Asymptotically minimax regret for exponential and curved exponential families

(summary)

Jun-ichi Takeuchi* Andrew R. Barron†

November 10th, 1997

1 Introduction

We study the problem of data compression, gambling and prediction of a sequence $x^n = x_1 x_2 ... x_n$ from a certain alphabet $\mathcal{X}$, in terms of regret and redundancy 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 [14, 15] in the general smooth families. In particular for one-dimensional exponential families, they also extend the works of Clarke and Barron [5, 6] to deal with the full natural parameter space rather than compact sets interior to it.

This paper's main concern is the regret of a coding or prediction strategy. 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 respect to a family of probability mass function $S = \{p(\cdot|\theta) : \theta \in \Theta\}$ and a set of the sequences $W_n \subseteq \mathcal{X}^n$ (denoted by $\bar{r}(W_n)$) is defined as

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

where $\hat{\theta}$ is the maximum likelihood estimate 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 $\log(1/q(x_{i+1}|x^i)) - \log(1/p(x_{i+1}|x^i, \theta))$. The maximin regret for set W_n (denoted by $\underline{r}(W_n)$) is defined as

$$\sup_{q \in \mathcal{P}(W_n)} \inf_{r \in \mathcal{P}(\mathcal{X}^n)} E_q(\log \frac{p(x^n|\hat{\theta})}{r(x^n)}),$$

*C&C Media Research Laboratories, NEC Corporation, 4-1-1 Miyazaki, Miyamae, Kawasaki, Kanagawa 216, Japan. e-mail tak@ccm.CL.nec.co.jp (This work was done while Takeuchi was staying at Yale University.)

†Department of Statistics, Yale University, 24 Hillhouse Avenue, New Haven, CT 06511, USA. e-mail barron@stat.yale.edu

where $\mathcal{P}(W_n)$ is the set of all probability mass function over W_n and E_q denotes the expectation with respect to q. It is known that $\bar{r}(W_n) = \underline{r}(W_n)$ holds [13, 15]. 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 [15] 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 $\mathcal{G}$. An important point in the above is that $\mathcal{G}$ is taken there to be Θ itself, i.e. we do not have to have any restriction for the sequence x^n. For obtaining this asymptotically minimax regret, they use sequences of Bayes mixtures with prior distributions that weakly converge to the Jeffreys prior (the prior proportional to the square root of the determinant of Fisher information matrix):

$$m_n(x^n) = \int p(x^n|\theta) w_n(d\theta),$$

where $\{w_n(d\theta)\}$ is a sequence of prior measures over Θ. The reason why one needs such variants of the Jeffreys prior is as follows: If we use the Jeffreys prior, the risk is asymptotically higher than the minimax value, for x^n such that $\hat{\theta}$ is near the boundary of Θ. We use priors which have higher density near the boundaries than the Jeffreys prior, to give more prior attention to these boundary regions and thereby pull the risk down to the asymptotically minimax level.

In this paper, we generalize the results of [15] to the case where S is an exponential family or the related curved exponential family.

For the multi-dimensional exponential family, variants of Jeffreys mixture are minimax, when $\mathcal{G}$ is a compact subset included in the interior of Θ. For the curved exponential family, the ordinary Jeffreys mixture for the subjective curved family is not minimax, even if $\mathcal{G}$ is a compact subset 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 subjective curved family. It is remarkable that this procedure is applicable to general smooth families. For the one-dimensional exponential family, we succeed to obtain variants of Jeffreys mixture which are minimax for any subset $\mathcal{G}$ under certain conditions.

We also consider the problem of minimax expected regret (redundancy). The minimax expected regret for the subset $\mathcal{G}$ of Θ (denoted by $\bar{R}_n(\mathcal{G})$) is defined as

$$\inf_q \sup_{\theta \in \mathcal{G}} E_\theta(\log \frac{1}{q(x^n)} - \log \frac{1}{p(x^n|\theta)}).$$

Also, the maximin expected regret for the parameter set $\mathcal{G}$ (denoted by $\underline{R}_n(\mathcal{G})$) is defined as

$$\sup_w \inf_q \int E_\theta(\log \frac{1}{q(x^n)} - \log \frac{1}{p(x^n|\theta)}) w(\theta) d\theta,$$

where supremum is taken for any prior measure w. It is known that $\bar{R}_n(\mathcal{G}) = \underline{R}_n(\mathcal{G})$ holds [8, 11, 9].

