Asymptotically Minimax Regret by Bayes Mixtures for Non-exponential Families

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 problems of data compression, gambling and prediction of a sequence $x^n = x_1x_2...x_n$ from an alphabet $\mathcal{X}$, in terms of regret with respect to various families of probability distributions. It is known that the regret of the Bayes mixture with respect to a general exponential families asymptotically achieves the minimax value when variants of Jeffreys prior are used, under the condition that the maximum likelihood estimate is in the interior of the parameter space. We discuss a modification of Jeffreys prior which has measure outside the given family of densities, to achieve minimax regret with respect to non-exponential type families, e.g. curved exponential families and mixture families. These results also provide characterization of Rissanen's stochastic complexity for those classes.**

I. Introduction

We study the problem of data compression, gambling and prediction of a sequence $x^n = x_1x_2...x_n$ from a certain alphabet $\mathcal{X}$ (not restricted to discrete one), in terms of regret with respect to a general exponential family, a related curved exponential family and also a general smooth family. In particular, we evaluate the regret of the Bayes mixture density and show that it asymptotically achieves their minimax values when variants of Jeffreys prior are used. These results are generalizations of the work by Xie and Barron [20], [21] in the general smooth families.

This paper's concern is the regret of a coding or prediction. This regret 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 $\mathcal{X}^n$. We can also use q for prediction and gambling, that is, its conditionals $q(x_{i+1}|x^i)$ provide a distribution for the coding or prediction of the next symbol given the past. The minimax regret with the target class (of probability densities) $S = \{p(\cdot|\theta) : \theta \in \Theta\}$ and a set of the sequences $W_n \subseteq \mathcal{X}^n$ (denoted by $\bar{r}(W_n)$) is defined as

$$\bar{r}(W_n) = \inf_q \sup_{x^n \in W_n} \Big(\log \frac{1}{q(x^n)} - \log \frac{1}{p(x^n|\hat{\theta}(x^n))} \Big),$$

where $\hat{\theta} = \hat{\theta}(x^n)$ is the maximum likelihood estimate of θ given x^n. Here, the regret $\log(1/q(x^n)) - \log(1/p(x^n|\hat{\theta}))$ in the data compression context is also called the (pointwise) redundancy: the difference between the code length based on q and the minimum of the codelength $\log(1/p(x^n|\theta))$ achieved by distributions in the family. Also, $\log(1/q(x^n)) - \log(1/p(x^n|\theta))$ is the sum of the incremental regrets of prediction $\log(1/q(x_{i+1}|x^i)) - \log(1/p(x_{i+1}|x^i,\theta))$. In this paper, we consider minimax problems for sets of sequences such that $W_n = \mathcal{X}^n(\mathcal{G}) = \{x^n : \hat{\theta} \in \mathcal{G}\}$, where $\mathcal{G}$ is a certain nice subset (satisfies $\bar{\mathcal{G}} = \bar{\mathcal{G}^\circ}$) of Θ.

When S is the class of discrete memoryless sources, Xie and Barron [21] proved that the minimax regret asymptotically equals $(d/2)\log(n/2\pi) + \log C_J(\mathcal{G}) + o(1)$, where d equals the size of alphabet minus 1 and $C_J(\mathcal{G})$ is the integral of the square root of the determinant of Fisher information matrix over $\mathcal{G}$. In this paper, we discuss the generalization of the results of [21] to the case where S is an exponential family or the related curved exponential family.

For multi-dimensional exponential families, variants of Jeffreys mixture are minimax, when $\mathcal{G}$ is a compact set included in the interior of Θ. For curved exponential families, the ordinary Jeffreys mixture for the concerned curved family is not minimax, even if $\mathcal{G}$ is a compact set included in the interior of Θ. However, we can obtain the minimax result by using a sequence of prior measures whose supports are the exponential family to which the curved family is embedded, rather than the concerned curved family. It is remarkable that this idea is applicable to general smooth families by the enlargement of the original family using exponential tilting. The idea for this enlargement in addressing minimax regret originates in preliminarily form in [17], [4] as informally discussed in [3]. The literature [18] gives discussion in the context of Amari's information geometry [2]. In this paper, we discuss the formal regularity conditions we assume and the relation between the method for curved exponential families and that for general smooth families. In particular, we show that our strategy works for general mixture families.

