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 an 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 multinomial 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] obtained 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 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.

2 Preliminaries

We review the definition of FSMX model [11]. Let alphabet \( \mathcal{X} \) be a set \( \{0,1, \ldots, k\} \). Define \( \mathcal{X}' = \mathcal{X} \setminus \{0\} \).

Let \( T \) be a subset of \( \mathcal{X}' \cup \mathcal{X} \cup \mathcal{X}' \cup \ldots \), where \( \lambda \) 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 \( x_1, x_2 \) and \( \lambda \)). Such set \( T \) is called a context tree. For a context tree \( T \), define \( \partial T \) as

\[
\partial T \overset{\text{def}}{=} \{ xs : x \in \mathcal{X}, s \in T \} \cup \{ \lambda \} \setminus T.
\]
It can be shown that $\partial T$ is a complete postfix set of $X$, i.e. no element of $\partial T$ is a postfix of another element and their length satisfies Kraft inequality with equality. For example, let $T_1 = \{A, 0\}$ ($X = \{0, 1\}$), then we have $\partial T_1 = \{1, 10, 00\}$, which is a complete postfix set. For $s \in X^*$, we let $\tau(s)$ denote an element of $\partial T$ which matches a postfix of $s$. Let $d \triangleq \max_{s \in \partial T} |s|$ ($|s|$ is length of $s$). When $|s| \geq d$, $\tau(s)$ exists and is unique. For example, in the case of $T_1$ in the above, we have $\tau(11) = 1$, $\tau(10) = 10$, $\tau(100) = 00$, $\tau(101) = 1$, $\tau(000) = 0$ and $\tau(001) = 1$.

Let $\eta^x$ denote the probability that $x$ is generated at the context $s$. We let $\eta_s$ denote a $k$-dimensional vector ($\eta^1_s, \eta^2_s, \ldots, \eta^k_s$) and $\eta$ a $k|\partial T|$-dimensional vector ($\eta^1_s, \ldots, \eta^k_s, \ldots, \eta^{|\partial T|}_s, \ldots, \eta^{|\partial T|}_s$). Here, we let $\eta^0_s \triangleq 1 - \sum_{x \in \mathcal{X}'} \eta^x_s$. Define the parameter space as

$$
H_s \triangleq \{ \eta : \forall x \in \mathcal{X}', \eta^x_s \geq 0 \land \sum_{x \in \mathcal{X}'} \eta^x_s \leq 1 \}.
$$

Let $\mathbb{E}_\eta$ denote the expected value of a function or random variable with respect to a family of prior distributions. Let $\mathbb{E}_0$ be the same expectation but with respect to a uniform distribution over $\{1, \ldots, m\}$.

Let $\eta^x, \eta^0_s$, $\eta^x$, and $\eta^0$ denote the probability mass function and density of the sequence $x^n$ and the distribution $\mathcal{P}(X)$ of the sequence $X$, respectively. We assume that there exists an initial state $\tau(x^0_{d+1})$ and let $n^x$ denote the number of occurrences of the state $x$ at the context $s$. We denote the probability mass function of the sequence $X^n$ by $p(x^n | \eta, x^0_{d+1})$.

We define a class of Markov sources as occurrences of context $(\eta^x, \eta^0_s)$ and a set of the sequences $\mathbb{P}_\eta$. We have

$$
\mathbb{P}_\eta = \sup_{q \in \mathcal{P}(\mathcal{X})} \mathbb{E}_\eta \log \frac{p(x^n | \eta, x^0_{d+1})}{r(x^n | \eta, x^0_{d+1})}.
$$

The regret $\bar{r}(q, W_n)$ is the difference between the code length based on $q$ and the minimum of the code length $\log(1/p(x^n | \eta))$ achieved by distributions in the family. Also, $\log(1/q(x^n | x^0_{d+1})) - \log(1/p(x^n | \eta))$ is the sum of the incremental regret of prediction $\log(1/q(x^n | x^0_{d+1})) - \log(1/p(x^n | \eta))$.