For asymptotics of this minimax expected regret, the results by Clarke and Barron [6] are known. They considered fairly general classes of i.i.d. processes and showed that the minimax expected regret asymptotically equals

$$(d/2)\log(n/2\pi e) + \log C_J(\mathcal{G}) + o(1),$$

where $\mathcal{G}$ must be a compact subset of Θ°. In work preceding [15], Xie and Barron [14] evaluated the minimax expected regret for the class of discrete memoryless sources and showed that sequences of slightly varied Jeffreys mixtures achieve the minimax value asymptotically for the probability simplex Θ. The answer for the minimax regret and the minimax expected regret are similar. We give analogous conclusions for both measures of regret for one-dimensional exponential families.

For obtaining the above minimax results, we employ the Laplace integration method, which was used by Clarke and Barron [5, 6] in order to evaluate the expected regret of the Bayes procedures. Especially in [6], they succeeded to uniformly evaluate the expected regret by the Laplace integration for a compact subset $\mathcal{G}$ of Θ°. However in our task for the one-dimensional case, a subset $\mathcal{G}$ can be arbitrary. This requires very careful application of the Laplace integration.

For determining the minimax value for curved exponential families, we use Rissanen's recipe [12] for evaluating the regret of the maximum likelihood code [13]. His recipe requires that the central limit theorem about MLE $\hat{\theta}$ uniformly holds for the parameter space $\mathcal{G}$. We show that it does hold for curved exponential families, when $\mathcal{G}$ is compact.

The maximum likelihood code 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 problem 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

$$m_N(x_n|x^{n-1}) = \int p(x_n|\theta)w_N(d\theta|x^{n-1}),$$

where we let $w_N(d\theta|x^{n-1})$ denote the posterior measure of θ given x^{n-1}.

2 Definitions

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

Definition 1 (Exponential Family) *Let ν be a σ-finite measure on the Borel subsets of $\Re^d$ and $\mathcal{X}$ be the support of ν. Define $\Theta \equiv \{\theta : \theta \in \Re^d, \int_{\mathcal{X}} \exp(\theta \cdot x)\nu(dx) < \infty\}$. Define a function ψ and a probability density p on $\mathcal{X}$ with respect to ν by $\psi(\theta) \equiv \log \int_{\mathcal{X}} \exp(\theta \cdot x)\nu(dx)$ and $p(x|\theta) \equiv \exp(\theta \cdot x - \psi(\theta))$. We refer to the set $S(\Theta) \equiv \{p(x|\theta)|\theta \in \Theta\}$ as an exponential family of densities.*

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.

Under this definition, the regret should be $\log(1/q(x^n)\nu(dx^n)) - \log(1/p(x^n|\hat{\theta})\nu(dx^n))$, where q is a probability density with respect to the measure ν, but that equals $\log(1/q(x^n)) - \log(1/p(x^n|\hat{\theta}))$. Hence, we can use the same definitions of regret given in the previous section.

When Θ is an open set, $S(\Theta)$ is said to be a regular exponential family. Many popular exponential families are regular, but we assume that $S(\Theta)$ is steep. This is a weaker condition than "regular". (When for all $\theta \in \Theta - \Theta^\circ$, $E_\theta(|x|) = \infty$ holds, then $S(\Theta)$ is said to be steep.) 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 steep exponential families, define expectation parameter η as $\eta(\theta) = E_\theta(x)$. It is known that the map : $\theta \mapsto \eta$ is one-to-one and analytic on Θ°. Also, $\eta_i = \partial\psi(\theta)/\partial\theta^i$ holds. We also use the terminology $\theta(\eta)$ as inverse function of $\eta(\theta)$. Note that $p(x^n|\theta) = \exp(n(\theta \cdot \bar{x} - \psi(\theta)))$ holds, where $\bar{x} = \sum_{t=1}^n x_t/n$. ($x_t$ 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 $\bar{x}$.