To obtain the above minimax results, we employ the Laplace integration method, which was used by Clarke and Barron [6], [7] in order to evaluate the expected regret of the Bayes procedures. Especially in [7], they succeeded to uniformly evaluate the expected regret by the Laplace integration for a compact subset $\mathcal{G}$ of Θ°.

The normalized maximum likelihood is an alternative way to obtain the minimax regret, which is defined as

$$\hat{m}_n(x^n) = \frac{p(x^n|\hat{\theta})}{\int_{W_n} p(x^n|\hat{\theta})dx^n}.$$

This is known to be strictly minimax, but it is difficult to calculate its conditionals (important for prediction prob-

lem and data compression algorithm) $\hat{m}_N(x_n|x^{n-1}) = \hat{m}_N(x^n)/\hat{m}_N(x^{n-1})$ (assuming $n \leq N$). On the other hand, we can obtain the conditionals of Bayes mixture by the integration $m_N(x_n|x^{n-1}) = \int p(x_n|\theta) w_N(d\theta|x^{n-1})$, $w_N(d\theta|x^{n-1})$ denote the posterior measure of θ given x^{n-1}.

II. Preliminaries

Let $(\mathcal{X}, \mathcal{B}, \nu)$ be a measurable space with a reference measure ν. Let $S = \{p(\cdot|\theta) : \theta \in \Theta\}$ denote a parametric family of probability densities over $\mathcal{X}$ with respect to ν. We let $p(x^n|\theta)$ denote $\prod_{i=1}^n p(x_i|\theta)$. Also, we let $\nu(dx^n)$ denote $\prod_{i=1}^n \nu(dx_i)$. Here, we are treating models for independently identically distributed (i.i.d.) random variables. We let P_θ denote the distribution function with density $p(\cdot|\theta)$ and E_θ denote expectation with respect to P_θ.

Assume that $\Theta \subseteq \Re^d$ and $\bar{\Theta} = \bar{\Theta^\circ}$ hold. That is, the closure of Θ matches the closure of its interior. Here $\bar{A}$ and A° respectively denote the closure and the interior of $A \subseteq \Re^k$.

We introduce the empirical Fisher information given x^n and the Fisher information:

$$\begin{aligned} \hat{J}_{ij}(\theta) &= \hat{J}_{ij}(\theta, x^n) = \frac{-1}{n} \frac{\partial^2 \log p(x^n|\theta)}{\partial\theta^i \partial\theta^j}, \\ J(\theta) &= E_\theta \hat{J}_{ij}(\theta, x^n). \end{aligned}$$

The exponential family is defined as follows. [5], [1], [2]

Definition 1 (Exponential Family): Given a Borel measurable function $T : \mathcal{X} \to \Re^d$, define

$$\Theta_a \equiv \Big\{\theta : \theta \in \Re^d, \int_{\mathcal{X}} \exp\big(\theta \cdot T(x)\big)\nu(dx) < \infty\Big\},$$

where $\theta \cdot T(x)$ denotes the inner product of θ and $T(x)$. Define a function ψ and a probability density p on $\mathcal{X}$ with respect to ν by $\psi(\theta) \equiv \log \int_{\mathcal{X}} \exp(\theta \cdot T(x))\nu(dx)$ and $p(x|\theta) \equiv \exp\big(\theta \cdot T(x) - \psi(\theta)\big)$. We refer to the set $S(\Theta) \equiv \{p(x|\theta)|\theta \in \Theta \subseteq \Theta_a\}$ as an exponential family of densities.

When Θ_a is an open set, $S(\Theta)$ is said to be a regular exponential family. Many popular exponential families are regular. We let $J(\theta)$ denote Fisher information matrix of θ. For exponential families, the components of J are given by

$$J_{ij}(\theta) = \frac{\partial^2 \psi(\theta)}{\partial\theta^i \partial\theta^j}. \tag{1}$$

For regular exponential families, define expectation parameter η as $\eta(\theta) = E_\theta(T(x))$. It is known that the map : $\theta \mapsto \eta$ is one-to-one and analytic on Θ_a. Also, $\eta_i = \partial\psi(\theta)/\partial\theta^i$ holds. We also use notation $\theta(\eta)$ as inverse function of $\eta(\theta)$. Note that $p(x^n|\theta) = \exp(n(\theta \cdot \bar{t} - \psi(\theta)))$ holds, where $\bar{t} = \sum_{t=1}^n T(x_t)/n$. ($x_t$ denotes the t-th element of sequence $x^n = x_1 x_2 ... x_n$). It is known that the maximum likelihood estimate of η given x^n equals $\bar{t}$.

