Stochastic Complexity for Tree Models

Jun'ichi Takeuchi
Department of Informatics, Kyushu University,
Fukuoka, Fukuoka, Japan

Andrew R. Barron
Department of Statistics, Yale University,
New Haven, Connecticut, USA

***Abstract*—We study the problem of data compression, gambling and prediction of strings $x^n = x_1x_2...x_n$ in terms of coding regret, where the tree model is assumed as a target class. We apply the minimax Bayes strategy for curved exponential families to this problem and show that it achieves the minimax regret without restriction on the data strings. This is an extension of the minimax result by (Takeuchi et al. 2013) for models of kth order Markov chains and determines the constant term of the Stochastic Complexity for the tree model.**

I. Introduction

We study the problem of data compression, gambling and prediction of strings $x^n = x_1x_2...x_n$ in terms of coding regret, where the tree model [21] is assumed as a target class.

A tree model is a parametric model of Markov sources defined by a context tree. It is used for data compression algorithms such as the Context Tree Weighting (CTW) Methods [22] and CONTEXT [11]. Note that the tree model is not an exponential family [1] unless it is an FSMX (Finite State Machine X) model [21], as pointed out in [19].

The Stochastic Complexity (SC) for a target class $S = \{p(\cdot|\theta) : \theta \in \Theta \subset \Re^d\}$ is defined as the codelength of the minimax code for S in terms of coding regret, which is the difference of the loss (codelength) incurred and the loss of an ideal coding for each string. A coding scheme for strings of length n is equivalent to a probabilistic mass function $q(x^n)$ on $\mathcal{X}^n$. It is well known that the normalized maximum likelihood (NML) is the exact minimax code [13], but in this study, we design an asymptotic minimax code by Bayesian mixtures mainly with Jeffreys prior over S, which is the prior density proportional to $|J(\theta)|^{1/2}$, where $|J(\theta)|$ is the determinant of the Fisher information $J(\theta)$ of θ.

It is known that such strategies for various target classes achieve the minimax regret for $W_n = \{x^n : \hat{\theta} \in K \subset \Theta\}$:

$$\frac{d}{2}\log\frac{n}{2\pi} + \log\int_K |J(\theta)|^{1/2}d\theta + o(1), \tag{1}$$

where $\hat{\theta}$ is maximum likelihood estimate (MLE) of θ. If the target class is an exponential family [4], [1], including models of Markov chains, then the Jeffreys prior with modification on the boundary of the parameter space is asymptotically minimax. In particular about Markov models, the minimax Bayesian strategy without restriction on data strings ($K = \Theta$) is demonstrated in [20], where the target class is a class of kth order Markov chains, which corresponds to a special case of the FSMX model and an exponential family in an asymptotic sense. For this target class, direct evaluation of the codelength of the NML is given in [6], Though the constant term's expression by [6] is different from that by [20], it is confirmed in [14] that both are equivalent. Note that it is easy to extend the result in [20] to the FSMX model's case.

In contrast, if the target class is not exponential type, then any Bayesian mixture of the target class cannot achieve the minimax regret [3], [15], [16], [17]. Hence we need a different technique. In [16], we discussed two minimax strategies for non-exponential families; one is for curved exponential families and the other is for general smooth families. Since the tree model is a curved exponential family in general [19], we employ the former strategy, which uses prior distributions over the exponential family in which the target curved exponential family is embedded. We prove that it achieves the minimax regret for the whole set of data strings (Theorem 2 in Section VI). This is an extension of Theorem 1 of [20] and an affirmative solution to a conjecture in [18].

II. Preliminaries

Let $\mathcal{X} = \{0, 1, ..., D\}$ be an alphabet and $p(x^n|\theta)$ a probability mass function of $x^n = x_1x_2...x_n \in \mathcal{X}^n$, parametrized by $\theta \in \Theta \subset \Re^d$. Assume $p(x^n|\theta)$ defines a stationary stochastic process. In particular, our concern is in the case where $p(x^n|\theta)$ corresponds to a tree source [21]. Since a tree source is a Markov source, we assume that an initial string $x^0_{-k+1} = x_{-k+1}x_{-k+2}...x_0$ is given before x^n, where k is the order of the Markov source. In this paper, we let $p(x^n|\theta)$ denote the conditional $p(x^n|x^0_{-k+1}, \theta)$ given x^0_{-k+1}. Define empirical Fisher information $\hat{J}(\theta, x^n)$ as the Hessian of $-(1/n)\log p(x^n|\theta) = -(1/n)\log p(x^n|x^0_{-k+1}, \theta)$ with respect to θ and Fisher information $J(\theta)$ by $J(\theta) = \lim_{n\to\infty} E_\theta \hat{J}(\theta, x^n)$, where E_θ denotes the expectation with respect to $p(x^n_{-k+1}|\theta)$.