The maximin regret for set $W_n$ (denoted by $r_n(W_n)$) is defined as

$$
r_n(W_n) \triangleq \sup_{q \in \mathcal{P}(\mathcal{X})} \inf_{r \in \mathcal{P}(\mathcal{X})} \mathbb{E}_\eta \log \frac{p(x^n | \eta)}{r(x^n | \eta)}.
$$

It is known that $r_n(W_n) = r_n(W_n)$ holds [7, 15].

We define Jeffreys' prior as $w_J(\eta) \triangleq \sqrt{\det J(\eta)} / C_J$, where $J$ is Fisher information matrix with respect to $\eta$ and $C_J \triangleq \int \sqrt{\det J(\eta)} d\eta$.

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 $\eta^x$ and $\eta^y$, by $J_{(x, y)}(\eta)$. Then, we can derive from (1),

$$
J_{(x, y)}(\eta) = \frac{\partial^2 p_\eta}{\partial \eta^x \partial \eta^y} = \frac{\partial^2 p_\eta}{\partial \eta^x \partial \eta^y} \left( \frac{\delta_{xy}}{(\eta^x)^2} + \frac{\delta_{xy}}{(\eta^y)^2} \right),
$$

where we let $p_\eta \triangleq \eta_s / n_s$ and $\eta_s \triangleq \eta^x_s / n_s$ also. Also $\delta_{xy}$ and $\delta_x$ denote Kronecker's delta. Fisher information is denoted as

$$
J_{(x, y)}(\eta) = \lim_{n \to \infty} \mathbb{E}_\eta \frac{J_{(x, y)}(\eta) \eta_s}{C_J},
$$

where $\eta_s(\eta)$ is the stationary probability of the state $s$ determined by $p(\cdot | \eta)$. $E_{\eta}$ denote the expectation with respect to $p(\cdot | \eta)$.

Let $D_{(\alpha)}(\eta) \triangleq \prod x \in \mathcal{X} (\eta^x)^{1-\alpha}$ (Dirichlet function), then we have

$$
w_{J}(\eta) = \frac{\prod x \in \mathcal{X} \eta^x^{1/2} D_{(1/2)}(\eta)}{C_J}.
$$

We define another prior density as $w_{(\alpha)}(\eta) \triangleq \prod x \in \mathcal{X} (\eta^x)^{1-\alpha} D_{(1/2)}(\eta)$, where $C_{(\alpha)} \triangleq \int D_{(\alpha)}(\eta) d\eta$. Note that $w_{(\alpha)}(\eta) / w_{J}(\eta) \to \infty$ holds as $\eta$ approaches the boundaries of $H$, if $0 < \alpha < 1/2$ holds.

3 Results

3.1 Minimax Regret

We define a modified Jeffreys prior as

$$
w_{(\alpha)} \triangleq (1 - n^{-b}) w_J + n^{-b} w_{(\alpha)},
$$
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.

$$\bar{r}(m_n, \mathcal{A}_n) = \frac{\|T\|k}{2} \log n + \log C_J + o(1)$$

(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 $p(x_{n+1}|x^n, \eta) = \frac{p(x_{n+1}|\eta)}{p(x^n|\eta)}$. The prediction probability based on Bayes mixture with prior $w$ is given by $p_w(x_{n+1}|x^n) = \int p(x_{n+1}|x^n, \eta)p(\eta|x^n)d\eta = \int \eta(x^n)d\eta p(x_{n+1}|\eta)$, where $p(\eta|x^n)$ denotes the posterior density of $\eta$. 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 length of the sequence.

