Asymptotically Minimax Regret for Models with Hidden Variables

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 string $x^n = x_1x_2...x_n$ from an alphabet $\mathcal{X}$, in terms of regret with respect to models with hidden variables including general mixture families. When the target class is a non-exponential family, a modification of Jeffreys prior which has measure outside the given family of densities was introduced to achieve the minimax regret [8], under certain regularity conditions. In this paper, we show that the models with hidden variables satisfy those regularity conditions, when the hidden variables' model is an exponential family. In paticular, we do not have to restrict the class of data strings so that the MLE is in the interior of the parameter space for the case of the general mixture family.**

I. Introduction

We study the problem of data compression, gambling and prediction of a string $x^n = x_1x_2...x_n$ from a certain alphabet $\mathcal{X}$ (not restricted to be discrete), in terms of regret with respect to models with hidden variables including general mixture families. In particular, we evaluate the regret of Bayes mixture densities and show that it asymptotically achieves their minimax values when variants of Jeffreys prior are used. This is some extension of the result for the mixture family addressed in Takeuchi & Barron [11], [13].

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 string A coding scheme for the string 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.

We give a brief review on the known results for this problem When S is the class of discrete memoryless sources, Xie and Barron [18] proved that the minimax regret asymptotically equals $(d/2)\log(n/2\pi) + \log C_J + o(1)$, where d equals the size of alphabet minus 1 and C_J is the integral of the square root of the determinant of Fisher information matrix over the whole parameter space. For the stationary Markov model with finite alphabet, the analogous result is known [16]. Next, consider the set of strings restricted as $\mathcal{K}_n = \mathcal{X}^n(K) = \{x^n : \hat{\theta} \in K\}$, where K is a certain nice subset (satisfies $\bar{K} = \bar{K^\circ}$) of Θ. For multi-dimensional exponential families, variants of Jeffreys mixture are minimax. The ordinary Jeffreys mixture for the concerned curved family is not minimax, even if K is a compact set included in the interior of Θ. However, the minimax result can be obtained 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 an enlargement of the original family using exponential tilting. The idea for this enlargement in addressing minimax regret originates in preliminarily form in [10], [3] as informally discussed in [2]. The literature [15] gives discussion in the context of Amari's information geometry [1].

Recently, more formal statement for that method was given in [11], where an example of the mixture family case was argued. In this paper, we extend it to the case of models with hidden variables assuming the hidden variables' model is an exponential family. Further, we show that the asymptotic minimax procedure for the full family of strings can be obtained by modifying the procedure in [11].

II. Preliminaries

Let $(\mathcal{X}, \mathcal{F}, \nu)$ be a measurable space. 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

978-1-4799-5186-4/14/$31.00 ©2014 IEEE

the Fisher information:

$$\hat{J}_{ij}(\theta) = \hat{J}_{ij}(\theta, x^n) = \frac{-1}{n}\frac{\partial^2 \log p(x^n|\theta)}{\partial\theta_i \partial\theta_j},$$
$$J_{ij}(\theta) = E_\theta \hat{J}_{ij}(\theta, x).$$

The exponential family is defined as follows [4], [1].

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

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

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 $\{p(x|\theta)|\theta \in K \subseteq \Theta\}$ as an exponential family of densities.

When Θ is an open set, $S(\Theta)$ is said to be a regular exponential family. Many popular exponential families are regular. For exponential families, the entries of the Fisher information are given by $\partial^2\psi(\theta)/\partial\theta_i\partial\theta_j$. For regular exponential families, define the 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. Let $\mathcal{H}$ denote $\eta(\Theta^\circ)$ and $\mathcal{W}$ denote the closure of the convex hull of $T(\mathcal{X})$, then $\mathcal{H} = \mathcal{W}^\circ$ holds, when S is a steep exponential family ($E_\theta|T(x)| = \infty$ for all $\theta \in \Theta \setminus \Theta^\circ$).

Also, $\eta_i = \partial\psi(\theta)/\partial\theta^i$ holds. Note that $p(x^n|\theta) = \exp(n(\theta \cdot \bar{T} - \psi(\theta)))$ holds, where $\bar{T} = \sum_{t=1}^n T(x_t)/n$. (Here x_t denotes the t-th element of the sequence x^n.) It is known that the maximum likelihood estimate of η given x^n equals $\bar{T}$.