For a subset $\mathcal{G}$ of Θ, we let $C_J(\mathcal{G}) = \int_{\mathcal{G}} |J(\theta)|^{1/2} d\theta$. The Jeffreys prior ([11]) over $\mathcal{G}$ (denoted by $w_{\mathcal{G}}(\theta)$) is defined as

$$w_{\mathcal{G}}(\theta) = \frac{|J(\theta)|^{1/2}}{C_J(\mathcal{G})}.$$

We define the Jeffreys mixture for $\mathcal{G}$ (denoted by $m_{\mathcal{G}}$) as $\int_{\mathcal{G}} p(x^n|\theta) w_{\mathcal{G}}(\theta) d\theta$.

We introduce the curved exponential family. Let $S_e = \{p_e(x^n|u) : u \in \mathcal{U}\}$ be the $\bar{d}$-dimensional exponential family. Using a smooth function $\phi : \Re^d \to \Re^{\bar{d}}$, we define a subfamily of S as follows:

$$S_c = \{p_c(\cdot|\theta) = p_e(\cdot|\phi(\theta)) : \theta \in \Theta\},$$

where Θ is an open set of $\Re^d$ and $\bar{d} \geq d$. This S_c is referred to as a curved exponential family embedded in S_e. We let $\hat{\theta}$ denote the maximum likelihood estimate of θ given x^n:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} p_c(x^n|\theta).$$

Definition 2 (Mixture Family): For $i = 0, 1, \ldots, d$, let $p_i(x)$ be a probability density function over $\mathcal{X}$. Define

$$p(x|\theta) = \sum_{i=0}^d \theta_i p_i(x),$$

where $\theta \in \Theta_a = \{\theta \in \boldsymbol{R}^d : 0 \leq \sum_{i=1}^d \theta_i \leq 1 \text{ and } \forall i, \theta_i \geq 0\}$ and $\theta_0 = 1 - \sum_{i=1}^d \theta_i$. Then, the set $\{p(\cdot|\theta) : \theta \in \Theta \subseteq \Theta_a\}$ is referred to as a mixture family of densities.

III. Lower Bound on Minimax Regret

Here, we give a lower bound on minimax regret with the target class being a general smooth family. We employ the assumptions described below.

Assumption 1: The density $p(x|\theta)$ is twice continuously differentiable in θ for all x, and there is a function of $\delta(\theta) > 0$ so that for each i, j,

$$E_\theta \sup_{\theta' : |\theta' - \theta| \leq \delta(\theta)} |\hat{J}(\theta', x)|^2$$

is finite and continuous as a function of θ.

Assumption 2: The Fisher information $J(\theta)$ is continuous and coincides with the matrix of which the (i, j)-entry is

$$E_\theta \frac{\partial \log p(x|\theta)}{\partial\theta^i} \frac{\partial \log p(x|\theta)}{\partial\theta^j}.$$

Assumption 3: For all $\theta, \theta' \in \Theta$, the Kullback Leibler divergence $D(\theta|\theta')$ is finite and for an arbitrary $\epsilon > 0$, the following holds.

$$\inf_{(\theta,\theta') : |\theta - \theta'| > \epsilon} D(\theta|\theta') > 0$$

Assumption 4: For an arbitrary compact set $K \subseteq \Theta^\circ$, the MLE is uniformly consistent in K.

$$\sup_{\theta \in K} P_\theta(|\hat{\theta}(x^n) - \theta| > \epsilon) = o(1/\log n).$$

Remark: These 4 assumptions hold for regular exponential families and appropriately defined mixture families.

We can prove the following.

Theorem 1: Let $S = \{p(\cdot|\theta) : \theta \in \Theta\}$ be a d-dimensional family of probability densities. We suppose that Assumptions 1-4 hold for S. We let K be an arbitrary subset of Θ satisfying $C_J(K) < \infty$ and $\bar{K} = \bar{K^\circ}$. The following holds.

$$\liminf_{n \to \infty} (\bar{r}_n(\mathcal{X}^n(K)) - \frac{d}{2} \log \frac{n}{2\pi}) \geq \log C_J(K).$$

We omit the proof in this paper.