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 length 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 $\theta$ than $\eta$. Note that $p(x^n|\eta)$ is rewritten as follows:

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

where we let $\theta_x^s = \log(\eta_x^s/\eta_0)$ and $\psi(\theta_s) = -\log \eta_0 = \log(1 + \sum_{x \in \mathcal{X}'} \exp \theta_x^s)$. Note that $(\partial/\partial \theta_x^s) \psi(\theta_s) = \eta_x^s$ holds. We let $\Theta_s = \{\eta_s \in H_s^\alpha\}$, then $\Theta_s = \mathbb{R}^{\|X'\|}$. Let $\Theta(T) = \bigcap_{s \in \partial T} \Theta_s = \mathbb{R}^{\|\partial T\|\|X'\|}$. It is known that the map $\eta_s \mapsto \theta_s$ on $H_s^\alpha$ is one to one and analytic (see [2]).

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

**Theorem 2** Let $K$ be a compact set included in the interior of $H$ and $b_0$ be an arbitrary number. Then,

$$\bar{\eta}_n^s - \eta^s_n = \frac{1}{n_s + (k + 1)/2} \frac{\partial \log(w_J(\eta)/w(1/2)(\eta))}{\partial \theta_x^s} \bigg|_{\eta = \tilde{\eta}} = O\left(\frac{\sqrt{n}}{n}\right)$$

holds, uniformly for all sequences $x_{-d+1}...x_1x_2...$ such that $\eta \in K$ holds for all $n \geq b_0$.

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 $\eta \sim \tilde{\eta}$ and coincides with the well known fact that $\tilde{\eta} = \hat{\eta}$ 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

$$\frac{\partial}{\partial \theta_x^s} = \sum_{y \in \mathcal{X}'} \frac{\partial \eta_x^s}{\partial \theta_x^s} \frac{\partial}{\partial \eta_x^s} = \sum_{y \in \mathcal{X}'} \frac{\partial \eta_x^s}{\partial \theta_x^s} \frac{\partial}{\partial \eta_x^s} = \sum_{y \in \mathcal{X}'} \frac{\partial \eta_x^s}{\partial \theta_x^s} \frac{\partial}{\partial \eta_x^s}.$$

Hence, we have

$$\bar{\eta}_n^s - \eta^s_n = \sum_{y \in \mathcal{X}'} \frac{k\eta_x^s}{2n_s + k + 1} \left. \frac{\partial \log p_c(t|\eta)}{\partial \eta_x^s} \right|_{\eta = \tilde{\eta}}$$

$$= \frac{\sqrt{n}}{n_s \sqrt{n}}.$$

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 $x_0$, equals 0. By Theorem 2, the approximation of the Jeffreys mixture for this case is given by

$$\bar{\eta}_n^1 \sim \frac{n_0 + 0.5}{n_0 + 1} + \frac{1}{n_0 + 1} \left( \frac{1 - \bar{\eta}_n^1}{2} - \frac{\bar{\eta}_n^1}{\bar{\eta}_n^1 + \bar{\eta}_n^0} \right).$$
Note that this depends on not only $\hat{\eta}_0$ but $\hat{\eta}_1^0$ and that the difference between $\hat{\eta}_0$ 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 $\tilde{r}(q, x_{n-d+1}) \text{ def } \log(p(x^n|\tilde{\eta})/q(x^n|x_{n-d+1}^{0})) - \log(n/2\pi)$. Since value of $\tilde{r}(q, x_{n-d+1})$ depends on $x_{n-d+1}$, we generated a number of $x_{n-d+1}$ using pseudo random number with respect to $p(x^n|\tilde{\eta})$. Table 1 shows a part of our results, where the first row and the first column indicate the values of $\eta_0^0$ and $\eta_1^0$, respectively, which we generate $x_{n-d+1}$ with respect to. Each cell except them indicates average value of $\tilde{r}(q, x_{n-d+1})$. 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 $p(x^n|\tilde{\eta})$ is i.i.d., when $\eta_0^0 = 1 - \eta_1^0$ holds. One might expect that $q_A$ would perform better than $q_L$ for these cases, but it didn't.

