat{\eta} \in H^{(2n^{-a})}$.

In the remaining of this section, we describe the proof of Lemma 1. Recall that n_s^x denotes the number of generation of x at the state s ($\in \partial T$) in the sequence $x^n = x_1x_2...x_n$. Further, for every $t, u \in \partial T$, we let n_t^u denote the number of transition from the state $t \in \partial T$ to the state $u \in \partial T$ in the sequence $x^n = x_1x_2...x_n$. The equation $n_s^x = n_s^{\tau(sx)}$ holds.

Similarly, we let $\forall x \in \mathcal{X}$, $\forall s \in \partial T$, $\eta_s^{\tau(sx)} \overset{\text{def}}{=} \eta_s^x$. We define

$$D_s \overset{\text{def}}{=} \{\tau(sx) : x \in \mathcal{X}\}.$$

The set D_s consists of the states which can be reached by one transition from the state s. Then for all $s \in \partial T$ and for all $s' \in D_s$, $\eta_s^{s'}$ is defined. We can define the state transition probability matrix as

$$\Pi_{s's} = \begin{cases} \eta_s^{s'}, & \text{when } s' \in D_s, \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

First, we will show the following.

Proposition 1 *If $\eta_t^x > \epsilon$ holds for each $t \in \partial T$ and each $x \in \mathcal{X}$, then $p_c(t|\eta) > \epsilon^d$ holds.*

Proof: Note that the probabilities $p_c(t|\eta)$ $(t \in \partial T)$ satisfy the following linear equations:

$$p_c(t|\eta) = \sum_{t' \in \partial T} \Pi_{tt'} p_c(t'|\eta). \tag{11}$$

For each $t \in \partial T$ and $x \in \mathcal{X}$, we have $\eta_t^x > \epsilon$ by the assumption. Let $y^d = t$. For each pair $(t, t') \in (\partial T)^2$, $\tau_d(t'y^d) = t$ holds. Further, for all $y^d \in \mathcal{X}^d$ and all $i \in \{0, 1, ..., d-1\}$, $\eta_{\tau_d(t'y^i)}^{y_{i+1}} > \epsilon$ holds. This implies that each element of Π^d is larger than ϵ^d, i.e. each $p_c(t|\eta)$ is larger than ϵ^d. **Q.E.D.**

Lemma 2 *There exist certain positive number C_1, such that*

$$\forall s \in \partial T,\ \forall t \in \partial T,\ \forall x \in \mathcal{X}',\ \forall \epsilon > 0,\ \forall \eta \in H^{(\epsilon)}, \quad \left|\frac{\partial \log p_c(s|\eta)}{\partial \eta_t^x}\right| \le \frac{C_1}{\epsilon^r}$$

holds, where $r = d(|\partial T| - 1)$.

Outline of Proof: We renumber the state as $\partial T = \{s_1, s_2, ... s_q\}$, where we let $q = |\partial T|$. Define a matrix A as

$$A_{ij} = {\Pi^d}_{s_i s_j}$$

and a vector $\boldsymbol{\mu}$ as $\boldsymbol{\mu} = (p_c(s_1|\eta), ..., p_c(s_q|\eta))^T$. (Note that when $\eta \in H^{(\epsilon)}$, $A_{ij} > \epsilon^d$ holds.) Then, we have

$$(I - A)\boldsymbol{\mu} = 0$$

(I denotes the unit matrix). Let Δ_{ij} be the cofactor with respect to the (i, j)-component of $I - A$. By some manipulation, we have

$$p_c(s_i|\eta) = \mu_i = \frac{\Delta_{1i}}{\sum_j \Delta_{1j}}.$$

Note that μ_i is a rational function of η. Further, we can prove the following:

$$\forall \eta \in H^{(\epsilon)},\ \forall s \in \partial T,\ \ \Delta_{1i} \ge \epsilon^{d(q-1)}.$$

We have

$$\frac{\partial \log p_c(s|\eta)}{\partial \eta_t^x} = \frac{1}{\Delta_{1i}} \frac{\partial \Delta_{1i}}{\partial \eta_t^x} - \frac{1}{\sum_j \Delta_{1j}} \frac{\partial \sum_j \Delta_{1j}}{\partial \eta_t^x}.$$

Note that the derivative of Δ_{ij} is bounded upper by a certain constant. Therefore, we have

$$\forall t \in \partial T, \forall x \in \mathcal{X}', \forall \eta \in H^{(\epsilon)}, \left|\frac{\partial \log p_c(s|\eta)}{\partial \eta_t^x}\right| \le \frac{C}{\epsilon^{d(q-1)}}.$$

Q.E.D.

Now, we can prove Lemma 1.