We review the definition of tree source [21]. Let T be a finite subset of $\mathcal{X}^* \stackrel{\text{def}}{=} \{\lambda\} \cup \mathcal{X} \cup \mathcal{X}^2 \cup ...$, where λ denotes a null string. 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 a set T is referred to as a context tree. For a context tree T, define $\partial T \stackrel{\text{def}}{=} \{xs : x \in \mathcal{X}, s \in T\} \setminus T$. Here, ∂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 a string $s \in \mathcal{X}^*$, let $c(s)$ denote the element of ∂T which matches a postfix of s, if it exists. We refer to $c(s)$ as the context of s (or the state for s). Let $k \stackrel{\text{def}}{=} \max_{s\in\partial T} |s|$ ($|s|$ is length of s). When $|s| \geq k$, $c(s)$ uniquely exists. An

978-1-4799-5999-0/14/$31.00 ©2014 IEEE

information source in which the probability of the successive character is defined for each context, is referred to as a tree source. If the set of contexts ∂T satisfies a condition that $c(sx)$ for any $s \in \partial T$ and any $x \in \mathcal{X}$ is determined (i.e. c defines a state transition function), then the tree source is referred to as an FSMX source. We give examples of context trees for an FSMX source and a non-FSMX tree source.

Example 1: Assume $\mathcal{X} = \{0, 1\}$. Let $T_1 = \{\lambda, 1, 10, 0\}$, then we have $\partial T_1 = \{00, 11, 01, 110, 010\}$, which is a complete postfix set (See Figure 1). This tree defines a state

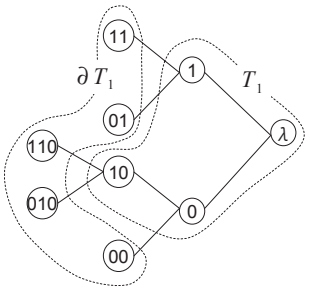

Fig. 1. A Context Tree for an FSMX model

transition function and the source is an FSMX one.

Example 2: Removing '11' and '01' from ∂T_1 ('1' from T_1), we obtain T_2 in Figure 2. If '0' is generated at the context

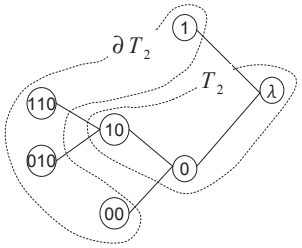

Fig. 2. A Context Tree for a non-FSMX tree source

'1', we cannot determine whether the machine has transferred to the context '110' or '010'. Hence this tree does not define an FSMX source.

Let us introduce the tree model given a context tree T. Let ℓ denote the number of contexts $|\partial T|$. Let $w_{x|s}$ denote the probability that x is generated at the context s. For each $s \in \partial T$, define a D-dimensional vector $w_s = (w_{1|s}, w_{2|s}, ..., w_{D|s})^t$. Let w denote a $D\ell$-dimensional vector $(w_{s_1}^t, w_{s_2}^t, , ..., w_{s_{|\partial T|}}^t)^t$. Define the range of parameter w as

$$\mathcal{W}_s = \Big\{ w_s : \forall x \in \mathcal{X},\ w_{x|s} \geq 0 \ \text{ and } \ \sum_{x=1}^{D} w_{x|s} \leq 1 \Big\}$$

and $\mathcal{W} = \mathcal{W}(T) \stackrel{\text{def}}{=} \prod_{s \in \partial T} \mathcal{W}_s$, where we assume that $w_{0|s} = 1 - \sum_{x=1}^{D} w_{x|s}$ is a dependent variable.

Let x_m^n denote a string $x_m x_{m+1} ... x_n \in \mathcal{X}^{n-m+1}$ $(m \leq n)$ and x^n a string x_1^n. Assume that we have an initial string x_{-k+1}^0 in advance. Define the probability mass function for the sequence x^n as

$$p_T(x^n | x_{-k+1}^0, w) = \prod_{i=0}^{n-1} w_{x_{i+1}|c(x_{-k+1}^i)}.$$

and the tree model [21], [8] based on T as

$$S(T) \stackrel{\text{def}}{=} \{ p_T(\cdot|\cdot, w) : w \in \mathcal{W}(T) \}. \tag{2}$$

When c defines a state transition function, then we refer to $S(T)$ as an FSMX model [21], [8]. The class of kth order Markov chains (case of $T = T^{(k)} \stackrel{\text{def}}{=} \bigcup_{i=0}^{k-1} \mathcal{X}^i$) is an example of FSMX model. Given x_{-k+1}^n, for every $sx \in \partial T \times \mathcal{X}$, let

$$\tau'_{sx} = \#\{ t : sx = x_{t-|s|}^t \text{ and } 1 \leq t \leq n \}$$

and let $\tau'_s = \sum_{x \in \mathcal{X}} \tau'_{sx}$. Then

$$p_T(x^n | x_{-k+1}^0, w) = \prod_{sx \in \partial T \times \mathcal{X}} (w_{x|s})^{\tau'_{sx}},$$

holds. By this, MLE of w is denoted as $\hat{w}_{x|s} = \tau'_{sx}/\tau'_s$, where we define $\hat{w} = \arg\max_{w \in \mathcal{W}} p_T(x^n | x_{-k+1}^0, w)$.