<table>
<thead>
<tr>
<th>q_A</th>
<th>0.1</th>
<th>0.3</th>
<th>0.5</th>
<th>0.7</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>q_L</td>
<td>0.99</td>
<td>1.16</td>
<td>1.30</td>
<td>1.35</td>
<td>1.39</td>
</tr>
<tr>
<td>0.3</td>
<td>1.60</td>
<td>1.45</td>
<td>1.30</td>
<td>1.18</td>
<td>0.99</td>
</tr>
<tr>
<td>0.5</td>
<td>1.12</td>
<td>1.17</td>
<td>1.26</td>
<td>1.30</td>
<td>1.30</td>
</tr>
<tr>
<td>0.7</td>
<td>1.60</td>
<td>1.56</td>
<td>1.50</td>
<td>1.45</td>
<td>1.45</td>
</tr>
<tr>
<td>0.9</td>
<td>1.22</td>
<td>1.29</td>
<td>1.35</td>
<td>1.40</td>
<td>1.40</td>
</tr>
<tr>
<td></td>
<td>1.60</td>
<td>1.58</td>
<td>1.55</td>
<td>1.59</td>
<td>1.59</td>
</tr>
</tbody>
</table>

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

Note that $1.296 \leq \log C_J \leq 1.305$ 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 information, by which we have the following asymptotics:

$$\frac{\int p(x^n|\eta)w_{2}\eta J(\eta) \text{ d} \eta}{p(x^n|\eta)} \sim \frac{\sqrt{\det(J(\eta))}}{C_J \sqrt{\det(J(x^n, \eta))}} \frac{(2\pi)^{k|\partial T|/2}}{n^{k|\partial T|/2}},$$

where $J(x^n, \eta)$ is the empirical Fisher information. When the model is an exponential family, $J(x^n, \tilde{\eta}) = J(\tilde{\eta})$ 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:

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

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 $H^{(2n-\epsilon)}(T)$. Here, we let

$$H^{(\epsilon)} = H^{(\epsilon)}(T) \text{ def } \{\eta : \forall x \in \partial T, \forall x \in X, \eta^x > \epsilon\}$$

and $\epsilon$ is a certain small positive number. For the sequences which do not belong to $H^{(2n-\epsilon)}$, we obtain smaller regret than the minimax value with the help from the second term of $u_n$, $n^{-b}u_n(x, \eta^x)$. 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), 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.

$$\forall n \geq 1, \forall \epsilon : ne^d > 2, \forall x^n, s_0 : \hat{\eta} \in H^{(2\epsilon)} \implies n_s > ne^d - 1$$

and

$$|\log \frac{p_s}{p_c(x^n|s)}| < C_1 \frac{1}{ne^d}$$

**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 $p_c(s|x)$ are very simple.

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

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

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

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 $n_{s}^{x}$ 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 $n_{s}^{x} = n_{s}^{x(x)}$ holds.
Similarly, we let \( \forall x \in X, \forall s \in \partial T, \eta^s(x) \overset{\text{def}}{=} \eta^s_t \). We define
\[
D_s \overset{\text{def}}{=} \{ \tau(sx) : x \in X \}.
\]
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 \( s' \in D_s \), \( \eta^s_t \) is defined. We can define the state transition probability matrix as
\[
\Pi(s,s') = \begin{cases} \eta^s_t, & \text{when } s' \in D_s, \\ 0, & \text{otherwise.} \end{cases}
\]
(10)

First, we will show the following.

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

**Proof:** Note that the probabilities \( p_c(t|\eta) (t \in \partial T) \) satisfy the following line equation:
\[
p_c(t|\eta) = \sum_{t' \in \partial T} \Pi(s,t')p_c(t'|\eta).
\]
(11)

For each \( t \in \partial T \) and \( x \in X \), we have \( \eta^x_t > \epsilon \) by the assumption. Let \( y^d = t \). For each pair \( (t',t) \in (\partial T)^2 \), \( \tau_a(t'y^d) = t \). Further, for all \( y^d \in X^d \) and all \( i \in \{0,1,...,d-1\} \), \( \eta^t_{ij} > \epsilon \) holds. This implies that each element of \( \Pi^d \) is larger than \( \epsilon \), i.e. each \( p_c(t|\eta) \) is larger than \( \epsilon^d \).

