Robustly Minimax Codes for Universal Data Compression

Jun-ichi Takeuchi* Andrew R. Barron†

Abstract— We introduce a notion of 'relative redundancy' for universal data compression and propose universal codes which asymptotically achieve the minimax value of the relative redundancy. The relative redundancy is a hybrid of redundancy and coding regret (pointwise redundancy), where a family of information sources and a family of codes are assumed. The minimax code for relative redundancy is an extension of the modified Jeffreys mixture, which was introduced by Takeuchi and Barron and is minimax for regret.

Keywords— universal coding, redundancy, regret, Bayes mixture, Jeffreys prior, robust learning

1 Introduction

We introduce a notion of 'relative redundancy' for universal data compression and sequential prediction, and propose procedures which asymptotically achieve the minimax value of the relative redundancy.

For universal data compression, the notion of redundancy, which is the difference between the expected code length of the code in concern and the expected code length of the Shannon code defined by the true source, has been often used as a criterion. On the other hand these days, there is a different criterion called 'regret' or 'pointwise redundancy', for which any true information source is not assumed [10]. Instead, a family of codes, which is the family of opponents for the code in concern, is assumed. Given a data sequence, the regret of a code is defined as the difference between the code length by the code in concern and the code length by the hindsight best code among the assumed family of codes. The relative redundancy is a hybrid of those two notions of redundancy.

For the relative redundancy, we assume both a family of true sources and a family of codes. (In the case of ordinary redundancy, the both are same to each other.) The relative redundancy is the difference between the expected code length of the code in concern and the expected code length of the best code among the assumed family of codes, where expectation is taken with respect to the true source, an element of the family of true sources. We are especially interested in the case that the family of true sources is non parametric and properly includes the family of codes. This setting requires the codes of a certain robustness.

This kind of robust setting was proposed in the field of computational learning theory and is known as 'robust PAC (Probably Approximate Correct) learning model' [8]. Also, the similar setting is used for the problem of sequential prediction[15]. It is related to universal coding and a certain aspect of it is a generalization of universal coding. However, in those contexts, any minimax procedure (upto the constant order) is not known to date.

When extending the notion of redundancy as the above, the minimax codes for traditional redundancy obtained by Bayes procedure with Jeffreys prior [5, 6] are no longer minimax. Similarly, the minimax codes for regret obtained by normalizing maximum likelihood [10, 9] or Bayes procedures [11, 4, 12] are not minimax. Takeuchi and Barron [12] showed that the minimax codes for redundancy are not minimax for the regret in usual cases and they obtained minimax codes, which are mixtures of enlarged family of codes. This enlargement is obtained by using the difference between the Fisher information and the empirical Fisher information. Usually, the above code is not minimax for our relative redundancy and we need another modification. Minimax codes for the relative redundancy are a mixture over an enlarged family of codes, which is obtained by utilizing the difference between the empirical covariance of score functions and the Fisher information, adding the difference between the Fisher information and the empirical Fisher information. It might be interesting that the value of the asymptotic minimax relative redundancy is same as that of the asymptotic minimax redundancy upto constant order.

2 Relative Redundancy

Let $\mathcal{X}$ be a measurable space and ν be a reference measure on $\mathcal{X}$. We define $\nu(dx^n) \stackrel{\text{def}}{=} \prod_{t=1}^n \nu(dx_t)$. We refer to p as the density of a stochastic process, if p satisfies $\int p(x_1)\nu(dx) = 1$ and $\int p(x^{n+1})\nu(dx_{n+1}) = p(x^n)$.

For now we assume that $\mathcal{X}$ is discreet and $\nu(\{x\}) = 1$ for $x \in \mathcal{X}$. Let q be a density of stochastic process. We can construct a code for the set of data $\mathcal{X}^n$ whose code length is given by $-\log q(x^n)$. Here, log is the natural logarithm. We measure code length by 'nat'. Conversely, when there exists a uniquely decodable code for $\mathcal{X}^n$ with code length $l(x^n)$, then $\int \exp(-l(x^n))\nu(dx^n) \leq 1$ (Kraft's inequality) holds. Hence, we refer to the density q as a code. We let C be a family of codes. We assume that C is a smooth parametric family: $C \stackrel{\text{def}}{=} \{p(\cdot|\theta) : \theta \in \Theta \subset \Re^d\}$. The circumstances for general $\mathcal{X}$ and ν are similar.

Let S be a certain family of densities of stochastic processes. We assume that data sequence $x^n \in \mathcal{X}^n$ is drawn from a certain $p \in S$, i.e. we refer to S as the

* RWCP (Real World Computing Partnership) c/o C&C Media Res. Labs., NEC Corp., Kawasaki, Kanagawa 216-8555, Japan. e-mail tak@ccm.CL.nec.co.jp

† Dept. of Statistics, Yale Univ., P.O. Box 208290, New Haven, CT 06520, USA. e-mail Andrew.Barron@yale.edu

family of information sources.