For an FSMX model, define $\sigma_i = c(x_{-k+1}^i)$, then $\sigma_i = c(\sigma_{i-1} x_i)$ and we have the sequence of contexts $\sigma_0^n = \sigma_0 \sigma_1 ... \sigma_n$ induced by x_{-k+1}^n. Let τ_{st} for $st \in (\partial T)^2$ denote the number of appearances of the pattern st in σ_0^n and $\tau_s = \sum_t \tau_{st}$. Then, $\tau'_{sx} = \tau_{sc(sx)}$ and $\tau'_s = \tau_s$ hold. Here, we define $w_{t|s}$ for each $st \in (\partial T)^2$ by extending the domain of $w_{x|s}$ as follows. If there exists an $x \in \mathcal{X}$ such that $c(sx) = t$, let $w_{t|s} = w_{x|s}$, otherwise $w_{t|s} = 0$. Then the probability mass function can be denoted as

$$p_T(x^n | \sigma_0, w) = \prod_{sx \in \partial T \times \mathcal{X}} (w_{x|s})^{\tau'_{sx}} = \prod_{st \in (\partial T)^2} (w_{t|s})^{\tau_{st}},$$

where we define $0^0 = 1$. Hence we can define $\hat{w}_{t|s} = \tau_{st}/\tau_s$. Then $\hat{w}_{c(sx)|s} = \hat{w}_{x|s}$ holds.

Note that any tree model is a subspace of a certain FSMX model. In fact, a tree model $S(T)$ is a subspace of FSMX model $S(T^{(k)})$, where $k \stackrel{\text{def}}{=} \max_{s \in \partial T} |s|$. The following provides a non-trivial example.

Example 3: Consider T_1 and T_2 in Examples 1 and 2. We have $S(T_2) = \{ p(\cdot|\cdot, w) \in S(T_1) : w_{1|01} = w_{1|11} \}$. That is, $S(T_2)$ is a subspace of $S(T_1)$.

For a tree model $S(T)$, (sx, ty)-entry (entry for $w_{x|s}$ and $w_{y|t}$) of the empirical Fisher information is

$$\hat{J}_{sx,ty}(w, x^n) = \frac{\delta_{st} \tau'_s}{n} \Big(\frac{\delta_{xy} \hat{w}_{x|s}}{(w_{x|s})^2} + \frac{\hat{w}_{0|s}}{(w_{0|s})^2} \Big),$$

where δ_{st} is Kronecker's delta. Hence the Fisher information is

$$J_{sx,ty}(w) = \delta_{st} \mu_s(w) \Big(\frac{\delta_{xy}}{w_{x|s}} + \frac{1}{w_{0|s}} \Big),$$

where $\mu_s(w)$ is the expectation of τ'_s/n with respect to $p_T(x_{-k+1}^n | w)$.

Next, we introduce the minimax regret for the tree model case following [20]. Let q be a conditional probability mass function of x^n given $s_0 = x_{-k+1}^0$. We consider the following

conditional regret of q for a target class $S(T)$ and a set of strings $W_n \subset \mathcal{X}^n$ given s_0:

$$r_n(q, W_n|s_0) = \max_{x^n \in W_n} \log \frac{p(x^n|s_0, \hat{w})}{q(x^n|s_0)},$$

The minimax regret is defined as $r_n(q, W_n|s_0) = \inf_q r_n(q, W_n|s_0)$.

III. Exponential and Curved Exponential Families

First we review the exponential family of i.i.d. probability densities. If the probability density function is given as

$$p(z|\theta) = \exp(\theta \cdot V(z) - \psi(\theta)), \tag{3}$$

then the class $\{p(\cdot|\theta) : \theta \in \Theta\}$ is referred to as an exponential family of probability densities [4] [1]. Here, $V(z)$ is a d-dimensional real valued random variable, θ a d-dimensional parameter, and $\theta \cdot V(z) = \sum_{i=1}^d \theta_i V_i(z)$. Let $\partial_i \stackrel{\text{def}}{=} \partial/\partial\theta_i$. Define $\eta_i \stackrel{\text{def}}{=} \partial_i \psi(\theta)$, then we have $\eta_i = E_\theta V_i(z)$. Further, $J_{ij}(\theta) = \partial_j \eta_i$ holds. The parameters η and θ are referred to as expectation parameter and canonical parameter, respectively. For the exponential family (3), the probability density function of a string $z^n = z_1 z_2 ... z_n$ is given as $p(z^n|\theta) \stackrel{\text{def}}{=} \prod_t p(z_t|\theta)$ $= \exp(n(\theta \cdot \bar{V} - \psi(\theta)))$, where $\bar{V} \stackrel{\text{def}}{=} (1/n)\sum_t V(z_t)$. Since $\partial_i \log p(z^n|\theta) = n(\bar{V}_i - \eta_i)$ holds, the MLE of η given z^n, equals the sufficient statistics $\bar{V}$.

