2nd International Symposium on
Information Geometry and its Applications
December 12-16, 2005, Tokyo
Pages 000–000

STATISTICAL CURVATURE AND STOCHASTIC COMPLEXITY

JUN-ICHI TAKEUCHI, ANDREW R. BARRON, AND TSUTOMU KAWABATA

1. Introduction

We discuss the relationship between the statistical embedding curvature [1, 2] and the logarithmic regret [11] (regret for short) of the Bayesian prediction strategy (or coding strategy) for curved exponential families and Markov models. The regret of a strategy is defined as the difference of the logarithmic loss (code length) incurred by the strategy and that of the best strategy for each data sequence among a considered class of prediction strategies. (The considered class is referred to as a reference class.) Since a prediction strategy is equivalent to a probability distribution, the class of prediction strategy is equivalent to a statistical model. Note that the logarithmic loss (equivalent to code length) by the minimax strategy is equal to Rissanen's stochastic complexity (SC). SC is generalization of Minimum Description Length [8, 3] and plays an important role in statistical inference such as model selection, universal prediction, universal coding, etc.

For this matter, it can be shown that the Bayesian strategy with Jeffreys prior (Jeffreys strategy for short) asymptotically achieves SC upto the constant term, when the reference class is an exponential family[12, 13, 16]. This is due to the fact that the logarithmic loss of Bayes mixture strategy is affected by the exponential curvature of the considered class. Hence, the Jeffreys strategy does not achieve the SC in general, if the reference class is not an exponential family. For a curved exponential family case, in order to obtain the minimax regret, we give a method to modify the Jeffreys mixture by assuming a prior distribution on the exponential family in which the curved family is embedded.

We also consider the expected version of regret (known as redundancy in information theory field). When the true probability distribution belongs to the reference class, the Jeffreys strategy asymptotically achieves the minimax redundancy, irrelevant to the curvature of the reference class as shown by Clarke and Barron [6]. However, if the true probability distribution does not belong to the reference class, the situation differs and the redundancy of Jeffreys strategy is affected by both exponential and mixture curvatures of the reference class.

Finally, we study the exponential curvature of a class of Markov sources defined by a context tree (tree model). Tree models are classified to FSMX models and non FSMX models. It is known that FSMX models are exponential families in asymptotic sense. We are interested in the problem if non FSMX models are exponential families or not. We show that a certain kind of non FSMX tree model is curved in terms of exponential curvature.

Almost all parts of this material is based on [13, 14, 4, 15].

2. Stochastic Complexity

We review the notion of stochastic complexity [10] and regret [11] for universal coding and universal prediction.

2.1. **Stochastic Complexity and Regret.** Regret is a performance measure used for the problem of data compression, gambling and prediction, and is defined as the difference of the loss incurred and the loss of an ideal coding or prediction strategy for each sequence. A coding scheme for the sequence of length n is equivalent to a probabilistic mass function $q(x^n)$ on $\mathcal{X}^n$. 'An ideal strategy' is selected from a certain class of probabilistic mass functions C, which we call a **reference class**.

The worst case regret of a strategy q with respect to a reference class $C = \{p(\cdot|u) : u \in U\}$ and a set of the sequences $W_n \subseteq \mathcal{X}^n$ is defined as

$$r_n(q, W_n, C) \stackrel{\text{def}}{=} \sup_{x^n \in W_n} \Big(\log \frac{1}{q(x^n)} - \log \frac{1}{p(x^n|\hat{u})} \Big),$$

where $\hat{u}$ is the maximum likelihood estimate (MLE) given x^n.

We consider minimax problems for sets of sequences such that $W_n = \mathcal{X}^n(K) \stackrel{\text{def}}{=} \{x^n : \hat{u} \in K\}$, where K is a certain nice subset (satisfies $\bar{K} = \bar{K^\circ}$) of U. For this problem, it is known that the normalized maximum likelihood

$$\hat{m}_n(x^n) \stackrel{\text{def}}{=} \frac{p(x^n|\hat{u})}{\int_{W_n} p(x^n|\hat{u}) dx^n}$$

achieves the minimax regret strictly [11]. Rissanen defined stochastic complexity of a sequence x^n with respect to the reference class C as the code length obtained by $\hat{m}_n$ [10]. That is, the stochastic complexity equals the minimax code length. Evaluation of stochastic complexity under certain regularity conditions is given as

$$\log \frac{1}{\hat{m}_n(x^n)} = \log \frac{1}{p(x^n|\hat{u})} + \frac{d}{2} \log \frac{n}{2\pi} + \log \int_K \sqrt{\det J(u)} du + o(1),$$

where $J(u)$ denotes the Fisher information matrix of u [10, 19, 12, 4, 13, 3].