IV. Minimax Bayes Procedures

Below we describe the asymptotically minimax Bayes procedures for each case we considered.

A. Exponential Families

We are interested in the regret of mixture strategies. The Jeffreys mixture is the mixture by the prior proportional to $|J(\theta)|^{1/2}$. We denote the Jeffreys mixture over a subset K of Θ by m_n. The value $C_J(\mathcal{G})$ is the normalization constant for the Jeffreys prior over the set $\mathcal{G}$.

For the exponential family including the multinomial Bernoulli and FSM (Finite State Machine), it is known that a sequence of Jeffreys mixtures achieves the minimax regret asymptotically [21], [16], [17], [19]. For the multinomial exponential family case except for multinomial Bernoulli and FSM, these facts are proven under the condition that $\mathcal{G}$ is a compact subset included in the interior of Θ.

We briefly review outline of the proof for that case. Let $\{\mathcal{G}_n\}$ be a sequence of subsets of Θ such that $\mathcal{G}_n^\circ \supset \mathcal{G}$. Suppose that $\mathcal{G}_n$ reduces to $\mathcal{G}$ 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

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

where the remainder $o(1)$ tends to zero uniformly over all sequences with MLE in $\mathcal{G}$. 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:

$$\frac{m_{J,n}(x^n)}{p(x^n|\hat{\theta})} \sim \frac{|J(\hat{\theta})|^{1/2}}{C_J(\mathcal{G})|\hat{J}(\hat{\theta},x^n)|^{1/2}}\frac{(2\pi)^{d/2}}{n^{d/2}}.$$

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

B. Curved Exponential Families

When the model S 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 $\hat{J}(\theta,x^n) - J(\theta)$ orthogonal (in terms of Fisher metric) to the model S is its embedding exponential curvature.

When the target class is a curved exponential family, it is easy to see this fact. Assume that S_c is embedded in a $\bar{d}$-dimensional exponential family with the natural parameter $u \in \mathcal{U} \subset \Re^{\bar{d}}$ ($\bar{d} > d$):

$$S_e = \{p_e(x|u) = \exp(u\cdot T(x) - \psi(u)) : u \in \mathcal{U}\}.$$

That is, we let

$$p_c(\cdot|\theta) = p_e(\cdot|\phi(\theta)),$$

where ϕ is a (4 times differentiable) function $\Theta \to \mathcal{U}$. Then

$$\begin{aligned}\hat{J}_{ij}(\theta,x^n) &= -\frac{\partial^2\phi(\theta)}{\partial\theta_i\partial\theta_j}\cdot(\bar{t}-\eta(\theta)) \\ &+ \sum_{k,l=1}^{\bar{d}}\frac{\partial\phi_k(\theta)}{\partial\theta_i}\frac{\partial\phi_l(\theta)}{\partial\theta_j}\cdot\frac{\partial^2\psi(u)}{\partial u_k\partial u_l}\Big|_{u=\phi(\theta)}.\end{aligned}$$

By taking expectation of both sides, we can see the last term is the Fisher information of $J_{ij}(\theta)$. Hence we have

$$\hat{J}_{ij}(\hat{\theta},x^n) = -\frac{\partial^2\phi(\hat{\theta})}{\partial\theta_i\partial\theta_j}\cdot(\bar{t}-\eta(\hat{\theta})) + J_{ij}(\hat{\theta}), \tag{3}$$

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

First, assume that S_c is not curved in S_e (in the natural parameter space), then $u = \phi(\theta)$ forms a plane in $\mathcal{U}$, i.e. the vectors $\partial^2\phi/\partial\theta_i\partial\theta_j$ $(i,j = 1,...,d)$ are certain linear combinations of the vectors $\partial\phi/\partial\theta_i$ $(i = 1,...,d)$. Then noting

$$\frac{\partial\phi}{\partial\theta_i}\Big|_{\theta=\hat{\theta}}\cdot(\bar{t}-\eta(\hat{\theta})) = 0,$$

we have

$$\frac{\partial^2\phi}{\partial\theta_i\partial\theta_j}\Big|_{\theta=\hat{\theta}}\cdot(\bar{t}-\eta(\hat{\theta})) = 0.$$

This implies $\hat{J}(\hat{\theta},x^n) = J(\hat{\theta})$. On the other hand, this is not guaranteed when S is curved.