Exponential families include many common statistical models such as Gaussian distributions, Poisson distributions, Bernoulli sources and etc. We explain some examples of exponential family.

Example 1 (Bernoulli sources) *Let $\mathcal{X} = \{0,1\}$ and $\nu(\{x\}) = 1$ for $x = 0, 1$. Then, we have $\psi(\theta) = \log(1 + e^\theta)$, which is finite for all $\theta \in \Re$. Hence, $\Theta = \Re$. We have $p(1|\theta) = \exp(\theta - \psi(\theta)) = e^\theta/(1 + e^\theta)$ and $p(0|\theta) = \exp(-\psi(\theta)) = 1/(1 + e^\theta)$. Also we have*

$$J(\theta) = \frac{e^\theta}{(1 + e^\theta)^2}.$$

Example 2 (Poisson distributions) *Let $\mathcal{X} = \{0, 1, ...\}$, and $\nu(\{x\}) = 1/x!$. We have*

$$\psi(\theta) = \log \sum_x \frac{e^{\theta x}}{x!} = e^\theta.$$

Hence, $\Theta = \Re$ and $J(\theta) = e^\theta$.

Example 3 (Inverse Gaussian distributions) *The density of inverse Gaussian distribution with respect to Lebesgue measure is*

$$p(x|c, \mu) = (\frac{c}{2\pi x^3})^{1/2} \exp(c \cdot (-\frac{x}{2\mu^2} - \frac{1}{2x} + \frac{1}{\mu})),$$

where $\mu > 0$, $c > 0$, and $x > 0$. Hereafter, we fix c. It may be arbitrary, but we let $c = 1$ for simplicity. Let $\theta = -1/2\mu^2$. We have

$$\begin{aligned} p(x|\theta) &= (\frac{1}{2\pi x^3})^{1/2} \exp(\theta x + \sqrt{-\theta} - \frac{1}{2x}) \\ &= (\frac{1}{2\pi})^{1/2} \exp(\theta x + \sqrt{-\theta} - \frac{1}{2x} - \frac{3}{2}\log x). \end{aligned}$$

Hence, we can see $\Theta = (-\infty, 0]$, $\nu(dx) = \exp(-1/2x - (3/2) \cdot \log x)dx$ and $\psi(\theta) = \sqrt{-\theta}$. Also, we have

$$J(\theta) = \frac{1}{-4\theta\sqrt{-\theta}}.$$

Note that the inverse Gaussian family is an example of not regular but steep exponential families.

We let $C_J(\mathcal{G}) = \int_{\mathcal{G}} \sqrt{\det(J(\theta))}d\theta$. The Jeffreys prior ([10]) over $\mathcal{G}$ (denoted by $w_{\mathcal{G}}(\theta)$) is defined as

$$w_{\mathcal{G}}(\theta) = \frac{\sqrt{\det(J(\theta))}}{C_J(\mathcal{G})}.$$

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

Finally, we introduce the curved exponential family. Let $S = \{p(x^n|\theta) : \theta \in \Theta\}$ be the $\bar{d}$-dimensional steep exponential family. Using a smooth function $\phi : \Re^d \to \Re^{\bar{d}}$, we define a subfamily of S as follow:

$$M = \{p_c(x^n|u) = p(x^n|\phi(u)) : u \in \mathcal{U}\},$$

where $\mathcal{U}$ is a certain open set of $\Re^d$ and $\bar{d} \geq d$. This M is referred to as a curved exponential family embedded in S. We let $\hat{u}$ denote the maximum likelihood estimate of u given x^n:

$$\hat{u} = \arg\max_{u \in \mathcal{U}} p_c(x^n|u).$$

3 Lower Bounds

3.1 Exponential Families

The following holds for d-dimensional steep exponential families.

$$\liminf_{n\to\infty}(\underline{r}(\mathcal{X}^n(\mathcal{G})) - \frac{d}{2}\log\frac{n}{2\pi}) \geq \log C_J(\mathcal{G}). \tag{2}$$

Note that this holds for any nice $\mathcal{G}$.