The multinomial model is an important example of the regular exponential family.

Example 1 (multinomial model): Let $\mathcal{X} = \{0, 1, \ldots, d\}$, $\nu(\{x\}) = 1$ for $x \in \mathcal{X}$, and $T(x) = (\delta_{1x}, \delta_{2x}, \ldots, \delta_{dx})$, where δ_{ij} is the Kronecker's delta. Define $p(x|\theta) = \exp(\theta \cdot T(x) - \psi(\theta)) = e^{\theta_x}/(1 + \sum_{x=1}^d e^{\theta_x})$, where $\theta = (\theta_1, \theta_2, \ldots, \theta_d) \in \Theta = \Re^d$. Then, we have $\psi(\theta) = \log(1 + \sum_{x=1}^d e^{\theta_x})$, which is finite for all $\theta \in \Theta$. Note that $\eta_x = p(x|\theta)$ and $\theta_x = \log(\eta_x/(1 - \sum_{x=1}^d \eta_x))$.

For a subset K of Θ, we let $C_J(K) = \int_K |J(\theta)|^{1/2}d\theta$. The Jeffreys prior ([5]) over K (denoted by $w_K(\theta)$) is defined as $w_K(\theta) = |J(\theta)|^{1/2}/C_J(K)$. We define the Jeffreys mixture for K (denoted by m_K) as $\int_K p(x^n|\theta)w_K(\theta)d\theta$.

Definition 2 (Model with Hidden Variables): Let $q(y|\theta)$ be a density of a d-dimensional exponential family over $\mathcal{Y}$. Define a class S of probability density functions over $\mathcal{X}$ by

$$S = \Big\{p(x|\theta) = \int \kappa(x|y)q(y|\theta)\nu_y(dy) \;\Big|\; \theta \in \Theta\Big\}, \tag{1}$$

where $\kappa(x|y)$ is a fixed conditional probability density function of x given y, and ν_y is the reference measure for $q(y|\theta)$.

If $q(y|\theta)$ is the multinomial model over $\mathcal{Y} = \{0, 1, \ldots, d\}$, then $p(x|\theta)$ forms a mixture family as

$$p(x|\theta) = \sum_{y=1}^d \eta_y\kappa(x|y) + (1 - \sum_{y=1}^d \eta_y)\kappa(x|0).$$

Lower Bound on Minimax Regret

A lower bound on minimax regret with the target class being a general smooth family is shown as follows. 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 $\delta(\theta) > 0$ so that for each i, j, $E_\theta \sup_{\theta':|\theta'-\theta|\le\delta(\theta)} |\hat{J}_{ij}(\theta', x)|^2$ 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 every $\epsilon > 0$, $\inf_{(\theta,\theta'):|\theta-\theta'|>\epsilon} D(\theta|\theta') > 0$ holds.

Assumption 4: For every 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 in appropriately defined models with hidden variables.

We have 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 $\bar{K} = \overline{K^\circ}$. The following holds.

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

Finally in this section, we state the following useful inequalities for the model with hidden variables.

Lemma 1: Given a data string x^n, let $\hat{\theta}$ denote the MLE for a model with hidden variables $p(x^n|\theta)$ defined by (1). Then, the following holds for all $x^n \in \mathcal{X}^n$.

$$\forall\theta \in \Theta,\ \frac{1}{n}\log\frac{p(x^n|\hat{\theta})}{p(x^n|\theta)} \le D(q(\cdot|\hat{\theta})|q(\cdot|\theta)) \tag{2}$$
$$\forall\theta \in \Theta,\ \hat{J}(\theta, x^n) \le G(\theta) \tag{3}$$

where $G(\theta)$ is the Fisher information of θ for $q(y|\theta)$, and $D(q(\cdot|\hat{\theta})|q(\cdot|\theta))$ denotes the Kullback-Leibler divergence from $q(\cdot|\hat{\theta})$ to $q(\cdot|\theta)$.