Proof of Lemma 1: Let s_0 denote the state determined by x_{-d+1}^0 (initial state) and s_e the state $\tau(x^n)$. First, we treat a special case in which $s_0 = s_e$ holds. In this case, we have

$$\forall s \in \partial T, \sum_{t \in \partial T} n_s^t = \sum_{t \in \partial T} n_t^s, \tag{12}$$

since the number of all transition from the state s equals the number of all transition to the state s. Hence, we have

$$\sum_{t \in \partial T} \hat{\eta}_t^s \hat{p}_t = \sum_{t \in \partial T} \frac{n_t^s}{n_t} \frac{n_t}{n} = \sum_{t \in \partial T} \frac{n_t^s}{n} = \frac{n_s}{n} = \hat{p}_s.$$

This implies $\hat{p}_s = p_c(s|\hat{\eta})$.

When $s_0 \ne s_e$, let $x_{n+1}^{n+\alpha}$ be a minimum path from the state s_e to s_0 (α does not exceed d). By adding a sequence $x_{n+1}^{n+\alpha}$ to the sequence x^n, we have $\tau(x^{n+\alpha}) = s_0$. Let $\tilde{p}_s$ denote the relative frequency of the state s in $x^{n+\alpha}$ and $\tilde{\eta}_s^t$ the maximum likelihood estimate of η_s^t given $x^{n+\alpha}$. Then, we have $\tilde{p}_s = p_c(s|\tilde{\eta})$. Let ϕ_s^t denote the number of transition from the state s to the state t in the sequence $x_n....x_{n+\alpha}$. Here, $\phi_s^t = 0$ or 1, since x_{n+1}^α is a minimum pass from s_e to s_0. Let $\phi_s = \sum_t \phi_s^t$. We have $\tilde{\eta}_s^t = (n_s^t + \phi_s^t)/(n_s + \phi_s)$. Hence

$$\tilde{\eta}_s^t \ge \frac{n_s^t}{n_s + 1} = \frac{\hat{\eta}_s^t}{1 + 1/n_s} \ge \frac{\hat{\eta}_s^t}{2} > \epsilon,$$

where we use the fact that $n_s = \sum_t n_s^t \ge 1$ (if $n_s^t = 0$ holds for all $t \in \partial T$, then we have $n_s \le 1$ for all $s \in \partial T$).

By (a) of Proposition 1, we have $\tilde{p}_s = p_c(s|\tilde{\eta}) > \epsilon^d$. Hence, $n_s > n\epsilon^d - 1$ holds. This is (8).

Hence, we have

$$\tilde{\eta}_s^t \ge \frac{\hat{\eta}_s^t}{1 + 1/n_s} > \frac{\hat{\eta}_s^t}{1 + 1/(n\epsilon^d - 1)}.$$

Hence,

$$\hat{\eta}_s^t < \tilde{\eta}_s^t (1 + \frac{1}{n\epsilon^d - 1}) \le \tilde{\eta}_s^t + \frac{1}{n\epsilon^d - 1} < \tilde{\eta}_s^t + \frac{2}{n\epsilon^d} = \tilde{\eta}_s^t + \frac{2}{n\epsilon^d}.$$

Also, we have $\tilde{\eta}_s^t < n_s^t + 1/n_s = \hat{\eta}_s^t + 1/n$. Therefore, we have

$$|\tilde{\eta}_s^t - \hat{\eta}_s^t| < 2/n\epsilon^d.$$

By Taylor's theorem, we have

$$\log p_c(s|\tilde{\eta}) - \log p_c(s|\hat{\eta}) \\ = \sum_{t\in\partial T,\, x\in\mathcal{X}'} \left.\frac{\partial \log p_c(s|\eta)}{\partial \eta_t^x}\right|_{\eta=h} (\tilde{\eta}_t^x - \hat{\eta}_t^x),$$

where h is a point between $\tilde{\eta}$ and $\hat{\eta}$. Since $\tilde{\eta}, \hat{\eta} \in H^{(\epsilon)}(T)$, $h \in H^{(\epsilon)}(T)$ holds. Hence by Lemma 2, we have

$$\left| \left.\frac{\partial \log p_c(s|\eta)}{\partial \eta_t^x}\right|_{\eta=h} \right| \leq \frac{C}{\epsilon^r}.$$

Hence, we have

$$-\frac{2Cd|\partial T|}{n\epsilon^{r+d}} \leq \log \frac{p_c(s|\tilde{\eta})}{p_c(s|\hat{\eta})} \leq \frac{2Cd|\partial T|}{n\epsilon^{r+d}}. \qquad (13)$$

Since $\tilde{p}_s = (n_s + \phi_s)/(n+\alpha)$ holds, we have

$$\tilde{p}_s \geq \frac{n_s}{n+\alpha} = \frac{\hat{p}_s}{1+\alpha/n} \geq \frac{\hat{p}_s}{1+d/n}$$

and

$$\tilde{p}_s \leq \frac{n_s+1}{n} = \hat{p}_s + \frac{1}{n} = \hat{p}_s(1+\frac{1}{n\hat{p}_s}) = \hat{p}_s(1+\frac{1}{n_s}).$$