The inequality (2) is shown by using the following which we can show by Laplace integration.

$$\liminf_{n\to\infty} \inf_{x^n:\hat{\theta}\in\mathcal{G}'} (\log\frac{p(x^n|\hat{\theta})}{m_{\mathcal{G}}(x^n)} - \frac{d}{2}\log\frac{n}{2\pi}) \geq \log C_J(\mathcal{G}),$$

where $\mathcal{G}'$ is any compact set interior to $\mathcal{G}$.

3.2 Curved Exponential Families

Though we obtained the lower bound for exponential families via the direct evaluation of the lower bound for the Bayes mixture, it is difficult for the curved exponential family. Hence, we utilize the theorem about the maximum likelihood code by Rissanen [12], for determining the minimax value for this case. According to that theorem, the minimax and maximin regret equals

$$\frac{d}{2}\log\frac{n}{2\pi} + \log\int\sqrt{\det(J(\theta))}d\theta + o(1).$$

under certain conditions. The main condition here is that the central limit theorem about the maximum likelihood estimate uniformly holds for the concerned class of probability distributions and is not trivial for curved exponential families (other conditions are also required, but they clearly hold for curved exponential family). We can confirm that this condition holds for the case of curved exponential family with compact parameter space interior to the whole parameter space of the family. For that proof, we use the theorem by Bhattacharya [2] (see also [3]).

We can prove that any natural Bayes mixture with prior on the curved family M does not achieve the above upper bound. Hence we need a certain different idea, which we describe in Subsection 4.2.

4 Upper Bounds

In all cases studied in this paper the minimax value of (pointwise) regret equals

$$\frac{\dim(\mathcal{G})}{2}\log\frac{n}{2\pi} + \log C_J(\mathcal{G}) + o(1). \tag{3}$$

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

4.1 Multi-dimensional Exponential Families

Let $\mathcal{G}$ be a nice compact subset of Θ°. 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 $n \to \infty$, where $C_J(\mathcal{G}_n)$ reduces to $C_J(\mathcal{G})$. If the rate of that reduction is sufficiently slow, then

$$\limsup_{n\to\infty}\Big(\sup_{x^n:\hat{\theta}\in\mathcal{G}} \log\frac{p(x^n|\hat{\theta})}{m_{\mathcal{G}_n}(x^n)} - \frac{d}{2}\log\frac{n}{2\pi}\Big) \le \log C_J(\mathcal{G}) \tag{4}$$

holds. Since the upper bound here matches the lower bound, our strategy is minimax and we have determined the minimax value.

4.2 Curved Exponential Families

Let $\mathcal{U}_s$ be a certain nice set of $\Re^d$ ($\bar{\mathcal{U}_s}^\circ = \bar{\mathcal{U}_s}$). Assume that $\mathcal{U}_s$ is a compact set interior to $\mathcal{U}$. We consider the minimax problem for the set $W_n = \{x^n : \hat{u} \in \mathcal{U}_s\}$

Here, we use a prior that lives not just on the subfamily M but also has a small component on the larger family S. The resultant procedure is based on the mixture

$$m_n^{(curve)}(x^n) = (1-\epsilon_n)\int p_c(x^n|u)w_{\mathcal{U}_n}(u)du + \epsilon_n \int p(x^n|\theta)w_{\mathcal{G}}(\theta)d\theta. \tag{5}$$

The first component is Jefrreys prior $w_{\mathcal{U}_n}$ on an open set $\mathcal{U}_n$ shrinking to $\mathcal{U}_s$ (this is absolutely continuous with respect to Lebesugue measure du on $\mathcal{U}$). The second component of the prior is absolutely continuous with respect to Lebesugue measure on Θ, with a positive density on a set $\mathcal{G}$ that contains a neighborhood of the image of $\mathcal{U}_s$ in Θ by map ϕ. The weight ϵ_n tends to zero at a polynomial rate.

The conclusion in this case is that this modified Jeffreys mixture is asymptotically minimax and maximin and that it achieves the value (3) asymptotically in agreement with the value achieved by normalized maximum likelihood as shown by Rissanen for stochastic complexity.

4.3 One-dimensional Exponential Families

For one-dimensional exponential families with natural parameter space Θ with integrable $\sqrt{J(\theta)}$, we identify two main types of boundary or tail behavior. The natural parameter space Θ forms an interval with right end point b either finite ($b < \infty$) or infinite ($b = \infty$). Here we focus on the behavior on the right side of the interval. (The behavior on the left side is analogous.)