In particular, when $q(y|\theta)$ is the multinomial model, the following holds

$$\frac{p(x^n|\hat{\theta})}{p(x^n|\theta)} \le \exp(nD(q(\cdot|\hat{\theta})|q(\cdot|\theta))) = \prod_{y\in\mathcal{Y}}\frac{\hat{\eta}_y^{n\hat{\eta}_y}}{\eta_y^{n\hat{\eta}_y}}, \tag{4}$$

where $\eta_y = q(y|\theta)$ and $\hat{\eta}_y = q(y|\hat{\theta})$.

This lemma can be shown by a convexity argument. First we proved it for the mixture family case as in [12]. Then, Hiroshi Nagaoka pointed out that (2) holds for this case and gave a proof from the view point of information geometry. We gave an extended statement and a different proof. See [14] for the detail.

III. Minimax Bayes for Restricted Sets

We discuss the minimax strategy under the condition that data strings are restricted so that the maximum likelihood estimate is in a compact set interior to the parameter space.

A. Exponential Families

For exponential families it is known that a sequence of Jeffreys mixtures achieves the minimax regret asymptotically [18], [9], [10], [16]. For the multinomial and Finite State Machine models this holds for the full parameter set $K = \Theta$, whereas for general exponential families these facts are proven provided that K is a compact subset included in Θ°.

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

$$\log \frac{p(x^n|\hat{u})}{m_{J,n}(x^n)} = \frac{d}{2} \log \frac{n}{2\pi} + \log C_J(K) + o(1), \tag{5}$$

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 by the following asymptotic formula, which holds uniformly in K_n by Laplace integration:

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

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.

B. General Smooth Families

When the target class is not an exponential family, we form an enlargement of S by exponential tilting using linear combinations of the entries of $\hat{J}(\theta) - J(\theta)$. Actually, we introduce its normalized version as

$$V(x^n|\theta) = J(\theta)^{-1/2} \hat{J}(\theta) J(\theta)^{-1/2} - I. \tag{6}$$

Let $\mathcal{B} = (-b/2, b/2)^{d \times d}$ for some $b > 0$. The enlargement is formed as

$$\bar{p}(x^n|u) = p(x^n|\theta) e^{n(\beta \cdot V(x^n|\theta) - \psi(\theta,\beta))}, \tag{7}$$

where u denotes the pair (θ, β), β is a matrix in $\mathcal{B}$, $V(x^n|\theta) \cdot \beta$ denotes $\mathrm{Tr}(V(x^n|\theta)\beta^\dagger) = \sum_{ij} V_{ij}(x^n|\theta)\beta_{ij}$, and $\psi(\theta, \beta)$ is the logarithm of the required normalization factor, defined as

$$\psi(\theta, \beta) \stackrel{\text{def}}{=} \log \int p(x|\theta) \exp(V(x|\theta) \cdot \beta) \nu(dx).$$

Then, we define $\bar{S} = \{\bar{p}(\cdot|u) : u \in \Theta \times \mathcal{B}\}$. In [11], we employed $\hat{J}(\theta) - J(\theta)$, which was sufficient for the restricted strings set cases. To prove the minimax bound without the restriction on the set of strings, we introduce (6).

We have $\partial\psi(\theta, \beta)/\partial\beta_{ij} = E_{\theta,\beta} V_{ij}(x|\theta)$ and

$$\frac{\partial^2 \psi(\theta, \beta)}{\partial \beta_{ij} \beta_{kl}} = E_u V_{ij}(x|\theta) V_{kl}(x|\theta) - E_u V_{ij}(x|\theta) E_u V_{kl}(x|\theta).$$

The latter is the covariance between $V_{ij}(x|\theta)$ and $V_{kl}(x|\theta)$. Let $\mathrm{Cov}(\theta, \beta)$ denote this matrix whose (ij, kl)-entry is $\partial^2\psi(\theta, \beta)/\partial\beta_{ij}\beta_{kl}$. This is a covariance matrix of $V(x|\theta)$. Let $\lambda_K^* = \lambda^*(K)$ denote the maximum of the largest eigenvalue of $\mathrm{Cov}(\theta, \beta)$ for $\theta \in K$ and $\beta \in \mathcal{B}$.