**Q.E.D.**

**Lemma 2** There exist certain positive number \( C_1 \), such that
\[
\forall s \in \partial T, \forall t \in \partial T, \forall x \in X', \forall \epsilon > 0, \forall \eta \in H(c), \frac{1}{\eta^d} \left| \frac{\partial \log p_c(s|\eta)}{\partial \eta^t} \right| \leq C_1 \frac{\epsilon^d}{\epsilon^r}
\]
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
\[
A_{ij} = \Pi^d_{s_is_j}
\]
and a vector \( \mu \) as \( \mu = (p_c(s_1|\eta),...,p_c(s_q|\eta))^T \). (Note that when \( \eta \in H(c), A_{ij} > \epsilon^d \) holds.) Then, we have
\[
(I - A)\mu = 0
\]
\( (I \) denotes the unit matrix). Let \( \Delta_{ij} \) be the cofactor with respect to the \((i,j)\)-component of \( I - A \). By some manipulation, we have
\[
p_c(s_i|\eta) = \mu_i = \frac{\Delta_{ii}}{\sum_j \Delta_{ij}}.
\]
Note that \( \mu_i \) is a rational function of \( \eta \). Further, we can prove the following:
\[
\forall \eta \in H(c), \forall s \in \partial T, \Delta_{ii} \geq \epsilon^{d(q-1)}.
\]

We have
\[
\frac{\partial \log p_c(s|\eta)}{\partial \eta^t} = \frac{1}{\Delta_{ii}} \frac{\partial \Delta_{ii}}{\partial \eta^t} - \frac{1}{\sum_j \Delta_{ij}} \frac{\partial \sum_j \Delta_{ij}}{\partial \eta^t}.
\]
Note that the derivative of \( \Delta_{ij} \) is bounded upper by a certain constant. Therefore, we have
\[
\forall t \in \partial T, \forall x \in X', \forall \eta \in H(c), \left| \frac{\partial \log p_c(s|\eta)}{\partial \eta^t} \right| \leq \frac{C}{\epsilon^{d(q-1)}}.
\]

**Q.E.D.**

Now, we can prove Lemma 1.

**Proof of Lemma 1:** Let \( s_0 \) denote the state determined by \( x_0^{n+1} \) (initial state) and \( s_c \) the state \( \tau(x^n) \). First, we treat a special case in which \( s_0 = s_c \) holds. In this case, we have
\[
\forall s \in \partial T, \sum_{t \in \partial T} n^s_t = \sum_{t \in \partial T} n^s_t,
\]
(12)
since the number of all transition from the state \( s \) equals the number of all transition to the state \( s \). Hence, we have
\[
\sum_{t \in \partial T} \hat{n}^s_t \hat{n}_t = \sum_{t \in \partial T} \frac{n^s_t n_t}{n} = \sum_{t \in \partial T} \frac{n^s_t n_t}{n} = n_s = \hat{n}_s.
\]
This implies \( \hat{p}_s = p_c(s|\eta) \).

When \( s_0 \neq s_c \), let \( x_n^{n+1} \) be a minimum path from the state \( s_c \) to \( s_0 \) (does not exceed \( d \)). By adding a sequence \( x_n^{n+1} \) to the sequence \( x^n \), we have \( \tau(x^{n+1}) = s_0 \). Let \( \hat{n}_s \) denote the relative frequency of the state \( x^n \). Then, we have \( \hat{n}_s = p_c(s|\eta) \). Let \( \phi^d \) denote the number of transition from the state \( s \) to the state \( t \) in the sequence \( x_n,...,x_{n+1} \). Here, \( \phi^d = 0 \) or 1, since \( x_{n+1} \) is a minimum pass from \( s_0 \) to \( s_c \). Let \( \phi^d = \sum_t \phi^d_t \). We have \( \hat{n}_s^t = (n^s_t + \phi^d_t)/(n_s + \phi^d) \). Hence
\[
\hat{n}_s^t = \frac{n^s_t}{n_s + 1} = \frac{\hat{n}_s^t}{1 + 1/n_s} \geq \frac{\hat{n}_s^t}{2} > \epsilon,
\]
where we use the fact that \( n_s = \sum_t n^s_t \geq 1 \) (if \( n^s_t = 0 \) holds for all \( t \in \partial T \), then we have \( n_s \leq 1 \) for all \( s \in \partial T \)).