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.

$$\bar{m}_n(x^n) = (1-n^{-r})m_{J,n}(x^n) + n^{-r}\int p_e(x^n|u)w(u)du, \tag{4}$$

where $m_{J,n}$ is the Jeffreys mixture over $\mathcal{G}_n \supset \mathcal{G}$ and $w(u)$ is a certain probability density on $\mathcal{U}$.

This can be shown as follows. Let $G_{n,\delta} = \{x^n | |\bar{x}-\eta(\hat{\theta})| \le \delta$ and $\hat{\theta}\in\mathcal{G}\}$ and $G_{n,\delta}^c = \{x^n | |\bar{x}-\eta(\hat{\theta})| > \delta$ and $\hat{\theta}\in\mathcal{G}\}$. If $x^n \in G_{n,\delta}$, then $||J(\hat{\theta}) - \hat{J}(\hat{\theta},x^n)|| \le B\delta$ holds, where B is a certain positive number determined by the function ϕ and the subspace $\mathcal{G}$. Since $J(\theta)$ is continuous and $\mathcal{G}$ is compact, the minimum eigenvalue of $J(\theta)$ is bounded below by a positive number for all $\theta\in\mathcal{G}$. Hence we can show

$$\sup_{x^n\in G_{n,\delta}}\frac{|\hat{J}(\hat{\theta})|}{|J(\hat{\theta})|} \le 1 + C\delta,$$

where C is a certain real number. Then by the Laplace integration we have

$$\begin{aligned}\inf_{x^n\in G_{n,\delta}}\frac{\bar{m}_n(\hat{\theta})}{p(x^n|\hat{\theta})} &\ge \inf_{x^n\in G_{n,\delta}}\frac{(1-n^{-r})m_n(x^n)}{p(x^n|\hat{\theta})} \\ &\ge \frac{(1-n^{-r})(1+o(1))(2\pi)^{d/2}}{C_J(\mathcal{G}_n)(1+C\delta)n^{d/2}},\end{aligned} \tag{5}$$

where $o(1)$ is a quantity converging to 0 as n goes to infinity.

To handle the sequence $x^n \in G_{n,\delta}^c$, let $\tilde{\eta}$ be the point between $\bar{x}$ and $\eta(\hat{\theta})$ such that $|\tilde{\eta}-\eta(\hat{\theta})| = \delta$. Then we have

[1] If $\mathcal{G}$ is the entire space for the statistical model, we cannot define the superset of $\mathcal{G}$ 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 [21], [17], [19].

$D(\bar{p}(\cdot|\tilde{u})|p(\cdot|\hat{\theta})) \geq A\delta^2$, where $D(p|q)$ denotes Kullback-Leibler divergence from p to q, $\tilde{u}$ the u corresponding to $\tilde{\eta}$, and A a certain positive number determined by $\bar{S}$ and K. From this, we can show

$$\frac{1}{n}\log\frac{\bar{p}(x^n|\tilde{u})}{p(x^n|\hat{\theta})} > A\delta^2.$$

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

$$\inf_{x^n\in G^c_{n,\delta}}\frac{\bar{m}_n(\hat{\theta})}{p(x^n|\hat{\theta})} \geq \frac{n^{-r}\int\bar{p}(x^n|u)w(u)d\theta}{p(x^n|\hat{\theta})} > n^{-r-d}e^{A\delta^2 n},$$

where we have evaluated $\int\bar{p}(x^n|u)w(u)d\theta$ by the integration in the $n^{-1/2}$-neighborhood around $\tilde{u}$. With letting $\delta = n^{-b}$ with $0 < b < 1/2$, together with (5), we have

$$\inf_{x^n:\hat{\theta}\in\mathcal{G}}\frac{\bar{m}_n(\hat{\theta})}{p(x^n|\hat{\theta})} \geq \frac{(1+o(1))(2\pi)^{d/2}}{C_J(K)n^{d/2}},$$

which yields an upper bound on the minimax regret.