3. Minimax Strategy by Bayes Mixtures

We are interested in the regret of mixture strategies. The Jeffreys mixture is the mixture by the prior proportional to $\sqrt{\det(J(u))}$. We denote the Jeffreys mixture over the set K by m_n. The value $C_J(K)$ is the normalization constant for the Jeffreys prior over the set K.

For the exponential families including the multinomial Bernoulli and FSM (Finite State Machine), it is known that a sequence of Jeffreys mixtures achieves the minimax regret asymptotically [20, 12, 13, 16]. For the multinomial exponential family case except for multinomial Bernoulli and FSM, these facts are proven under the condition that K is a compact subset included in the interior of U.

We briefly review outline of the proof for that case. Let $\{\mathcal{G}_n\}$ be a sequence of subsets of U such that $\mathcal{G}_n^\circ \supset K$. Suppose that $\mathcal{G}_n$ reduces to K as $n \to \infty$. Let $m_{J,n}$ denote the Jeffreys mixture for $\mathcal{G}_n$. If the rate of that reduction is sufficiently slow, then we have

$$\log \frac{p(x^n|\hat{u})}{m_{J,n}(x^n)} \sim \frac{d}{2} \log \frac{n}{2\pi} + \log C_J(K) + o(1), \tag{1}$$

where the remainder $o(1)$ tends to zero uniformly over all sequences with MLE in K. This implies that the sequence $\{m_{J,n}\}$ is asymptotically minimax. This is verified using the following asymptotic formula resulted by the Laplace integration, which holds uniformly:

$$\frac{m_{J,n}(x^n)}{p(x^n|\hat{u})} \sim \frac{\sqrt{\det(J(\hat{u}))}}{C_J(K)\sqrt{\det(\hat{J}(x^n, \hat{u}))}} \frac{(2\pi)^{d/2}}{n^{d/2}},$$

where $\hat{J}(x^n, u)$ is the empirical Fisher information. When C is an exponential family, $\hat{J}(x^n, \hat{u}) = J(\hat{u})$ holds. Hence, the above expression asymptotically equals the minimax value of regret mentioned in the former section.[1]

When the model C is not exponential type, the situation differs. The Jeffreys mixture is not guaranteed to be minimax, because the empirical Fisher information is not close to the Fisher information at the MLE in general. Note that the component of $\hat{J}(x^n, u) - J(u)$ orthogonal to the model C is its embedding exponential curvature.

When the reference class C is a curved exponential family, it is easy to see this. Assume that C is embedded in a $\bar{d}$-dimensional exponential family ($\bar{d} > d$)

$$S = \{\bar{p}(x|\theta) = \exp(\theta^i x_i - \psi(\theta)) : \theta \in \Theta\},$$

i.e.

$$p(\cdot|u) = \bar{p}(\cdot|\phi(u)),$$

where ϕ is a (4 times differentiable) function $U \to \Theta$. Then, we have

$$\hat{J}_{ab}(x^n, \hat{u}) = -\frac{\partial^2 \phi^i}{\partial u^a \partial u^b}\Big|_{u=\hat{u}} (\bar{x}_i - \eta_i(\hat{u})) + J_{ab}(\hat{u}), \tag{2}$$

where we let $\bar{x}$ denote $(1/n)\sum_{t=1}^n x_t$ and $\eta(u)$ the expectation parameter of S at $\theta = \phi(u)$.

First, assume that C is not curved in S (in the natural parameter space), then $\theta = \phi(u)$ forms a plane in Θ, i.e. the vectors $\partial^2\phi/\partial u^a \partial u^b$ $(a, b = 1, ..., d)$ are certain linear combinations of the vectors $\partial\phi/\partial u^a$ $(a = 1, ..., d)$. On the other hand,

$$\frac{\partial \phi^i}{\partial u^a}\Big|_{u=\hat{u}} (\bar{x}_i - \eta_i(\hat{u})) = 0$$

holds. Hence, we have

$$\frac{\partial^2 \phi^i}{\partial u^a \partial u^b}\Big|_{u=\hat{u}} (\bar{x}_i - \eta_i(\hat{u})) = 0.$$

This implies that $\hat{J}(x^n, \hat{u}) = J(\hat{u})$.