Traditionally such enlargements of the families as in (7) arise in local asymptotic expansion of likelihood ratio used in demonstration of *local asymptotic normality* [6]. In Amari's information geometry [1] it is a *local exponential family bundle*.

When the target class is not an exponential family and $K \subset \Theta^\circ$, we employ the mixtures

$$\bar{m}_n(x^n) = (1 - n^{-r}) m_{J,n}(x^n) + n^{-r} \int \bar{p}(x^n|u) w(u) du. \tag{8}$$

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 $\mathcal{B}$.

In the analysis, the consideration of β in a neighborhood of a small multiple of $V(x^n|\theta)$ is sufficient to accomplish our objectives under the assumptions addressed below. Here, we define two neighborhoods of θ' as

$$B_\epsilon(\theta') = \{\theta : (\theta - \theta')^\dagger J(\theta')(\theta - \theta') \le \epsilon^2\}, \tag{9}$$

$$\hat{B}_\epsilon(\theta') = \{\theta : (\theta - \theta')^\dagger \hat{J}(\theta')(\theta - \theta') \le \epsilon^2\}. \tag{10}$$

Assumption 5: For the prior density function we use, we assume the following semi-continuity, that is, there exists a positive number $\kappa_w = \kappa_w(K)$ such that $w(\theta') \ge (1 - \kappa_w\epsilon) w(\theta)$ holds for all small $\epsilon > 0$, for all $\theta \in K$, and for all θ' in $B_\epsilon(\theta)$.

Define a set of good sequences $\mathcal{G}_{n,\delta}$ and a set of not good sequences $\mathcal{G}_{n,\delta}^c$. Note that $\mathcal{G}_{n,\delta}^c = \mathcal{K}_n \setminus \mathcal{G}_{n,\delta}$.

$$\mathcal{G}_{n,\delta} = \{x^n : ||V(x^n|\hat{\theta})||_s \le \delta \text{ and } \hat{\theta} \in K\},$$

$$\mathcal{G}_{n,\delta}^c = \{x^n : ||V(x^n|\hat{\theta})||_s > \delta \text{ and } \hat{\theta} \in K\},$$

where $||A||_s$ for a symmetric matrix $A \in \Re^{d \times d}$ denotes the spectral norm of A defined as $||A||_s = \max_{z:|z|=1} |z^\dagger A z|$. Comparing with the Frobenius norm defined as $||A|| = (\mathrm{Tr}(AA^\dagger))^{1/2}$, $||A||/\sqrt{d} \le ||A||_s \le ||A||$ holds.

Assumption 6: We assume a kind of equi-semicontinuity for $\hat{J}(\theta)$, that is, there exist a positive number $\kappa_J = \kappa_J(K)$ such that for all small $\epsilon > 0$, for a certain $\delta_0 > 0$ that is independent of K, for all x^n in $\mathcal{G}_{n,\delta_0}$, for all $\tilde{\theta} \in \hat{B}_\epsilon(\hat{\theta})$, and for all $\theta \ne \hat{\theta}$,

$$\frac{(\theta - \hat{\theta})^\dagger \hat{J}(\tilde{\theta})(\theta - \hat{\theta})}{(\theta - \hat{\theta})^\dagger \hat{J}(\hat{\theta})(\theta - \hat{\theta})} \le 1 + \kappa_J \epsilon.$$

This is used to control the Laplace integration for $\bar{m}_n$ of our strategy for the good sequences. In fact, we can prove the following lemma, where Φ is the probability measure of standard normal distribution over $\Re^d$.

Lemma 2: Fix a compact set K' such that $K' \subset \Theta^\circ$ and $K \subset K'$. Suppose Assumptions 1-3, 5, and 6 hold and that $\hat{B}_\epsilon(\hat{\theta}) \subset K'$. Then, for all $\delta < \delta_0$, the following holds.

$$\inf_{x^n \in \mathcal{G}_{n,\delta}} \frac{m_{K'}(x^n)}{p(x^n|\hat{\theta})} \ge \frac{(1 - \kappa_w\epsilon)^{d/2} \Phi(\sqrt{n} B_\epsilon(0))}{(1 + \kappa_J\epsilon)^{d/2} (1 + \delta)^{d/2}} \frac{(2\pi)^{d/2}}{C(K') n^{d/2}}.$$

The proof is omitted.