Theorem 2: For a curved exponential family S_c, the following holds.

$$\inf_{x^n:\hat{\theta}\in\mathcal{G}}\frac{\bar{m}_n(x^n)}{p(x^n|\hat{\theta})} \geq \frac{(1+o(1))(2\pi)^{d/2}}{C_J(\mathcal{G})n^{d/2}}.$$

C. General Smooth Families

For more general smooth families we form a direct enlargement by a exponential tilting using linear combinations of the entries of the differences $V(x^n|\theta) = \hat{J}(\theta) - J(\theta)$. Let $\mathcal{B} = (-b/2, b/2)^{d\times d}$ for some $b > 0$. The enlargement is formed as

$$\bar{p}(x^n|u) = p(x^n|\theta)e^{n(\beta\cdot V(x^n|\theta)-\psi_n(\theta,\beta))}, \tag{6}$$

where u denotes the pair (θ,β), β is a matrix in $\mathcal{B}$, and $V(x^n|\theta)\cdot\beta$ denotes $\mathrm{Tr}(V(x^n|\theta)\beta^t) = \sum_{ij}V_{ij}(x^n|\theta)\beta^t)\beta_{ij}$. Then, we define the model $\bar{S} = \{\bar{p}(\cdot|u) : u \in \Theta\times\mathcal{B}\}$.

Traditionally such a family arises as in local asymptotic expansion of likelihood ratio, evaluated at a perturbation $\theta+\beta$ of a given θ, used in demonstration of *local asymptotic normality*. [10], [12], [13] In Amari's information geometry [1], [2] it is a local exponential family boundle.

The enlarged model $\bar{S}$ plays a role of the outside exponential family in the curved exponential family case, i.e. we employ the mixtures

$$\bar{m}_n(x^n) = (1-n^{-r})m_{J,n}(x^n) + n^{-r}\int\bar{p}(x^n|u)w(u)du. \tag{7}$$

Specifically, the prior $w(u)$ for $\bar{S}$ is defined as the direct product of the Jeffreys prior on S and the uniform prior on $\mathcal{B}$.

Here $\psi_n(\theta,\beta)$ is the logarithm of the required normalization factor, so that $p(x^n|\theta,\beta)$ sums (integrates with respect to ν^n) to the value 1 for every $\theta\in\mathcal{G}$ and every β in a neighborhood around 0.

In the analysis, the consideration of β in a neighborhood of a small multiple of $V(x^n|\theta) = \hat{J}(\hat{\theta}) - J(\hat{\theta})$ is sufficient to accomplish our objectives under the assumptions addressed below.

Assumption 5: For a certain $C_1 > 0$, the following holds.

$$\forall n\in\mathbb{N},\ \forall\theta\in\mathcal{G},\ \forall\beta\in\mathcal{B}, \tag{8}$$
$$\Big(\int p(x^n|\theta)\exp(nV(x^n|\theta)\cdot\beta)\nu_n(dx^n)\Big)^{1/n} < C_1.$$

Define a function $\psi_n(\theta,\beta)$ as

$$\psi_n(\theta,\beta) \stackrel{\text{def}}{=} \frac{1}{n}\log\int p(x^n|\theta)\exp(nV(x^n|\theta)\cdot\beta)\nu_n(dx^n).$$

Define a set of good sequences ${G'}_{n,\delta}$ and a set of not good sequences ${G'}^c_{n,\delta}$ similarly as for the curved exponential family case.

$$\begin{aligned}{G'}_{n,\delta} &= \{x^n : ||V(x^n|\hat{\theta})|| \leq \delta \text{ and } \hat{\theta}\in\mathcal{G}\},\\ {G'}^c_{n,\delta} &= \{x^n : ||V(x^n|\hat{\theta})|| > \delta \text{ and } \hat{\theta}\in\mathcal{G}\},\end{aligned}$$

where $||A||$ for $A\in\Re^{d\times d}$ denotes the Frobenius norm defined as $||A|| = (\mathrm{Tr}(AA^t))^{1/2}$.