Second, assume that C is a (really curved) curved exponential family. At least one of the vectors $\partial^2\phi/\partial u^a \partial u^b$ $(a, b = 1, ..., d)$ for a certain $u^* \in U$ has a component orthogonal to the tangent space of C. We let $V(u^*)$ denote the linear space spanned by all such components. This holds for a certain neighborhood B^* of u^*, since ϕ is 4 times differentiable. Hence, there exists a sequence x^n such that $\hat{u} \in B^*$ and the mixture geodesic connecting $\eta = \bar{x}$ and $\eta = \eta(\hat{u})$, is not orthogonal to $V(\hat{u})$ at $\eta = \eta(\hat{u})$ (assuming n is sufficiently large). Hence, there exists a sequence x^n such that $\det J(\hat{u}) / \det \hat{J}(x^n, \hat{u}) \neq 1$ holds, except for the case that variation of $J(u)$ to the direction of $V(u)$ preserves the value of $\det J(u)$ for every $u \in B^*$. In other words, except for such cases, the Jeffreys mixtures do not achieve the minimax regret asymptotically.

Even for the curved exponential family case, we can modify the Jeffreys mixture to achieve the minimax regret asymptotically. In fact, the series of the following mixtures is asymptotically minimax with respect to regret.

$$\bar{m}_n(x^n) = (1 - n^{-a}) \int p(x^n|u) w_J(u) du + n^{-a} \int \bar{p}(x^n|\theta) w(\theta) d\theta,$$

where $w(\theta)$ is a certain probability density on Θ.

This can be derived as follows. If $|\bar{x} - \eta(u)|$ is small, then the difference between $J(\hat{u})$ and $\hat{J}(x^n, \hat{u})$ is small. This implies that the first term of $\bar{m}_n$ is nearly

[1] If K is the entire space for the statistical model, we cannot define the superset of K and need a different technique, which was established for the cases of multinomial Bernoulli, FSM, and a certain type of one-dimensional exponential families. See [20, 13, 16].

equal to the minimax value. Otherwise (concretely, $|\bar{x} - \eta(\hat{u})| > n^{-1/4}$), we have $D(\bar{p}(\cdot|\hat{\theta})|p(\cdot|\hat{u})) > An^{-1/2}$, where $D(\cdot|\cdot)$ denotes Kullback-Leibler divergence, $\hat{\theta}$ the maximum likelihood estimate of θ in S, and A a certain positive number. This is equivalent to

$$\frac{1}{n} \log \frac{\bar{p}(x^n|\hat{\theta})}{p(x^n|\hat{u})} > An^{-1/2}.$$

Hence, we have $\bar{p}(x^n|\hat{\theta}) > \exp(An^{1/2})p(x^n|\hat{u})$. Noting this fact, we can show

$$n^{-a} \int \bar{p}(x^n|\theta)w(\theta)d\theta > n^{-a-d} \exp(An^{1/2})p(x^n|\hat{u}).$$

The right hand side is larger than $p(x^n|\hat{u})$ for sufficiently large n. Hence the regret of $\bar{m}_n(x^n)$ is smaller than the minimax value (actually we have negative regret).

4. Redundancy and Relative Redundancy

Redundancy is expectation version of regret. For a strategy q and a reference class C, the maximum redundancy is defined as

$$R_n(q, C) \stackrel{\text{def}}{=} \max_{r \in C} E_r \log \frac{r(x^n)}{q(x^n)},$$

where E_r denotes expectation with respect to r. The expectation in the right hand side is the Kullback-Leibler divergence from r to q. The minimax redundancy is $\bar{R}_n(C) \stackrel{\text{def}}{=} \min_q R_n(q, C)$. For this, it is known that the Jeffreys mixture is asymptotically minimax and $\bar{R}_n(C) = (d/2)\log(n/2\pi e) + \log C_J(\Theta) + o(1)$ holds [6, 19]. Note that it holds for general smooth families under certain regularity conditions. In other words, the redundancy of the Jeffreys mixture is asymptotically independent of the curvature of the family C.