Next we review the notion of curved exponential families. Consider the model embedded in an $\bar{d}$-dimensional exponential family $\bar{S} = \{\bar{p}(\cdot|u) : u \in \mathcal{U}\}$, where u is the natural parameter. Let $\theta \in \Theta$ be a d-dimensional vector ($d < \bar{d}$) and $\phi : \Theta \to \mathcal{U}$ a class C^∞ function, provided the rank of its Jacobian is d over Θ. Then, we refer to $S = \{p(\cdot|\theta) = \bar{p}(\cdot|\phi(\theta)) : \theta \in \Theta\}$ as a curved exponential family embedded in $\bar{S}$.

When S is an exponential family too, the Fisher information at MLE coincides with the empirical Fisher information at MLE. It can be confirmed as follows. First note the following.

$$\begin{aligned} \frac{\partial}{\partial\theta_a} \log p(z^n|\theta) &= \sum_i \frac{\partial u_i}{\partial \theta_a}(\bar{V}_i - \eta_i) \\ \hat{J}_{ab}(\theta, z^n) &\stackrel{\text{def}}{=} -\frac{1}{n}\frac{\partial^2}{\partial\theta_a \partial\theta_b} \log p(z^n|u) \\ &= -\sum_i \frac{\partial^2 u_i}{\partial\theta_a\partial\theta_b}(\bar{V}_i - \eta_i) + J_{ab}(\theta), \end{aligned} \tag{4}$$

where $J_{ab}(\theta)$ and $\hat{J}_{ab}(\theta, z^n)$ are Fisher information and empirical Fisher information with respect to θ, respectively. Since the log likelihood is maximized at $\hat{\theta}$, $\sum_i (\partial u_i/\partial\theta_a)(\bar{V}_i - \eta_i)$ $= 0$ holds at $\theta = \hat{\theta}$. If S is an exponential family, $\phi(\theta)$ forms a hyper plane in $\mathcal{U}$. Hence, $\partial^2 u/\partial\theta_a\partial\theta_b$ belongs to the linear space spanned by $\{\partial u/\partial\theta_a\}_{a=1,\cdots,d}$. Therefore, the first term of the third side of (4) equals zero.

Now we state the definition of an *asymptotic exponential family,* which is a refinement of the one given in [19].

Definition 1 (Asymptotic Exponential Family): For a parametric model $S = \{p(x^n|x_{-k+1}^0, \theta) : \theta \in \Theta\}$, assume that the probability density function is written as

$$p(x^n|\theta) = \exp(n(\theta \cdot V(x_{-k+1}^n) - \psi(\theta)) + U(x_{-k+1}^n|\theta)),$$

where V and U are certain functions of x_{-k+1}^n ($n = 1, 2, \cdots$). If for every compact set K interior to Θ,

$$\max_{i,j,x_{-k+1}^n} \max_{\theta \in K} \frac{1}{n}\Big|\frac{\partial^2 U(x_{-k+1}^n|\theta)}{\partial\theta_i\partial\theta_j}\Big| \to 0 \ (n \to \infty)$$

holds, then we refer to S as an asymptotic exponential family.

Remark 1: The (ordinary) exponential family is an asymptotic exponential family. Hence, when we say "a model is not an asymptotic exponential family," it implies that the model is not an exponential family.

Remark 2: For models of Markov sources, this definition is equivalent to the definition of the exponential family of Markov sources given by Nagaoka [10] and employed in [5].

Note that the FSMX model is an example of asymptotic exponential family. This fact can be confirmed by showing

$$\lim_{n\to\infty} \max_{x_{-k+1}^n : \hat{w} \in K} |J_{sx,ty}(\hat{w}) - \hat{J}_{sx,ty}(\hat{w}, x_{-k+1}^n)| = 0. \tag{5}$$

See [19], [20] for the proof. In [19], the proof is given for binary alphabet case. In [20], the proof is given for $T = T^{(k)}$ case. Since the key of the proof is an equality $\sum_t(\tau_{st} - \tau_{ts}) = \pm 1$ or 0, which holds for any FSM, those proofs work for all FSMX models. See also [24], [10].