Hence,

$$\frac{1}{1+1/n_s} \leq \frac{\hat{p}_s}{\tilde{p}_s} \leq 1 + \frac{d}{n},$$

i.e.

$$-\frac{1}{n_s} \leq \log \frac{\hat{p}_s}{\tilde{p}_s} \leq \frac{d}{n}$$

holds. Together with (13) and $\tilde{p}_s = p_c(s|\tilde{\eta})$, we have

$$\begin{aligned} -\frac{C_2}{n\epsilon^{r+d}} &< -\frac{2C_1 d|\partial T|}{n\epsilon^{r+d}} - \frac{1}{n_s} \\ &< \log \frac{\hat{p}_s}{p_c(s|\hat{\eta})} \leq \frac{2C_1 d|\partial T|}{n\epsilon^{r+d}} + \frac{d}{n} < \frac{C_2}{n\epsilon^{r+d}}, \end{aligned}$$

where we use (8) and let $C_2 = 2\max\{2C_1 d|\partial T|,\ d\}$. **Q.E.D.**

6 Concluding Remark

We have determined the minimax regret for FSMX models without any restriction on the sequences, using modified Jeffreys mixtures. The obtained regret is of the same form as that for the multinominal Bernoulli models. Also, we have reviewed the approximation formula of Jeffreys mixture for FSMX models. Its computational cost is $O(|\partial T|^3)$. This is not low enough, if we try to use it in CTW method or CONTEXT algorithm.

References

[1] K. Atteson, "The asymptotic redundancy of Bayes rules for Markov chains", *IEEE trans. Inform. Theory,* Vol. 45, No. 6, pp. 2104-2109, 1999.

[2] L. Brown, *Fundamentals of statistical exponential families*, Institute of Mathematical Statistics, 1986.

[3] M. Gotoh, T. Matsushima, S. Hirasawa, "A Generalization of B. S. Clarke and A. R. Barron's Asymptotics of Bayes Codes for FSMX Sources", *IEICE trans. on Fundamentals*, Vol. E81-A, No. 10, p.2123–2132, 1998.

[4] T. Kawabata & F. Willems, "A context tree weighting algorithm with an incremental context set," *IEICE Trans. on Fundamentals*, vol. E83-A, No. 10, pp. 1898–1903, 2000.

[5] J. Rissanen, "A universal data compression system," *IEEE trans. Inform. Theory,* Vol. 29, No. 5, pp. 656-664, 1983.

[6] J. Rissanen, "Fisher information and stochastic complexity," *IEEE trans. Inform. Theory,* vol. 40, pp. 40-47, 1996.

[7] Yu M. Shtarkov, "Universal sequential coding of single messages," *Problems of Information Transmission*, vol. 23, pp. 3-17, July 1987.

[8] J. Takeuchi, "Characterization of the Bayes estimator and the MDL estimator for exponential families," *IEEE trans. Inform. Theory*, vol. 43, No. 4, pp. 1165-1174, July 1997.

[9] J. Takeuchi & A. R. Barron, "Asymptotically minimax regret by Bayes mixtures", *Proc. of IEEE International Symposium on Inform. Theory*, 1998.

[10] J. Takeuchi & T. Kawabata, "Approximation of Bayes code for Markov sources", *Proc. of 1995 IEEE International Symposium on Inform. Theory*, p. 391, 1995.

[11] M. J. Weinberger, J. Rissanen and M. Feder, "A universal finite memory source", *IEEE trans. Inform. Theory,* Vol. 41. No. 3, pp. 643-652, 1995.

[12] F. Willems, Y. Shtarkov and T. Tjalkens, "The context tree weighting method: basic properties," *IEEE trans. Inform. Theory,* Vol. 41. No. 3, pp. 653-664, 1995.

[13] Q. Xie, *Minimax coding and prediction,* Doctoral Dissertation, Dept. of Statistics, Yale University, 1997.

[14] Q. Xie & A. R. Barron, "Minimax redundancy for the class of memoryless sources", *IEEE trans. Inform. Theory,* vol. 43, no. 2, pp. 646-657, 1997.

[15] Q. Xie & A. R. Barron, "Asymptotic minimax regret for data compression, gambling and prediction", *IEEE trans. Inform. Theory,* vol. 46, no. 2, pp. 431-445, 2000.

[16] H. Itoh & S. Amari, Geometry of information sources (in Japanese), *Proc. of SITA88*, pp. 57–60, 1988.

[17] T. Kawabata, "Bayes codes and context tree weighting method (in Japanese)," *Technical Report of IEICE*, IT93-121. 1994-03. pp. 7-12, 1994.

[18] J. Takeuchi, "On minimax regret with respect to families of stationary stochastic processes (in Japanese)", *Porc. of IBIS2000*, pp. 63–68, 2000.

[19] J. Takeuchi & T. Kawabata, "On data compression algorithms by Bayes coding for Markov sources (in Japanese)," *Proc. of SITA94,* pp. 513–516, 1994.