As for the definition of the maximum redundancy, note that we can rewrite it as

$$R_n(q, C) \stackrel{\text{def}}{=} \max_{r \in C} E_r \log \frac{r(x^n)}{q(x^n)} = \max_{r \in C} \max_{p \in C} E_r \log \frac{p(x^n)}{q(x^n)}.$$

Here, C in the second maximum is the reference class in the same sense for the definition of regret, while C in the first maximum defines the range of the true information source. These classes do not have to coincide. In particular, interesting is the robust case where the true information source r does not belong to the reference class, i.e. the case that the first C is larger than the second C. Now, we introduce the symbol S to denote the class of true information sources.

Then, we can extend the notion of redundancy to the relative redundancy, i.e. we define

$$RR_n(q, r, C) \stackrel{\text{def}}{=} E_r \log \frac{1}{q(x^n)} - \inf_{p \in C} E_r \log \frac{1}{p(x^n)},$$

which we refer to as the relative redundancy of q with respect to an information source r and a reference class C. We further define the worst case relative redundancy of a strategy q with respect to a class of information sources S and a reference class C as

$$RR_n(q, S, C) \stackrel{\text{def}}{=} \sup_{r \in S} \Big(E_r \log \frac{1}{q(x^n)} - \inf_{p \in C} E_r \log \frac{1}{p(x^n)} \Big).$$

When S equals C, the relative redundancy is reduced to the ordinary redundancy. This definition is relevant to Haussler's *robust PAC* (*Probably Approximately Correct*) *learning model* [7].

The minimax relative redundancy for the pair (S, C) is defined as

$$\bar{RR}_n(S, C) \stackrel{\text{def}}{=} \inf_q \sup_{r \in S} \Big(E_r \log \frac{1}{q(x^n)} - \inf_{p \in C} E_r \log \frac{1}{p(x^n)} \Big).$$

We are interested in the case where S is really larger than C, for example S is a non parametric class.

4.1. **Asymptotic Expansion of Relative Redundancy.** We evaluate the relative redundancy of Jeffreys mixture of the reference class C.

We define $\tilde{u}_r$ for r as

$$\tilde{u}_r \stackrel{\text{def}}{=} \arg\min_{u\in U} E_r \log \frac{1}{p(x|u)}.$$

We can rewrite the relative redundancy as

$$RR_n(q,r,C) = E_r \log \frac{p(x^n|\hat{u})}{q(x^n)} + E_r \log \frac{p(x^n|\tilde{u}_r)}{p(x^n|\hat{u})}. \tag{3}$$

Let us consider the case for $r \in C$. Then for the Jeffreys mixture m_J, the first term is evaluated as

$$E_r \log \frac{p(x^n|\hat{u})}{m_J(x^n)} \sim E_r \log \frac{(2\pi)^{d/2}\sqrt{\det(J(\hat{u}))}}{C_J n^{d/2}\sqrt{\det(\hat{J}(x^n,\hat{u}))}}.$$

When $r \in C$, $\log\sqrt{\det(\hat{J}(x^n,\hat{u}))}$ converges to $\log\sqrt{\det(J(\hat{u}))}$ (almost surely). Hence, the first term does not depend on r.

The second term of (3) is evaluated as

$$E_r \log \frac{p(x^n|\tilde{u}_r)}{p(x^n|\hat{u})} \sim -E_r \frac{\mathrm{tr} J(\tilde{u}_r)^{-1}\hat{I}(x^n,\tilde{u}_r)}{2} = -\frac{\mathrm{tr} J(\tilde{u}_r)^{-1} I(\tilde{u}_r)}{2}, \tag{4}$$

where we let

$$\begin{aligned} \hat{I}_{ij}(x^n,u) &\stackrel{\text{def}}{=} \frac{1}{n}\frac{\partial \log p(x^n|u)}{\partial u^i}\frac{\partial \log p(x^n|u)}{\partial u^j}, \\ I(u) &\stackrel{\text{def}}{=} E_{p(\cdot|u)}\hat{I}_{ij}(x^n,u). \end{aligned}$$

When $r \in C$, $E_u\hat{I}(x^n,u) = J(u)$ holds under certain regularity conditions. Hence, (4) equals $-d/2$. Then the Jeffreys mixture is an asymptotic equalizing rule, which is asymptotically minimax as shown by Clarke and Barron [6].