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

Hence, we have
\[
\hat{n}_s^t \geq \frac{\hat{n}_s^t}{1 + 1/n_s} > \frac{\hat{n}_s^t}{1 + 1/(ne^d - 1)}.
\]
Hence,
\[
\hat{n}_s^t < \frac{n_s^t(1 + 1/(ne^d - 1))}{n_s^t + 1/n_s} \leq \hat{n}_s^t + \frac{1}{ne^d - 1} < \hat{n}_s^t + \frac{2}{ne^d} = \hat{n}_s^t + \frac{2}{ne^d}.
\]
Also, we have \( \hat{n}_s^t < n_s^t + 1/n_s = \hat{n}_s^t + 1/n_s \). Therefore, we have
\[
|\hat{n}_s^t - n_s^t| < 2/ne^d.
\]
By Taylor’s theorem, we have
\[
\log p_c(s|\hat{\eta}) - \log p_c(s|\hat{\eta}) = \sum_{t \in \mathcal{T}, x \in \mathcal{X}'} \frac{\partial \log p_c(s|\hat{\eta})}{\partial \hat{\eta}_t^x} |_{\eta=h} (\hat{\eta}_t^x - \hat{\eta}_t^x),
\]
where \( h \) is a point between \( \hat{\eta} \) and \( \hat{\eta} \). Since \( \hat{\eta}, \hat{\eta} \in H^{(c)}(T), h \in H^{(c)}(T) \) holds. Hence by Lemma 2, we have
\[
\left| \frac{\partial \log p_c(s|\hat{\eta})}{\partial \hat{\eta}_t^x} \right|_{\eta=h} \leq C. e^c.
\]
Hence, we have
\[
-2C d|\partial T| \leq \log \frac{p_c(s|\hat{\eta})}{p_c(s|\hat{\eta})} \leq 2C d|\partial T| \cdot \frac{1}{ne^{r+d}}. \tag{13}
\]
Since \( \hat{p}_s = (n_s + \phi_s)/(n + \alpha) \) holds, we have
\[
\hat{p}_s \geq \frac{n_s}{n + \alpha} = \frac{\hat{p}_s}{1 + \alpha/n} \geq \frac{\hat{p}_s}{1 + d/n}
\]
and
\[
\hat{p}_s \leq \frac{n_s + 1}{n} = \hat{p}_s + \frac{1}{n} = \hat{p}_s(1 + \frac{1}{n\hat{p}_s}) = \hat{p}_s(1 + \frac{1}{n}).
\]
Hence,
\[
\frac{1}{1 + \frac{1}{n\hat{p}_s}} \leq \frac{\hat{p}_s}{\hat{p}_s} \leq 1 + \frac{d}{n},
\]
i.e.
\[
- \frac{1}{n} \leq \log \frac{\hat{p}_s}{\hat{p}_s} \leq \frac{d}{n}
\]
holds. Together with (13) and \( \hat{p}_s = p_c(s|\hat{\eta}) \), we have
\[
\frac{C_2}{ne^{r+d}} < - \frac{2C_1 d|\partial T|}{ne^{r+d}} \leq \frac{1}{n}\frac{d}{n} \leq \frac{2C_1 d|\partial T|}{ne^{r+d}} + \frac{d}{n} \leq \frac{C_2}{ne^{r+d}},
\]
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