Let λ be an element of Θ°. We let $\mathcal{G} = [\lambda, \infty) \cap \Theta$ and consider the minimax problem for the set $\mathcal{X}^n(\mathcal{G})$.

In the case that $b = \infty$ and that root of $J(\theta)$ slightly smaller than $1/2$ is integrable, we use priors $w_n(\theta)$ defined on $\mathcal{G}_n$ and proportional to $(J(\theta))^{(1-\alpha_n)/2}$, where α_n is any choice that tends to zero slower than $1/\log n$ and $\{\mathcal{G}_n\}$ is analogously defined as in the multi dimensional case. Then, this procedure is asymptotically minimax. This case includes Bernoulli sources and Poisson distributions. This method provides an alternative to the technique in Xie and Barron [14, 15].

In the case that the right endpoint of Θ is a finite b, we identify two situations for steep exponential families. In one case the right endpoint b is in Θ (non regular exponential family) and we use

$$w_n(d\theta) = (1-\epsilon_n)w_{\mathcal{G}_n}(\theta)d\theta + \epsilon_n\delta_b(d\theta), \tag{6}$$

where $\mathcal{G}_n = [\lambda_n, b)$ with $\lambda_n \leq \lambda$ and $w_{\mathcal{G}_n}$ is Jeffreys prior on $\mathcal{G}_n$ (absolutely continuous with respect to Lebesgue measure $d\theta$), the component δ_b is point mass at b and ϵ_n is any sequence converge to zero slower at rate $n^{-\beta}$ for some $\beta \leq 1/2$. If λ_n approaches λ sufficiently slowly, then the above strategy is asymptotically minimax. This case includes Inverse Gaussian family. It is remarkable that if we use prior $w(d\theta)$ which is absolutely continuous at point b (e.g. the prior (6) with $\epsilon_n = 0$), we have

$$\lim_{\hat{\theta} \to b-0} \log \frac{p(x^n|\hat{\theta})}{\int p(x^n|\theta) w(d\theta)} = \infty$$

for any n. Hence, the second component of (6) is necessary.

Finally for regular exponential families with finite endpoint, Θ is open and hence does not contain b. In this case, under a certain weak condition we confirm that Fisher information $J(\theta)$ diverges rapidly to infinity as θ approaches b yielding $\int \sqrt{J(\theta)} d\theta = \infty$.

5 Ideas for the Proofs

5.1 Laplace integration

The main tool we use in this work is the Laplace integration. Using Taylor's theorem we have

$$\begin{aligned}
\frac{m_{\mathcal{G}}(x^n)}{p(x^n|\hat{\theta})} &= \int \frac{p(x^n|\theta) w_{\mathcal{G}}(\theta)}{p(x^n|\hat{\theta})} d\theta \\
&\sim \int \exp\left(-\frac{n(\theta-\hat{\theta})^t \hat{J}(\hat{\theta})(\theta-\hat{\theta})}{2}\right) w_{\mathcal{G}}(\theta) d\theta \\
&\sim \frac{w_{\mathcal{G}}(\hat{\theta})}{\sqrt{\det(\hat{J}(\hat{\theta}))}} \frac{(2\pi)^{d/2}}{n^{d/2}} \\
&= \frac{\sqrt{\det(J(\hat{\theta}))}}{C_J(\mathcal{G})\sqrt{\det(\hat{J}(\hat{\theta}))}} \frac{(2\pi)^{d/2}}{n^{d/2}},
\end{aligned}$$

where $\hat{J}(\theta)$ is empirical Fisher information matrix (Hessian of $-\log p(x^n|\theta)/n$). For exponential families, $\hat{J}(\hat{\theta})$ equals Fisher information $J(\hat{\theta})$. This can be confirmed by noting (1) and

$$\log p(x^n|\theta) = n(\theta \cdot \bar{x} - \psi(\theta)),$$

where $\bar{x}$ is the average of x in x^n.