We define the two neighborhoods of θ' as

$$B_\epsilon(\theta') = \{\theta : (\theta-\theta')^t J(\theta')(\theta-\theta') \leq \epsilon^2\}, \tag{9}$$
$$\hat{B}_\epsilon(\theta') = \{\theta : (\theta-\theta')^t \hat{J}(\theta')(\theta-\theta') \leq \epsilon^2\}. \tag{10}$$

Assumption 6: We assume a kind of equi-semicontinuity for $\hat{J}(\theta)$, that is, there exist a $\kappa > 0$ and a $\delta_0 > 0$ such that for all small $\epsilon > 0$, for all x^n in ${G'}_{n,\delta_0}$, for all $\tilde{\theta}\in\hat{B}_\epsilon(\hat{\theta})$, and for all $\theta\neq\hat{\theta}$,

$$\frac{(\theta-\hat{\theta})^t\hat{J}(\tilde{\theta})(\theta-\hat{\theta})}{(\theta-\hat{\theta})^t\hat{J}(\hat{\theta})(\theta-\hat{\theta})} \leq 1+\kappa\epsilon.$$

This is used to control the Laplace integration for m_n of our strategy for the good sequences. In fact, we can prove the following lemma.

Lemma: 1: Under Assumptions 1-3, 5, and 6, for all $\delta < \delta_0$, the following holds.

$$\inf_{x^n\in{G'}_{n,\delta}}\frac{\bar{m}_n(x^n)}{p(x^n|\hat{\theta})} \geq \frac{(1+o(1))(1-n^{-r})}{(1+\zeta\delta)^{d/2}}\frac{(2\pi)^{d/2}}{C(\mathcal{G}_n)n^{d/2}},$$

where $\zeta > 0$ is determined by $J(\theta)$ and K.
The proof is done by the Laplace integration, which causes the error term of $o(1)$.

Assumption 7: There exists an $\epsilon > 0$, such that for all x^n in $G^c_{n,\delta}$, for all $\tilde{\theta}$ in $N_\epsilon(\hat{\theta})$, the following holds

$$||V(x^n|\tilde{\theta})|| \geq ||V(x^n|\hat{\theta})||/2.$$

Assumption 8: There exists an $\epsilon > 0$, such that for all x^n in $G^c_{n,\delta}$, and for all $\tilde{\theta}\in N_\epsilon(\hat{\theta})$, $2\hat{J}(\hat{\theta}) - \hat{J}(\tilde{\theta})$ is semi-positive definite.

These two assumptions are used to control the second term of our strategy for the not good sequences. They require that $\hat{J}(\theta)$ does not change so rapidly in the region for the integration. We can prove the following lemma.

Lemma: 2: Under Assumptions 1-3, 5, 7, and 8, the following holds

$$\inf_{x^n \in G'^c_{n,\delta}} \frac{\bar{m}_n(x^n)}{p(x^n|\hat{\theta})} \geq n^{-r-d} \exp(A\delta^2 n),$$

where A is a certain positive constant.

Letting $\delta = n^{-b}$ with $0 < b < 1/2$, by Lemmas 1 and 2, we can show the following Theorem.

Theorem 3: Under Assumptions 1-3 and 5-8,

$$\inf_{x^n : \hat{\theta} \in \mathcal{G}} \frac{\bar{m}_n(x^n)}{p(x^n|\hat{\theta})} \geq \frac{(1+o(1))(2\pi)^{d/2}}{C_J(\mathcal{G}) n^{d/2}}.$$

V. General Strategy applied to Curved Exponential Families

To understand the property of our strategy for general smooth families, we examine Assumptions 6-8 for the curved exponential family. First assume that the range of $T(x)$ is unbounded. For the curved exponential family by (3), we have

$$V_{ij}(x^n|\hat{\theta}) = -\frac{\partial^2 \phi(\hat{\theta})}{\partial \theta_i \partial \theta_j} \cdot (\bar{t} - \eta(\hat{\theta})).$$

This implies that derivatives of $V(x^n|\theta)$ at $\theta = \hat{\theta}$ can be arbitrarily large even when $||V(x^n|\hat{\theta})|| = 0$. To understand it, assume that $\partial^2\phi/\partial\theta_i\partial\theta_j = 0$ and higher order derivatives are not 0 at $\theta = \hat{\theta}$, and note that $|\bar{t} - \eta(\hat{\theta})|$ can be arbitrarily large, since the range of $T(x)$ is unbounded. Then, Assumption 6 does not hold when the range of $T(x)$ is unbounded.