When $r \notin C$, the situation differs. First, we must care the fact that

$$E_r \log \frac{\det(J(\hat{u}))}{\det(\hat{J}(x^n,\hat{u}))}$$

is not zero in general, caused by the exponential curvature of C. Especially when r is close to $p(\cdot|\tilde{u}_r)$, using Taylor expansion, we have

$$E_r \log \frac{\det(J(\hat{u}))}{\det(\hat{J}(x^n,\hat{u}))} \sim -\mathrm{tr} J(\tilde{u}_r)^{-1}(\underbrace{E_r\hat{J}(x^n,\tilde{u}_r) - J(\tilde{u}_r)}_{\text{e-curvature}}),$$

where the second factor is exponential curvature along with the direction of the mixture geodesic connecting r and $p(\cdot|\tilde{u}_r)$ at $p(\cdot|\tilde{u}_r)$. As for (4), we have

$$\begin{aligned} & \mathrm{tr} J(\tilde{u}_r)^{-1} E_r\hat{I}(x^n,\tilde{u}_r) \\ = \; & \mathrm{tr} J(\tilde{u}_r)^{-1}(E_r\hat{I}(x^n,\tilde{u}_r) - J(\tilde{u}_r) + J(\tilde{u}_r)) \\ = \; & d + \mathrm{tr} J(\tilde{u}_r)^{-1}(\underbrace{E_r(\hat{I}(x^n,\tilde{u}_r) - \hat{J}(x^n,\tilde{u}_r))}_{\text{m-curvature}} + \underbrace{E_r(\hat{J}(x^n,\tilde{u}_r) - J(\tilde{u}_r))}_{\text{e-curvature}}). \end{aligned}$$

Hence, $RR_n(m_J,r,C)$ is generally affected by both exponential and mixture curvatures, but when r is close to $p(\cdot|\tilde{u}_r)$, the influence of exponential curvature almost vanishes.

Similarly, we can analyze relative redundancy of the normalized maximum likelihood $\hat{m}$ (for C). For that case, since the first term of (3) does not depend on r, the fluctuation of $RR_n(\hat{m}, r, C)$ from the minimax redundancy consists of both exponential and mixture curvatures of C at $p(\cdot|u_r)$.

4.2. **Minimax Strategy for Relative Redundancy.** Similarly as the case of regret, we can propose a possible minimax strategy for relative redundancy. First define a d^2-dimensional vector valued random variable $V(x,\theta)$ as $V_{dj+i}(x^n, u) \stackrel{\text{def}}{=} \hat{I}_{ij}(x^n, u) - J_{ij}(u)$. Also define a d^2-dimensional vector valued random variable $U_{dj+i}(x^n, u) \stackrel{\text{def}}{=} \hat{J}_{ij}(x^n, u) - J_{ij}(u)$. Using these random variables, we define an extended class $\bar{C}$ of the reference class C:

$$\bar{C} \stackrel{\text{def}}{=} \{p_e(\cdot|u,\xi,\omega) = \frac{p(\cdot|u)e^{\xi \cdot U(\cdot,u) + \omega \cdot V(\cdot,u)}}{\lambda(u,\xi,\omega)} : u \in U,\ |\xi,\omega| \leq c_1\},$$

where $\lambda(u,\xi,\omega)$ is the normalization constant. Note that $\bar{C}$ is $(d+2d^2)$-dimensional and the original class C is smoothly embedded in the enlarged class $\bar{C}$. Then, we can conjecture that the following mixture is asymptotically minimax for relative redundancy under appropriate conditions, where we assume $\epsilon_n = o(1)$.

$$\bar{m}_n(x^n) \stackrel{\text{def}}{=} (1-\epsilon_n) m_J(x^n) + \epsilon_n \int p_e(x^n|u,\xi,\omega) w(u,\xi,\omega) du d\xi d\omega.$$

If $\mathcal{X}$ is a finite set, we can use the whole space formed by all probability mass functions on $\mathcal{X}$ as the enlarged class $\bar{C}$.