When the target class is an asymptotic exponential family $S = \{p(x^n|\theta) : \theta \in K \subset \Theta^\circ\}$, the modified Jeffreys mixture is asymptotically minimax. Define $\{K_i\}$ be a sequence of subsets of Θ° such that $K_{n+1}^\circ \supset K_n$ and K_n slowly approaches K as n increases. Let $w_n(\theta)$ denote the Jeffreys prior over K_n. Let $m_n(x^n|s_0) = \int p(x^n|s_0, w) w_n(\theta) d\theta$. Then by Laplace integration, we have

$$\frac{m_n(x^n|s_0)}{p(x^n|s_0, \hat{\theta})} \sim \frac{|J(\hat{w})|^{1/2}}{C_J(K_n)|\hat{J}(\hat{w}, x^n)|^{1/2}} \frac{(2\pi)^{\ell D/2}}{n^{\ell D/2}}.$$

For the asymptotic exponential family, $|J(\hat{w})|/|\hat{J}(\hat{w}, x^n)|$ is canceled and the mixture m_n asymptotically achieves SC.

Hence for the FSMX model, when data strings are restricted so that MLE is in a compact K interior to Θ, sequence of Jeffreys mixture is asymptotically minimax.

IV. Exponential Curvature of Tree Models

In the previous section, we note that the FSMX model is an asymptotic exponential family. In contrast, it was shown in [19] that the non-FSMX tree model is not an asymptotic exponential family. The following theorem was shown provided $\mathcal{X}$ is binary. Hence in this section, we assume $\mathcal{X}$ is binary, but we think it is not difficult to extend this theorem to finite alphabet case.

Theorem 1 (Takeuchi & Kawabata 2007): For a context tree T, $S(T)$ is an asymptotic exponential family, iff $S(T)$ is an FSMX model.

This theorem is proved by showing that $\hat{J}(\hat{w}) - J(\hat{w})$ for the non-FSMX tree model does not converge to zero. Since

$$\hat{J}_{sx,ty}(\hat{w}) - J_{sx,ty}(\hat{w}) = \delta_{st}(\hat{w}_s - \mu_s(\hat{w}))\Big(\frac{\delta_{xy}}{w_{x|s}} + \frac{1}{w_{0|s}}\Big),$$

that is reduced to the difference $\hat{w}_s - \mu_s(\hat{w})$.

Note that $S(T)$ is an FSMX model, iff sx does not belong to T for every $s \in \partial T$ and every $x \in \mathcal{X}$. Now, consider the tree $T = \{\lambda\} \cup \{s1 : s \in T_1\} \cup \{s0 : s \in T_2\}$, where T_1 and T_2 are context trees such that $T_1 \supset T^{(d)}$ and $T_2 \subset T^{(d-2)}$. Then for any $s0 \in \partial T$, its successor $s01$ belongs to T, hence $s01$'s context is not determined. This suggests that most tree models are not an asymptotic exponential family.

It means that the Jeffreys mixture is not asymptotically minimax for most tree models. Hence, we try to apply the asymptotic minimax strategy for the curved exponential family which was proposed and discussed in [3], [15], [16], [18].

V. Minimax Strategy for Curved Exponential Family

Here we review the minimax strategy for the curved exponential family described in [16]. Assume that the target class M is a curved exponential family introduced in Section III. Let K be a compact set included in Θ° and $\{K_n\}_n$ be a sequence of compact subsets of Θ such that $K_{n+1}^\circ \supset K_n$ and K_n shrinks to K as n goes to infinity.

In this case, the series of the following mixtures asymptotically achieves the minimax regret (1).

$$\bar{m}_n(z^n) = (1 - n^{-b}) m_{J,n}(z^n) + n^{-b} \int \bar{p}(z^n|u) w(u) du, \quad (6)$$

where $m_{J,n}$ is the Jeffreys mixture of $p(z^n|\theta)$ over $K_n \supset K$ and $w(u)$ is a certain probability density on $\mathcal{U}$.

This can be shown as follows. Let $\mathcal{G}_n = \{z^n | |\bar{V} - \eta(\hat{\theta})| \le n^{-q}$ and $\hat{\theta} \in K\}$ and $\mathcal{G}_n^c = \{z^n | |\bar{V} - \eta(\hat{\theta})| > n^{-q}$ and $\hat{\theta} \in K\}$. If $z^n \in \mathcal{G}_n$, then $||J(\hat{\theta}) - \hat{J}(\hat{\theta}, z^n)|| \le B n^{-q}$ holds, where B is a certain positive number determined by the function ϕ and the subspace K. Hence similarly to the asymptotic exponential family case, by the Laplace integration, it is easy to show that the first term of $\bar{m}_n$ works well.