Assumption 7: For any $\delta > 0$, there exists an $\epsilon > 0$, such that for all x^n in $\mathcal{G}^c_{n,\delta}$, for all $\tilde{\theta}$ in $B_\epsilon(\hat{\theta})$, $||V(x^n|\tilde{\theta})||_s \geq \zeta\delta$, holds, where ζ is a certain constant.

Assumption 8: There exists an $\epsilon > 0$, such that for all x^n in $\mathcal{K}$, and for all $\tilde{\theta} \in B_\epsilon(\hat{\theta})$, $2(\hat{J}(\hat{\theta}) + J(\hat{\theta})) - \hat{J}(\tilde{\theta})$ is positive semidefinite.

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

$$\inf_{x^n \in \mathcal{G}^c_{n,\delta}} \frac{\int \bar{p}(x^n|u)w(u)du}{p(x^n|\hat{\theta})} \geq (\xi^*_K)^{d^2} \delta^{d^2} n^{-d} \exp(A\delta^2 \xi^*_K n)$$

where A is a positive constant independent of K and $\xi^*_K = \min\{n/\lambda^*_K, 1\}$.

Remark: Here, λ^*_K/n is the maximum over all $u \in K \times \mathcal{B}$ of the largest eigenvalue of the covariance matrix of $V(x^n|\theta)$ with respect to $\bar{p}(x^n|u)$. If λ^*_K/n is large, we cannot expect appropriate likelihood gain.

From Lemmas 2 and 3, we have the following Theorem, which provides an equivalent upper bound as in [11].

Theorem 2: When a target class satisfies Assumptions 1-3 and 5-8, the strategy $\bar{m}_n$ defined as (8) asymptotically achieves the minimax regret, i.e.

$$\max_{x^n \in \mathcal{K}_n} \log \frac{p(x^n|\hat{\theta})}{\bar{m}_n(x^n)} \leq \frac{d}{2} \log \frac{n}{2\pi} + \log C_J(K) + o(1),$$

where r is appropriately chosen.

C. Models with Hidden Variables

For models with hidden variables, we can show the following equalities:

$$\begin{aligned} \hat{J}_{ij}(\theta, x^n) &= G_{ij}(\theta) - \tilde{E}_\theta(\bar{T}_i - \tilde{t}_i)(\bar{T}_j - \tilde{t}_j), \\ \frac{\partial \hat{J}_{ij}(\theta, x^n)}{\partial \theta_k} &= \frac{\partial G_{ij}(\theta)}{\partial \theta_k} - \tilde{E}_\theta(\bar{T}_i - \tilde{t}_i)(\bar{T}_j - \tilde{t}_j)(\bar{T}_k - \tilde{t}_k), \end{aligned}$$

where $\tilde{E}_\theta$ denotes the expectation by the posterior distribution of $T^n = T(y_1)...T(y_n)$ given x^n, $\bar{T} = \sum_{t=1}^n T(y_t)/n$, and $\tilde{t} = \tilde{E}_\theta \bar{T}$.

Here we assume that the range of $T(y)$ is bounded. Then from the above equations, when $\hat{\theta}$ is restricted in K interior to Θ, $|\partial \hat{J}(\theta, x^n)/\partial\theta|$ is uniformly bounded for all x^n. This means that $\hat{J}(\theta, x^n)$ is equi-continuous for all x^n as a function of θ and Assumptions 5-8 hold.

IV. Minimax Strategy for the Mixture Family without Restriction on Data Strings

Here we give the minimax strategy for the mixture family. Let θ denote the expectation parameter for the multinomial model for the hidden variable y, that is, we let $p(y|\theta) = \theta_y$, $\Theta = \{\theta \in \Re^d : \forall y \geq 1,\ \theta_y \geq 0, \sum_{y=1}^d \theta_y \leq 1\}$, and $\theta_0 = 1 - \sum_{y=1}^d \theta_y$. Define the interior set Θ_τ ($\tau > 0$) as

$$\Theta_\tau = \{\theta \in \Theta :\ \theta_y \geq \tau,\ y = 0, 1, \ldots, d\}.$$

Later, we argue the situation that τ converges to zero as n goes to infinity. Under that situation we utilize Lemmas 2 and 3. Then we have to control the behavior of $\kappa_J(K)$ and λ^*_K since K changes as n increases.