5. Curvature of Tree Model

In the field of source coding, Markov model is important. For example, CONTEXT algorithm [9] and the Context Tree Weighting method [18] employ a kind of Markov models, which is referred to as tree models. Here, we discuss about the exponential curvature of tree models. This is an important topic from view point of stochastic complexity.

A tree model is a parametric model of Markov sources, defined using tree structure of contexts (we refer to a last subsequence of a data sequence as its context). It is well known that a family which consists of all Markov chains of fixed order forms an exponential family. A tree model is a subspace of such an exponential family, hence in general, it forms a curved exponential family. On the other hand, it is known that a parametric class of Markov sources defined by a finite state machine (automaton), which is referred to as an FSMX model, is an exponential family. As for a tree model, there are two cases: 1) It is equivalent to an finite state machine. 2) It is not. For the case 1), the tree model is an exponential family, while for the case 2), it is unknown if it is an exponential family or not. We would like to study this matter.

Let us give a formal definition of the tree model. Suppose $\mathcal{X} = \{0, 1, 2, \cdots, m\}$ $(m \geq 1)$. Let $\mathcal{X}'$ denote $\mathcal{X} \backslash \{0\}$. Let T be a finite subset of $\mathcal{X}^* \stackrel{\text{def}}{=} \{\lambda\} \cup \mathcal{X} \cup \mathcal{X}^2 \cup \cdots$. When, 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 (the null sequence)), T is called a context tree. Define

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

Each element of ∂T is referred to as a leaf of T or a context. We refer to $L(T) \stackrel{\text{def}}{=} \max_{s \in \partial T} |s|$ as depth of T, where $|s|$ denotes a length of s.

For $x^i = x_1x_2, ..., x_i$, let $s(x^i)$ denote a postfix of x^i which belongs to ∂T. If $i \geq L(T)$, $s(x^i)$ is uniquely determined. Here, $s(x^i)$ is referred to as the context of x^i defined by T. By putting a probability distribution of $x \in \mathcal{X}$ to every element of

∂T, we can define a Markov source. It is referred to as a *tree source* [17]. For a tree T, assume that the context of sx for every $s \in \partial T$ and every $x \in \mathcal{X}$ is determined by s and x (even if $|sx| < L(T)$). Then the tree source defined by T is referred to as an FSMX source. This is equivalent to the condition that there exists a state transition function τ which maps a pair (s, x) to the context of sx defined by T. Hereafter, we assume $\mathcal{X} = \{0, 1\}$ (binary case). We give examples of the context tree in Figures 1 and 2. There exists a state transition function for T_{e1}, while there does not for T_{e2}.

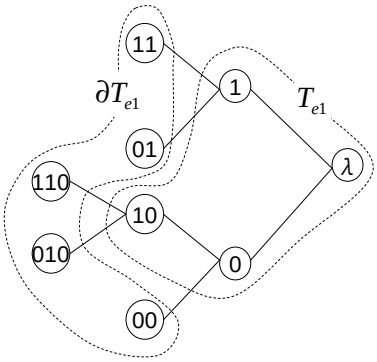

FIGURE 1. A Context Tree for an FSMX source

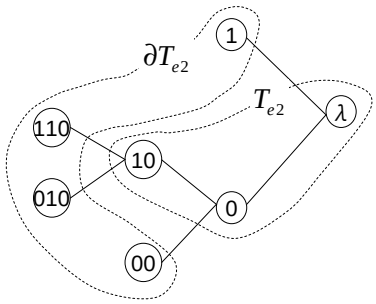

FIGURE 2. A Context Tree for a tree source

Now, we formally define a tree source. We let $x_i^j \stackrel{\text{def}}{=} x_i x_{i+1} \dots x_j$ for $i \leq j$ and L denote $L(T)$, hereafter. We define the probability mass function given an initial sequence x_{-L+1}^0 as follows

$$p(x^n | x_{-L+1}^0, h, T) = \prod_{i=0}^{n-1} h_{s(x_{-L+1}^i)}^{x_{i+1}},$$

where h_s^x denotes the probability that x is produced at the context s, i.e. $h_s^x \geq 0$ and $\sum_{x \in \mathcal{X}} h_s^x = 1$ are assumed. Let h be the $(|\mathcal{X}| - 1) \cdot |\partial T|$-dimensional vector whose components are h_s^a ($s \in \partial T$, $a \in \mathcal{X}'$) and $H(T)$ denote the range of h. The stochastic process $p(x^n | x_{-L+1}^0, h, T)$ (with a fixed h) is referred to as a tree source. An tree model $M(T)$ is a parametric class of tree sources as defined by