To handle $z^n \in \mathcal{G}_n^c$, let $\tilde{\eta}$ be the point between $\bar{V}$ and $\eta(\hat{\theta})$ such that $|\tilde{\eta} - \eta(\hat{\theta})| = n^{-q}$. Then we have $KL(\bar{p}(\cdot|\tilde{u}) | p(\cdot|\hat{\theta})) \ge A n^{-2q}$, where $KL(p|p')$ denotes Kullback-Leibler divergence from p to p', $\tilde{u}$ the u corresponding to $\tilde{\eta}$, and A a certain positive number determined by $\bar{S}$ and K. From this, we can show $\bar{p}(x^n|\tilde{u}) > \exp(A n^{1-2q}) p(x^n|\hat{\theta})$. We have

$$\inf_{z^n \in \mathcal{G}_n^c} \frac{\int \bar{p}(z^n|u) w(u) d\theta}{p(z^n|\hat{\theta})} > n^{-\bar{d}} e^{A n^{1-2q}}, \quad (7)$$

where we have evaluated $\int \bar{p}(z^n|u) w(u) d\theta$ by the integration in the $n^{-1/2}$-neighborhood around $\tilde{u}$.

VI. Stochastic Complexity of Tree Models

If the data string is restricted as $\hat{w}$ belongs to a compact K interior to Θ°, it is easy to apply the strategy stated in the previous section to the tree model, and obtain an asymptotic minimax result. However in this section, we give a stronger result, which is without restriction on data strings.

Let $S(\bar{T})$ be an FSMX model and $S(T)$ be a non-FSMX tree model embedded in $S(\bar{T})$, that is, we assume $T \subset \bar{T}$. Then the target class is $S(T)$. Note that we may employ the FSM closure [9] of $S(T)$ as $S(\bar{T})$. Let $w \in \mathcal{W}$ be parameters for $S(\bar{T})$ representing the transition probabilities and $u \in \mathcal{U}$ those for $S(T)$. For every $s \in \partial T$, define a subset of $\partial \bar{T}$ as

$$\mathcal{C}_s = \{t : \exists s' \in \mathcal{X}^*,\ t = s's \text{ and } t \in \partial \bar{T}\}.$$

Define a map ϕ from $\mathcal{U}$ to $\mathcal{W}$ as

$$\forall s \in \partial T,\ \forall t \in \mathcal{C}_s,\ \phi_t(u) = u_s.$$

The function $w = \phi(u)$ embeds $S(T)$ into $S(\bar{T})$. In this section, let $p(x^n|u)$ and $\bar{p}(x^n|w)$ denote $p_T(x^n|u)$ and $p_{\bar{T}}(x^n|w)$, respectively.

Let $\rho_J(u)$ denote the Jeffreys prior over $\mathcal{U}$, $\rho_{(\alpha)}(u)$ Dirichlet(α) prior ($\alpha < 1$), and $\bar{\rho}_J(w)$ the Jeffreys prior over $\mathcal{W}$. Note that

$$\begin{aligned} |J(u)| &= \prod_{s \in \partial T} \Big(\sum_{t \in \mathcal{C}_s} \bar{\mu}_t(\phi(u)) \Big)^D \prod_{x \in \mathcal{X}} (u_{x|s})^{-1}, \quad (8) \\ \rho_{(\alpha)}(u) &\propto \prod_{s \partial T} \prod_{x \in \mathcal{X}} (u_{x|s})^{-(1-\alpha)}, \end{aligned}$$

where $\bar{\mu}_t(w)$ is the stationary probability of $t \in \partial \bar{T}$ determined by w. For every $s \in \partial T$, let $\mu_s(u) = \sum_{t \in \mathcal{C}_s} \bar{\mu}(\phi(u))$.

Our minimax strategy is

$$\begin{aligned} \bar{m}_n(x^n) &= (1 - 2n^{-b}) \int p(x^n|u) \rho_J(u) du \quad (9) \\ &+ n^{-b} \int p(x^n|u) \rho_{(\alpha)}(u) du \quad (10) \\ &+ n^{-b} \int \bar{p}(x^n|w) \bar{\rho}_J(w) dw. \quad (11) \end{aligned}$$

Theorem 2: The strategy $\bar{m}_n$ asymptotically achieves the minimax regret for $S(T)$, i.e. for every $s_0 \in \partial T$,

$$\max_{x^n \in \mathcal{X}^n} \log \frac{p(x^n|\hat{\theta})}{\bar{m}_n(x^n)} = \frac{D\ell}{2} \log \frac{n}{2\pi} + \log \int |J(u)|^{1/2} du + o(1),$$

where $\alpha < 1/2$, $k = \max_{s \in \partial T} |s|$, $a < 1/k$, and $b < \iota = (1/2 - \alpha)a$.