To treat the strings with the maximum likelihood estimate being near the boundary, we employ the technique introduced by Xie & Barron [18], which utilizes the Dirichlet(α) prior $w_{(\alpha)}(\theta) \propto \prod_{i=0}^d \theta_i^{-(1-\alpha)}$ with $\alpha < 1/2$. Note that this prior with $\alpha = 1/2$ has the same form as the Jeffreys prior for the multinomial model, and has higher density than the latter when θ approaches the boundary of Θ.

Our asymptotic minimax strategy is the mixture

$$\begin{aligned} \bar{m}_n(x^n) &= (1 - 2n^{-r}) m_J(x^n) \\ &+ n^{-r} \int \bar{p}(x^n|u)w(u)du \\ &+ n^{-r} \int p(x^n|\theta) w_{(\alpha)}(\theta) d\theta. \end{aligned} \tag{11}$$

Here, the first and second terms are for the strings with the MLE being away from the boundary, while the third term works when the MLE approaches the boundary.

We can show the following theorem.

Theorem 3: The strategy $\bar{m}_n$ defined as (11) for a mixture family as the target class asymptotically achieves the minimax regret, i.e.

$$\max_{x^n \in \mathcal{X}^n} \log \frac{p(x^n|\hat{\theta})}{\bar{m}_n(x^n)} \leq \frac{d}{2} \log \frac{n}{2\pi} + \log C_J(\Theta) + o(1),$$

where r is appropriately chosen.

Since we cannot give the complete proof in this manuscript, we give some discussion below. We assume that $\hat{\theta} \in \Theta_{2\tau}$ and that $\epsilon < \tau$. Then, we have $\tilde{\theta} \in \Theta_\tau$ for all $\tilde{\theta} \in B_\epsilon(\hat{\theta})$.

We examine Assumption 6. Note that, for $z \in \Re^d$,

$$z^\dagger \hat{J}(\theta, x) z = \frac{(\sum_i z_i(\kappa(x|i) - \kappa(x|0)))^2}{p(x|\theta)^2} \geq 0.$$

Further we have

$$\frac{\partial z^\dagger \hat{J}(\theta, x) z}{\partial \theta_k} = \frac{-2(\kappa(x|k) - \kappa(x|0))}{(p(x|\theta))} z^\dagger \hat{J}(\theta) z, \tag{12}$$

yielding

$$\frac{\partial z^\dagger \hat{J}(\theta, x^n) z}{\partial \theta_k} = \sum_t \frac{-2(\kappa(x_t|k) - \kappa(x_t|0))}{(p(x_t|\theta))} z^\dagger \hat{J}(\theta, x_t) z. \tag{13}$$

Hence we have for $\theta \in \Theta_\tau$,

$$\left| \frac{\partial z^\dagger \hat{J}(\theta, x^n) z}{\partial \theta_k} \right| \leq \frac{2 z^\dagger \hat{J}(\theta, x_t) z}{\tau}.$$

Then we have

$$e^{-2\sqrt{d}|\tilde{\theta} - \hat{\theta}|/\tau} \leq \frac{z^\dagger \hat{J}(\tilde{\theta}, x^n) z}{z^\dagger \hat{J}(\hat{\theta}, x^n) z} \leq e^{2\sqrt{d}|\tilde{\theta} - \hat{\theta}|/\tau}. \tag{14}$$