$$M(T) \stackrel{\text{def}}{=} \{p(\cdot|\cdot, h, T) : h \in H(T)\}.$$

Note that $M(T_{e2})$ is a subspace of $M(T_{e1})$, which is obtained by putting restriction $h_{11}^1 = h_{01}^1$ on $M(T_{e1})$.

We can easily show that an FSMX model is an exponential family (in asymptotic sense), by proving the difference between the empirical Fisher information and the Fisher information converges to 0 as n goes to infinity. As for a non FSMX tree model, the same inference does not work. We conjecture that all non FSMX models are really curved in terms of exponential curvature. In other words, we conjecture that the following two statements are equivalent: a) A tree model has a state transition function, b) A tree model forms an exponential family (in asymptotic sense). We have not yet proven the conjecture, but the following was shown [15].

Lemma 1. *Assume that a context tree T has a state transition function and that the context tree T' which is obtained by removing a parent node s of nodes $1s, 0s \in \partial T$ does not have a state transition function. Then, $M(T')$ is not an exponential family, even in asymptotic sense.*

This Lemma claims that a certain kind of tree model is really curved in terms of exponential curvature. Here, T_{e1} and T_{e2} (in Figures 1 and 2) are examples of T and T' in Lemma 1, i.e. $M(T_{e2})$ has non-zero exponential curvature.

References

[1] S. Amari, *Differential-geometrical methods in statistics* (*2nd pr.*), Springer-Verlag, 1990.

[2] S. Amari, "Statistical curvature," *Encyclopedia of Statistical Sciences,* vol. 8, pp. 642-646, Willey & Sons, 1994.

[3] A. R. Barron, J. Rissanen and B. Yu, "The minimum description length principle in coding and modeling," *IEEE trans. Inform. Theory,* 1998.

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

[5] B. Clarke & A. R. Barron, "Information-theoretic asymptotics of Bayes methods," *IEEE trans. Inform. Theory,* Vol. 36. No. 3, pp. 453-471, 1990.

[6] B. Clarke & A. R. Barron, "Jeffreys prior is asymptotically least favorable under entropy risk," *JSPI,* 41:37-60, 1994.

[7] D. Haussler, "Decision theoretic generalizations of the PAC model for neural net and other learning applications," *Inf. and Comp.*, 100(1), pp. 78-150, 1992.

[8] J. Rissanen, "Modeling by shortest data description," *Automatica,* vol. 14, pp. 465-471, 1978.

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

[10] J. Rissanen, "Fisher information and stochastic complexity," *IEEE trans. Inform. Theory,* vol. 40, no. 1, pp. 40-47, 1996.

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

[12] J. Takeuchi & A. R. Barron, "Asymptotically minimax regret for exponential families," *Proc. of the 20th Symposium on Information Theory and Its Applications* (*SITA'97*), pp. 665-668, 1997.

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

[14] J. Takeuchi & A. R. Barron, "Robustly minimax codes for universal data compression," *the 21st Symposium on Information Theory and Its Applications* (*SITA'98*), 1998.

[15] J. Takeuchi & T. Kawabata, "Exponential curvature and Jeffreys mixture prediction strategy for Markov model (in Japanese)," *Proc. of the 7th Workshop on Information-Based Induction Sciences,* 2004.

[16] J. Takeuchi, T. Kawabata, and A. R. Barron, "Properties of Jeffreys mixture for Markov sources," *Proc. of the fourth Workshop on Information-Based Induction Sciences* (*IBIS2001)*, pp. 327-332, 2001.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

Internet Systems Research Laboratories, NEC Corporation, 5-7-1 Shiba, Minato-ku, Tokyo 108-8001, JAPAN

Department of Statistics, Yale University, P.O. Box 208290, New Haven, CT 06520-8290, USA

Department of Information & Communications Engineering, University of Electro-Communications, 1-5-1 Choufugaoka, Chofushi, Tokyo 182-8585, JAPAN