Outline of the proof: Define the interior region of $\mathcal{U}$ as

$$\mathcal{U}_n = \{u : \forall s \in \partial T, \forall x \in \mathcal{X},\ u_{x|s} \ge 2n^{-a}\}.$$

When $\hat{u} \notin \mathcal{U}_n$, the regret of the second term of (9) is smaller than the minimax value for sufficiently large n. This is the effect of boundary modification for the Jeffreys prior.

Next, consider the strings with $\hat{u} \in \mathcal{U}_n$. Divide such strings to the set of good strings $\mathcal{G}_n$ and that of bad strings $\mathcal{G}_n^c$ as:

$$\begin{aligned} \mathcal{G}_n &= \{x^n : \hat{u} \in \mathcal{U}_n,\ \forall s \in \partial T,\ \forall t \in \mathcal{C}_s,\ |\hat{w}_t - \hat{u}_s| < n^{-q}\} \\ \mathcal{G}_n^c &= \{x^n : \hat{u} \in \mathcal{U}_n\} \setminus \mathcal{G}_n, \end{aligned}$$

assuming $q > a$ and $n^{-q} \le n^{-a}$.

When $x^n \in \mathcal{G}_n$, the regret of the first term of (9) achieves the minimax value asymptotically, because the empirical Fisher information uniformly converges to the Fisher information.

When the data string belongs to $\mathcal{G}_n^c$, the third term of our mixture works, since the maximum likelihood in $S(\bar{T})$ is significantly larger than that in $S(T)$. In fact

$$\begin{aligned}
\log \frac{\bar{p}(x^n|\hat{w})}{p(x^n|\hat{u})} &= \log \frac{\bar{p}(x^n|\hat{w})}{\bar{p}(x^n|\phi(\hat{u}))} \\
&= n \sum_{t\in\partial\bar{T}} \sum_{x\in\mathcal{X}} \frac{\bar{\tau}_{tx}}{n} \log \frac{\hat{w}_{x|t}}{\phi_{x|t}(\hat{u})} \\
&= n \sum_{s\in\partial T} \sum_{t\in\mathcal{C}_s} \sum_{x\in\mathcal{X}} \frac{\bar{\tau}_{tx}}{n} \log \frac{\hat{w}_{x|t}}{\hat{u}_{x|s}} \\
&= n \sum_{s\in\partial T} \sum_{t\in\mathcal{C}_s} \frac{\bar{\tau}_t}{n} \sum_{x\in\mathcal{X}} \hat{w}_{x|t} \log \frac{\hat{w}_{x|t}}{\hat{u}_{x|s}} \\
&\geq nn^{-ka} \sum_{s\in\partial T} \sum_{t\in\mathcal{C}_s} \sum_{x\in\mathcal{X}} \hat{w}_{x|t} \log \frac{\hat{w}_{x|t}}{\hat{u}_{x|s}} \\
&= n^{1-ka} \sum_{s\in\partial T} \sum_{t\in\mathcal{C}_s} KL(\hat{w}_t|\hat{u}_s).
\end{aligned}$$

Here we have used $\bar{\tau}_t/n \geq n^{-ka}$, which holds from Lemma 1 of [20]. Since there exists a pair of $s' \in \partial T$ and $t' \in \mathcal{C}_{s'}$ such that $|\hat{w}_{t'} - \hat{u}_{s'}| \geq n^{-q}$. Noting $\hat{w}_{x|t} \geq n^{-a} \geq n^{-q}$, we can show $KL(\hat{w}_{t'}|\hat{u}_{s'}) \geq n^{-2q}/4$. Therefore, we have

$$\log \frac{\bar{p}(x^n|\hat{w})}{p(x^n|\hat{u})} \geq \frac{n^{1-ka-2q}}{4},$$

which is $\bar{p}(x^n|\hat{w}) \geq \exp(n^{1-ka-2q}/4)p(x^n|\hat{u})$. Hence, similarly to (7) we have

$$\frac{\int \bar{p}(x^n|w)\bar{\rho}_J(w)dw}{p(x^n|\hat{u})} \geq C\exp(n^{1-ka-2q}/4)n^{-\bar{\ell}D}, \qquad (12)$$

where C is a certain constant. When $ka + 2q < 1$, the last expression is larger than the minimax level for all sufficiently large n. Therefore we have the theorem's claim.