Therefore, we have

$$\frac{\sqrt{\det(J(\hat{\theta}))}}{C_J(\mathcal{G})\sqrt{\det(\hat{J}(\hat{\theta}))}} = \frac{1}{C_J(\mathcal{G})},$$

which implies

$$\frac{m_{\mathcal{G}}(x^n)}{p(x^n|\hat{\theta})} \sim \frac{(2\pi)^{d/2}}{n^{d/2} C_J(\mathcal{G})}. \tag{7}$$

This asymptotics hold when $\hat{\theta}$ stays interior to $\mathcal{G}$. For the sequence for which $\hat{\theta}$ is near boundary of Θ, we use different techniques.

5.2 For curved families

For curved exponential families the above asymptotics do not hold for almost all x^n, since the difference between $\hat{J}(\hat{\theta})$ and $J(\hat{\theta})$ is not zero for almost all x^n. However, if $|\hat{\theta} - \phi(\hat{u})| \leq a_n = o(1)$ (recall that $\hat{\theta} = \arg\max_\theta p(x^n|\theta)$) as $n \to \infty$, then we can derive $\det(\hat{J}(\hat{\theta}))/\det(J(\hat{\theta})) = 1+O(a_n) = 1 + o(1)$, hence we have (7) for such sequences x^n. Hence, recalling $\epsilon_n = o(1)$, we have

$$\frac{m_n^{(curve)}(x^n)}{p_c(x^n|\hat{u})} \geq \frac{(1-\epsilon_n)\int p_c(x^n|u)w_{\mathcal{U}_n}(u)du}{p_c(x^n|\hat{u})} \sim \frac{(2\pi)^{d/2}}{n^{d/2}C_J(\mathcal{U})}. \tag{8}$$

For x^n with $|\bar{x} - \eta(\phi(\hat{u}))| \geq a_n$, we have

$$\frac{p(x^n|\hat{\theta})}{p_c(x^n|\hat{u})} = \frac{p(x^n|\hat{\theta})}{p(x^n|\phi(\hat{u}))} = \exp(nD(\hat{\theta}|\phi(\hat{u}))) \geq \exp(Cna_n^2),$$

where $D(\theta|\theta')$ is Kullback-Leibler divergence of $p(x|\theta')$ with respect to $p(x|\theta)$ and C is a certain positive real number. Also, we have the following from (7).

$$\frac{\int p(x^n|\theta)w_{\mathcal{G}}(\theta)d\theta}{p(x^n|\hat{\theta})} \geq \frac{C'}{n^{\bar{d}/2}}.$$

Therefore, we have

$$\frac{\int p(x^n|\theta)w_{\mathcal{G}}(\theta)d\theta}{p_c(x^n|\hat{u})} \geq \frac{C'\exp(Cna_n^2)}{n^{\bar{d}/2}}.$$

Now we let $a_n^2 = n^{-1/2}$, then we have

$$\frac{\int p(x^n|\theta)w_{\mathcal{G}}(\theta)d\theta}{p_c(x^n|\hat{u})} \geq \frac{C'\exp(Cn^{1/2})}{n^{\bar{d}/2}}.$$

Recalling ϵ_n is converging to zero at a certain polynomial rate, we see that

$$\frac{m_n^{(curve)}(x^n)}{p_c(x^n|\hat{u})} \geq \frac{\epsilon_n \int p(x^n|\theta)w_{\mathcal{G}}(\theta)d\theta}{p_c(x^n|\hat{u})} \geq \frac{C'\epsilon_n \exp(Cn^{1/2})}{n^{\bar{d}/2}} \geq \frac{V}{n^{d/2}}$$

holds for sufficiently large n, where V is an arbitrary large constant. Together with (8), we have obtained the minimax answer.

6 Extension to General Smooth families

Here we address the the extension of the idea for curved exponential family to general smooth family $M = \{p(x|u) : u \in \mathcal{U} \subset \Re^d\}$.

Let $J(u)$ be a Fisher information matrix of u and $\hat{J}(x^n|u)$ be an emprical Fisher information matrix of u. Define d^2-dimensional vector valued randum variable $q(x^n|u)$ as $q_{jd+i}(x^n|u) = \hat{J}_{ij}(x^n|u) - J_{ij}(u)$.