Therefore, if $|\tilde{\theta} - \hat{\theta}| \leq \tau/2\sqrt{d}$,

$$\frac{z^\dagger \hat{J}(\tilde{\theta}, x^n) z}{z^\dagger \hat{J}(\hat{\theta}, x^n) z} \leq 1 + \frac{2\sqrt{d} e |\tilde{\theta} - \hat{\theta}|}{\tau}$$

holds for all $z \in \Re^d \backslash \{0\}$. It implies that when $\epsilon \leq \lambda_{\min}\tau/2\sqrt{d}$ is satisfied, $z^\dagger \hat{J}(\tilde{\theta}, x^n)z/z^\dagger \hat{J}(\hat{\theta}, x^n)z \leq 1 + e\sqrt{d}\epsilon/\tau$ holds for all $\tilde{\theta} \in B_\epsilon(\hat{\theta})$, where $\lambda_{\min}$ is the minimum of the smallest eigenvalue of $J(\theta)$. This implies Assumption 6 holds with $\kappa_J(\Theta_\tau) \leq e\sqrt{d}/\tau$ for $\epsilon \leq \lambda_{\min}\tau/2\sqrt{d}$.

Next we examine Assumption 7. Note that there exists a unit vector $\bar{z} \in \Re^d$ such that $|\bar{z}^\dagger V(x^n|\hat{\theta})\bar{z}| = ||V(x^n|\hat{\theta})||_s$. Here we have two cases; i) $\bar{z}^\dagger V(x^n|\hat{\theta})\bar{z} = ||V(x^n|\hat{\theta})||_s$ and ii) $-\bar{z}^\dagger V(x^n|\hat{\theta})\bar{z} = ||V(x^n|\hat{\theta})||_s$.

First consider the case i), for which we have

$$\begin{aligned} ||V(x^n|\hat{\theta})||_s &= \bar{z}^\dagger J(\hat{\theta})^{-1/2}\hat{J}(\hat{\theta})J(\hat{\theta})^{-1/2}\bar{z} - 1 \\ &= \frac{z^\dagger \hat{J}(\hat{\theta})z}{z^\dagger J(\hat{\theta})z} - 1 \end{aligned}$$

with a certain $z \in \Re^d$. That is,

$$\frac{z^\dagger \hat{J}(\hat{\theta})z}{z^\dagger J(\hat{\theta})z} = 1 + ||V(x^n|\hat{\theta})||_s.$$

For the numerator in the left side, from (14) we have

$$z^\dagger \hat{J}(\tilde{\theta})z \geq e^{-2\sqrt{d}|\tilde{\theta}-\hat{\theta}|/\tau} z^\dagger \hat{J}(\hat{\theta})z.$$

As for the denominator, similarly as (14) we can show by (12)

$$e^{-2\sqrt{d}|\tilde{\theta}-\hat{\theta}|/\tau} \leq \frac{z^\dagger J(\tilde{\theta})z}{z^\dagger J(\hat{\theta})z} \leq e^{2\sqrt{d}|\tilde{\theta}-\hat{\theta}|/\tau} \tag{15}$$

for $\tilde{\theta} \in \Theta_\tau$. Hence, noting $||V(x^n|\hat{\theta})||_s \geq \delta$, we have

$$\begin{aligned} \frac{z^\dagger \hat{J}(\tilde{\theta})z}{z^\dagger J(\tilde{\theta})z} &\geq (1+\delta)e^{-4\sqrt{d}|\tilde{\theta}-\hat{\theta}|/\tau} \\ &\geq (1+\delta)\Big(1 - \frac{4\sqrt{d}|\tilde{\theta}-\hat{\theta}|}{\tau}\Big) \\ &= 1 + \delta - \frac{4\sqrt{d}(1+\delta)|\tilde{\theta}-\hat{\theta}|}{\tau}. \end{aligned}$$

Hence when $|\tilde{\theta} - \hat{\theta}| \leq \delta\tau/(8\sqrt{d}(1+\delta))$ we have $z^\dagger \hat{J}(\tilde{\theta})z/z^\dagger J(\tilde{\theta})z \geq 1 + \delta/2$, which implies $||V(x^n|\tilde{\theta})||_s \geq \delta/2$. For the case ii), we can show the similar conclusion. Hence, if $\epsilon < \delta\tau/(8\sqrt{d}(1+\delta))$, $||V(x^n|\tilde{\theta})|| \geq \delta/2$ holds.

To see Assumption 8 holds is easy because of (14).