This consideration shows that the general strategy does not work for this situation, although the mixture (4) properly works. One reason is in the difference between $G_{n,\delta}$ and ${G'}_{n,\delta}$. The former is defined in terms of $|\bar{t} - \eta(\hat{\theta})|$, while the latter is in terms of $|\hat{J}(\hat{\theta}) - J(\hat{\theta})|$, Important is that small $|\hat{J}(\hat{\theta}) - J(\hat{\theta})|$ does not imply small $|\bar{t} - \eta(\hat{\theta})|$,

We should note the difference between the enlargement model $\bar{p_c}(\cdot|\theta, \beta)$ and the outside exponential family $p_e(\cdot|u)$. In the situation we are considering, $\bar{p_c}(\cdot|\theta, \beta)$ does not extend to the direction of $\bar{t} - \eta(\hat{\theta})$ in the outside exponential family, which is needed to obtain higher likelihood of the mixture.

Note that this drawback is avoided if the range of $T(x)$ is bounded. In that case, $V(x^n|\theta)$ is equi-continuous for all x^n. Then, Assumptions 5-8 are satisfied.

Another case in which the general method works for an curved exponential family is that ϕ forms a quadratic hyper surface in the u-space of the outside exponential family. Suppose for example that $\phi(\theta) = (\theta, \theta^2)$ for $\bar{d} = 2$ and $d = 1$ case. Then, the derivative of $V(x^n|\theta)$ does not depend on $\bar{t}$, hence $V(x^n|\theta)$ is equi-continuous.

VI. Mixture Families

For the mixture family,

$$\frac{\partial^2 \log p(x|\theta)}{\partial\theta_i \partial\theta_j} = \frac{-(p_i(x) - p_0(x))(p_j(x) - p_0(x))}{(p(x|\theta))^2}.$$

holds, where we have

$$\Big|\frac{p_i(x) - p_0(x)}{p(x|\theta)}\Big| \leq \frac{p_i(x) + p_0(x)}{\sum_{i=0}^d \theta_i p_i(x)} \leq \frac{1}{\theta_i} + \frac{1}{\theta_0}.$$

The last expression is not more than $2\max_{\theta\in\mathcal{G}} \max_i \theta_i^{-1}$, which is finite since $\mathcal{G}$ is a compact set interior to the whole parameter set. Then, $\hat{J}(\theta, x^n)$ is bounded and Assumption 5 holds. Further, $\partial V(x^n|\theta)/\partial\theta_i$ is similarly bounded, so $V(x^n|\theta)$ is equi-continuous for all $x^n : \hat{\theta} \in \mathcal{G}$ and Assumption 6-8 hold. This implies the general mixture strategy works for mixture families with a compact $\mathcal{G}$ interior to Θ.

Acknowledgment

This research is supported by "the Aihara Project, the FIRST program from JSPS, initiated by CSTP," "JSPS Global COE program Education-and-Research Hub for Math-for-industry," and JSPS KAKENHI Grant Numbers 19300051 and 24500018.

References

[1] S. Amari, *Differential-geometrical methods in statistics* (*2nd pr.*), Lecture Notes in Statistics, Vol.28, Springer-Verlag, 1990.

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

[3] 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.

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

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

[6] B. Clarke & A. R. Barron, "Information-theoretic asymptotics of Bayes methods," *IEEE trans. on IT,* vol. 36. no. 3, pp. 453-471, 1990.

[7] B. Clarke & A. R. Barron, "Jeffreys prior is asymptotically least favorable under entropy risk," *J. Statistical Planning and Inference,* 41:37-60, 1994.

[8] D. Haussler, "A general minimax result for relative entropy," *IEEE trans. Inform. Theory,* vol. 43, no. 4, pp. 1276-1280, 1997.

[9] D. Haussler & A. R. Barron, "How well do Bayes methods work for on-line prediction of $\{\pm 1\}$ values," *Proceedings 1992 NEC Symposium on Computation and Cognition,* Chapter 4.

[10] I. Ibragimov & R. Hasminski, *Statistical estimation; Asymptotic theory,* Springer, New York, 1981.

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

[12] L. Le Cam, *Asymptotic Methods in Statistical Decision Theory* (Springer Series in Statistics), Springer New York, 1986.

[13] D. Pollard, Online notes, "http://www.stat.yale.edu/˜pollard/Books/Asymptopia/," 2010.

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

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

[16] 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.

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

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

[20] 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.

[21] 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.