VII. Discussion

From Theorem 1 with (8), we see that the constant term of the SC for $S(T)$ equals

$$\log \int_{\mathcal{U}} \prod_{s\in\partial T} \Big(\sum_{t\in\mathcal{C}_s} \bar{\mu}_t(\phi(u))\Big)^{D/2} \prod_{x\in\mathcal{X}} (u_{x|s})^{-1/2} du. \qquad (13)$$

Note that the stationary probabilities can be written [20] as

$$\mu_s(w) = \frac{\Delta_{ss}}{\sum_{t\in\partial T} \Delta_{tt}},$$

where Δ_{st} is the (s,t)-cofactor of the matrix of which entries are $\delta_{st} - w_{s|t}$ for every $st \in (\partial T)^2$. If $S(T)$ is an FSMX model, we can set $S(\bar{T}) = S(T)$, from which we have $\mathcal{C}_s = \{s\}$ and $w = \phi(u) = u$. Then (13) is reduced to the form

$$\log \int_{\mathcal{U}} \prod_{s\in\partial T} (\mu_s(w))^{D/2} \prod_{x\in\mathcal{X}} (u_{x|s})^{-1/2} du \qquad (14)$$

which is given in [20]. When $S(T)$ is the model of kth order Markov chains, Jacquet & Szpankowski directly evaluated the codelength of the NML by a combinatorial technique. Though the expression of the constant term in their result is different from (14). both are equivalent to each other. See [14] for the details.

Acknowledgment

The authors thank Hiroshi Nagaoka and Masahito Hayashi for fruitful discussions and the anonymous referees for their helpful comments. This research was supported in part by the Aihara Project, the FIRST program from JSPS, initiated by CSTP and in part by JSPS KAKENHI Grant Number 24500018.

References

[1] S. Amari & H. Nagaoka, *Methods of Information Geometry*, AMS & Oxford University Press, 2000.

[2] A. R. Barron, J. Rissanen, & B. Yu, "The minimum description length principle in coding and modeling," *IEEE trans. Inform. Theory,* Vol. 44 No. 6, pp. 2743 - 2760, 1998.

[3] A. R. Barron & J. Takeuchi, "Mixture models achieving optimal coding regret," *Proc. of 1998 Inform. Theory Workshop,* 1998.

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

[5] M. Hayashi & S. Watanabe, "Information Geometry Approach to Parameter Estimation in Markov Chains," arXiv:1401.3814v1, 2014.

[6] P. Jacquet & W. Szpankowski, "Markov types and minimax redundancy for Markov sources," *IEEE trans. Inform. Theory,* Vol. 50, No. 7, July 2004.

[7] H. Jeffreys, *Theory of probability, 3rd ed.* , Univ. of California Press, Berkeley, Cal, 1961.

[8] 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.

[9] A. Martìn, G. Seroussi, & M. J. Weinberger, "Linear time universal coding and time reversal of tree sources via FSM closure," *IEEE Trans. Inf. Theory,* vol. 50, no. 7, pp. 1442-1468, July 2004.

[10] H. Nagaoka, "The exponential family of Markov chains and its information geometry," *Proc. of the 28th Symposium on Information Theory and its Applications* (SITA2005), 2005.

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

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

[13] Yu M. Shtarkov, "Universal sequential coding of single messages," *Problems of Information Transmission*, vol. 23, pp. 3-17, July 1988.

[14] J. Takeuchi, "Fisher Information Determinant and Stochastic complexity for Markov Models," *Proc. of 2009 IEEE International Symposium on Information Theory,* pp. 1894-1898, Seoul, Korea, 2009.

[15] J. Takeuchi & A. R. Barron, "Asymptotically minimax regret by Bayes mixtures," *Proc. of 1998 IEEE ISIT,* p. 318, 1998.

[16] J. Takeuchi & A. R. Barron, "Asymptotically Minimax Regret by Bayes Mixtures for Non-exponential Families," *Proc. of 2013 IEEE ITW,* pp. 204-208, 2013.

[17] J. Takeuchi & A. R. Barron, "Asymptotically minimax regret for models with hidden variables," *Proc. of 2014 IEEE ISIT,*, to appear, 2014.

[18] J. Takeuchi, A. R. Barron, & T. Kawabata, "Statistical curvature and stochastic complexity," *Proc. of the 2nd Symposium on Information Geometry and Its Applications,* pp. 29–36, 2006.

[19] J. Takeuchi & T. Kawabata, "Exponential Curvature of Markov Models," *Proc. of 2007 IEEE International Symposium on Information Theory*, pp. 2891-2895, Nice, France, 2007.

[20] J. Takeuchi, T. Kawabata, & A. R. Barron, "Properties of Jeffreys mixture for Markov sources," *IEEE trans. Inform. Theory,* vol. 59, no. 1, pp. 438-457, 2013.

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

[22] F. Willems, Y. Shtarkov & T. Tjalkens, "The context tree weighting method: basic properties," *IEEE trans. Inform. Theory,* vol. 41. no. 3, pp. 653-664, 1995.

[23] 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.

[24] H. Itoh & S. Amari, "Geometry of information sources (in Japanese)," *Proc. of the 11th Symp. on Inform. Theory and Its Apps.*, pp. 57–60, 1988.