Finally, we evaluate $\xi^*_{\Theta_\tau} = \min\{n/\lambda^*_{\Theta_\tau}, 1\}$ in Lemma 3. Since $|\hat{J}_{ij}(\theta)|$ is bounded by $4/\tau^2$ for $\theta \in \Theta_\tau$ and since the smallest eigenvalue of $J(\theta)$ is lower bounded by a certain positive constant, $\lambda^*_{\Theta_\tau} = O(\tau^{-2})$, hence, $\xi^*_{\Theta_\tau}$ is lower bounded by $n\tau^2$ times a certain constant.

We set $\tau = n^{-(1-p)}(0 < p < 1)$ and $\delta = n^{-1/2+\gamma}(0 < \gamma < 1/2)$, where ϵ must satisfy $n^{-1/2+\iota} \leq \epsilon \leq \delta\tau$ $(0 < \iota < 1/2)$, in the sense of order, which is equivalent to $p > 1 - \gamma + \iota$. To make the second term of (11) have sufficient likelihood gain, it suffices to have

$$\liminf_{n\to\infty} \frac{\log(n\tau^2\delta^2)}{\log n} > 0,$$

which is equivalent to $p > 1 - \gamma$ in our setting. This is automatically satisfied under $p > 1 - \gamma + \iota$. When $\gamma > \iota$, there exists a p in $(0, 1)$ which satisfies $p > 1 - \gamma + \iota$.

In this setting, when $\hat{\theta}_x < 2n^{-(1-p)}$ for some x, we need the help from the third term in (11). From (4), we have

$$\frac{\int p(x^n|\theta)w_{(\alpha)}(\theta)d\theta}{p(x^n|\hat{\theta})} \geq \frac{\int \prod_{i=0}^d \theta_i^{n\hat{\theta}_i} w_{(\alpha)}(\theta)d\theta}{\prod_{i=0}^d \hat{\theta}_i^{n\hat{\theta}_i}}.$$

By Lemma 4 of [18], the right side is not less than

$$\frac{1}{Rn^{d/2-(1/2-\alpha)(1-p)}}$$

where R is a constant determined by m, α, and p. Let r be smaller than $(1/2-\alpha)(1-p)$, then for large n the third term in (11) is larger than $(2\pi)^{d/2}/(C_J n^{d/2})$.

Acknowledgment

The authors give their sincere gratitude to Hiroshi Nagaoka and the anonymous reviwers for 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] H. Jeffreys, *Theory of probability, 3rd ed.* , Univ. of California Press, Berkeley, Cal, 1961.
[6] D. Pollard, Online notes, "http://www.stat.yale.edu/~pollard/Books/Asymptopia/," 2010.
[7] J. Rissanen, "Fisher information and stochastic complexity," *IEEE trans. Inform. Theory,* vol. 40, pp. 40-47, 1996.
[8] Yu M. Shtarkov, "Universal sequential coding of single messages," *Problems of Information Transmission*, vol. 23, pp. 3-17, July 1988.
[9] 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.
[10] J. Takeuchi & A. R. Barron, "Asymptotically minimax regret by Bayes mixtures," *Proc. of 1998 IEEE ISIT,* p. 318, 1998.
[11] J. Takeuchi & A. R. Barron, "Asymptotically Minimax Regret by Bayes Mixtures for Non-exponential Families," *Proc. of 2013 IEEE ITW,* pp. 204-208, 2013.
[12] J. Takeuchi & A. R. Barron, "Some inequality for mixture families," http://www-kairo.csce.kyushu-u.ac.jp/~tak/papers/memo_r2.pdf, October 2013.
[13] J. Takeuchi & A. R. Barron, "Asymptotically minimax prediction for mixture families," *The 36th Symp. on Inform. Theory and Its Applications* (*SITA'13*), pp. 653-657, 2013. (without peer review, Japan domestic)
[14] J. Takeuchi & A. R. Barron, "Some inequality for models with hidden variables," http://www-kairo.csce.kyushu-u.ac.jp/~tak/papers/memo_r3.pdf, January 2014.
[15] 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.
[16] 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.
[17] 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.
[18] 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.