Asymptotically minimax regret by Bayes mixtures

Jun-ichi Takeuchi[1]
C&C Media Res. Labs., NEC Corporation
Kanagawa 2168555, Japan
Email tak@ccm.CL.nec.co.jp

Andrew R. Barron
Dept. of Statistics, Yale University
P.O. Box 208290, New Haven, CT 06520, USA
Email Andrew.Barron@yale.edu

We study the problem of data compression, gambling and prediction of a sequence $x^n = x_1x_2...x_n$ from a certain alphabet $\mathcal{X}$, in terms of regret [4] and redundancy with respect to a general exponential family, a general smooth family, and also Markov sources. In particular, we show that variants of Jeffreys mixture asymptotically achieve their minimax values. These results are generalizations of the work by Xie and Barron [5, 6] in the general smooth families. In particular for one-dimensional exponential families, they also extend the works of Clarke and Barron [1] to deal with the full natural parameter space rather than compact sets interior to it.

The worst case regret of a probability density q with respect to a d-dimensional family of probability densities $S = \{p(\cdot|\theta) : \theta \in \Theta\}$ and a set of the sequences $W_n \subseteq \mathcal{X}^n$ is defined as

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

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

We consider minimax problems for sets of sequences such that $W_n = \mathcal{X}^n(\mathcal{G}) \stackrel{\text{def}}{=} \{x^n : \hat{\theta} \in \mathcal{G}\}$, where $\mathcal{G}$ is a certain nice subset (satisfies $\bar{\mathcal{G}} = \bar{\mathcal{G}^\circ}$) of Θ. That is, we construct probability densities which minimize the worst case regret.

We let $J(\theta)$ denote the Fisher information matrix of θ. For $K \subset \Theta$, define $C_J(K) \stackrel{\text{def}}{=} \int_K \sqrt{\det(J(\theta))}d\theta$. The Jeffreys mixture is the mixture density by the prior proportional to $\sqrt{\det(J(\theta))}$. We denote the Jeffreys mixture over the set K by m_K. The value $C_J(K)$ is the normalization constant for the Jeffreys prior over the set K.

For (i.i.d.) exponential families, a sequence of Jeffreys mixtures achieves the minimax regret asymptotically, when $\mathcal{G}$ is a compact subset included in the interior 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 we have

$$\log \frac{p(x^n|\hat{\theta})}{m_{\mathcal{G}_n}(x^n)} \sim \frac{d}{2}\log\frac{n}{2\pi} + \log C_J(\mathcal{G}) + o(1), \qquad (1)$$

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

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

where $\hat{J}(x^n, \theta)$ is the empirical Fisher information. When θ is the natural parameter of an exponential family, $\hat{J}(x^n, \hat{\theta}) = J(\hat{\theta})$ holds.

[1]This work was done while Takeuchi was staying at Yale University.

However, when the model S is not exponential type, the Jeffreys mixture is not minimax, because the empirical Fisher information is not close to the Fisher information at the MLE for a small set of sequences. For such S, a modified Jeffreys mixture obtained by adding the small contribution from a mixture of the enlarged set which consists of densities

$$p_e(x|\theta,\xi) \stackrel{\text{def}}{=} p(x|\theta)\exp\Big(\sum_{i\leq j}\big(\hat{J}_{ij}(x,\theta) - J_{ij}(\theta)\big)\xi_{ij}\Big)/\Lambda(\theta,\xi),$$

where ξ is added parameters and $\Lambda(\theta, \xi)$ is the normalization constant. The added component deals with the sequences for which the empirical Fisher information differs from the Fisher information. When S is the curved exponential family embedded in the exponential family $\bar{S}$, we can use $\bar{S}$ as the enlarged model.

Certain exponential families that permit dependence (not i.i.d.) can also be handled, in particular for the first order Markov sources with the alphabet size 2, we obtain modifications of the Jeffreys mixture, which achieve the minimax regret for all sequences (including the case the MLE near the boundary), similarly to Xie and Barron's result for discreet memoryless sources [6].

Also for one-dimensional exponential families, we show how to modify the Jeffreys mixture for dealing with the sequences with the MLE locating near the naturally determined boundary of the range of parameters. For this task, we have classified the types of boundary or tail behavior for one-dimensional exponential families. Our modification must be designed depending on such types. The same modifications are also minimax for the traditional redundancy $E_\theta[\log(p(x^n|\theta)/q(x^n))]$.

To summarize the answer, the asymptotics like (1) given for the stochastic complexity in Rissanen [3] and given in Clarke and Barron [1] for related minimax redundancy (expected regret) remain valid for minimax regret when dealing with exponential families of various boundary behavior and with general smooth families and in each case is achieved by modifications of Jeffreys mixture in some cases analogous to those suggested by Xie and Barron [5, 6].

References

[1] B. Clarke & A. R. Barron, "Jeffreys prior is asymptotically least favorable under entropy risk," *JSPI*, 1994.

[2] L. Davisson & A. Leon-Garcia, "A source matching approach to finding minimax codes," *IEEE Trans. Inform. Theory*, 1980.

[3] J. Rissanen, "Fisher information and stochastic complexity," *IEEE trans. Inform. Theory*, 1996.

[4] Yu M. Shtarkov, "Universal sequential coding of single messages," *Problems of Information Transmission*, 1988.

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

[6] ———, "Asymptotic minimax regret for data compression, gambling and prediction," *IEEE trans. Inform. Theory*, 1998.