We define the enlarged family S :

$$S = \{p_e(x|u,v) = p(x|u)\exp(q(x|u)\cdot v - \psi(u,v)) : u \in \mathcal{U}, |v| \leq b\},$$

where

$$\psi(u,v) = \log \int p(x|u)\exp(q(x|u)\cdot v)\nu(dx)$$

and b is a certain positive number. Here we are assuming that $\psi(u,v) < \infty$ for $u \in \mathcal{U}_s \subset \mathcal{U}$ and $|v| \leq b$. We let $\theta = (u,v)$. Here, M is a smooth subfamily of S. Under this setting the Bayes procedure (5) is asymptotically minimax for the set $\{x^n : \hat{u} \in \mathcal{U}_s\}$. The proof is similar to the case of curved exponential families.

7 Minimax Expected Regret

For the lower bound on maximin expected regret, The result by Clarke and Barron [6] is known, i.e. for d-dimensional smooth families,

$$\liminf_{n\to\infty}(\underline{R}_n(\mathcal{G}) - \frac{d}{2}\log\frac{n}{2\pi e}) \geq \log C_J(\mathcal{G}).$$

holds. This can be applied to d-dimensional steep exponential families. We note that in [6] corresponding upper bounds were only obtained for $\mathcal{G}$ compact and in the interior of Θ. Here, we give tools to handle the boundary behavior. For lower bound, the work of [6] is sufficient to handle arbitrary $\mathcal{G}$.

Recall that the minimax expected regret is

$$\bar{R}_n(\mathcal{G}) = \inf_q \sup_{\theta\in\mathcal{G}} E_\theta(\log\frac{p(x^n|\theta)}{q(x^n)}).$$

We can transform it as

$$E_\theta(\log\frac{p(x^n|\theta)}{q(x^n)}) = E_\theta(\log\frac{p(x^n|\hat{\theta})}{q(x^n)}) + E_\theta(\log\frac{p(x^n|\theta)}{p(x^n|\hat{\theta})}).$$

Since we can evaluate an upper bound on $E_\theta(\log(p(x^n|\hat{\theta})/q(x^n)))$ by using the upper bound on minimax pointwise regret, if we obtain an upper bound on $E_\theta(\log(p(x^n|\theta)/p(x^n|\hat{\theta})))$, then we can evaluate the upper bound on $\bar{R}_n(\mathcal{G})$.

In fact for one-dimensional exponential families, we can show that the minimax strategies for pointwise regret are minimax for expected regret as well. The tool we use for this proof is one of large deviation inequalities [7, 4].

8 Conclusions

To summarize the answer,

$$\frac{d}{2}\log\frac{n}{2\pi} + \log\int\sqrt{\det(J(\theta))}d\theta$$

given for the stochastic complexity in Rissanen [12] and given in Clarke and Barron [6] for related minimax redundancy (expected regret) remains valid for minimax regret when dealing with exponential families of various boundary behavior and with curved exponential families and in each case is achieved by modifications of Jeffreys prior in some cases analogous to those suggested by Xie and Barron [14, 15].

References

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

[2] R. N. Bhattacharya, "Rates of weak convergence and asymptotic expansions for classical central limit thorems", *Ann. of Math. Statist.*, vol. 42, no. 1, 241-259, 1971.

[3] R. N. Bhattacharya and R. Rao, *Normal approximation and asymptotic expansions*, John Wiley & Sons, New York, 1976.

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

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

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

[7] I. Csiszàr, "Sanov property, generalized I-projection and a conditional limit theorem," *Ann. of probability*, vol. 12, No. 3, pp. 768-793, 1984.

[8] L. Davisson & A. Leon-Garcia, "A source matching approach to finding minimax codes," *IEEE Trans. Inform. Theory,* vol. 26, pp. 166-174, March 1980.

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

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

[11] T. Matsushima, H. Inazumi & S. Hirasawa, "A class of distortionless codes designed by Bayes decision theory," *IEEE Trans. Inform. Theory,* vol. 37, pp. 1288-1293, 1991.

[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] Q. Xie & A. R. Barron, "Minimax redundancy for the class of memoryless sources", *IEEE trans. Inform. Theory,* 1997.

[15] Q. Xie & A. R. Barron, "Asymptotic minimax regret for data compression, gambling and prediction," preprint, May 1996.