Let r be the true source, i.e. let r be an element of S. Let q be a code. Define the relative redundancy of the code q with respect to the true source r and the set of codes C as

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

Define the worst case relative redundancy of q for the pair (S,C) as

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

When S equals C, the relative redundancy coincides with the ordinary redundancy. Finally, the minimax relative redundancy for the pair (S,C) is defined as

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

The above definition is valid for general families of stochastic processes, but hereafter, we restrict C and S to families of i.i.d. processes, i.e. we assume that $p(x^n) = \prod_{t=1}^n p(x_t)$ for $p \in C \cup S$.

In that case, it is known that $R_n(C,C) = (d/2)\log(n/2\pi e) + \log C_J(\Theta) + o(1)$ holds [5, 6, 13]. Here, $o(1) \to 0$ and $C_J(\Theta) \stackrel{\text{def}}{=} \int_\Theta \sqrt{\det J(\theta)}d\theta$, where $J(\theta)$ is the Fisher information matrix defined as

$$J_{ij}(\theta) \stackrel{\text{def}}{=} E_\theta(-\frac{\partial^2 \log p(x|\theta)}{\partial\theta^i \partial\theta^j}).$$

In this paper, we show that $R_n(S,C) = (d/2)\log(n/2\pi e) + \log C_J(\Theta) + o(1)$ still holds for fairly general i.i.d. families of codes C.

We describe the definition of minimax regret $\bar{r}$ for reference, where W_n is a subset of $\mathcal{X}^n$:

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

For this, it is known [9, 14, 11, 4, 12, 3] that $\bar{r}_n(W_n,C) = (d/2)\log(n/2\pi) + \log C_J(\Theta) + o(1)$, where $W_n = \{x^n : \hat{\theta}(x^n) \in \Theta\}$ and $\hat{\theta}(x^n)$ is the maximum likelihood estimate given x^n.

3 Minimax Code

Let $\mathcal{P}$ be a set of all i.i.d. processes and C be a smooth parametric subset of $\mathcal{P}$: $C = \{p(\cdot|\theta) : \theta \in \Theta_c \subset \Theta\}$, where Θ_c is compact and included in Θ°. We refer to C as a family of codes. We define $\tilde{\theta}$ for $r \in \mathcal{P}$ as

$$\tilde{\theta} \stackrel{\text{def}}{=} \arg\min D(r|p(\cdot|\theta)),$$

where $D(r|p) \stackrel{\text{def}}{=} E_r \log(r(x)/p(x))$ (Kullback Leibler divergence). Note that we have $E_r \nabla \log p(x|\tilde{\theta}) = 0$, where we let ∇ denote the gradient with respect to θ. Define a set of probability densities as

$$S = \{r : \arg\min_{\theta\in\Theta} D(r|p(\cdot|\theta)) \in \Theta_c\}.$$

We will consider the relative redundancy for pair (S,C) (actually we must consider the problem for (S',C) where S' is a certain subset of S).

Let K_n be a compact subset of Θ such that $K_n^\circ \supset \Theta_c$ $(n \geq 0)$, $K_n \subset K_0^\circ$ $(n \geq 1)$ and K_n slowly shrinks to Θ_c as $n \to \infty$.

Now, we describe the conditions under which we construct the minimax code. Define a matrix $\hat{I}$ as

$$\hat{I}(x,\theta) \stackrel{\text{def}}{=} \frac{\partial p(x|\theta)}{\partial\theta^i}\frac{\partial p(x|\theta)}{\partial\theta^j}$$

and $\hat{I}(x^n,\theta) \stackrel{\text{def}}{=} (1/n)\sum_{t=1}^n \hat{I}(x_t,\theta)$. We assume $E_\theta \hat{I}(x,\theta) = J(\theta)$. Define a d^2-dimensional vector valued random variable $V(x,\theta)$ as $V_{dj+i}(x^n,\theta) \stackrel{\text{def}}{=} \hat{I}_{ij}(x^n,\theta) - J_{ij}(\theta)$. Note that $E_\theta V(x,\theta) = 0$. Define $I(r,\theta) \stackrel{\text{def}}{=} E_r \hat{I}(x,\theta)$. In particular, we let $I(r) = I(r,\tilde{\theta})$. Note that $I(p(\cdot|\theta)) = J(\theta)$.

Define a d^2-dimensional vector valued random variable

$$U_{dj+i}(x^n,\theta) \stackrel{\text{def}}{=} \hat{J}_{ij}(x^n,\theta) - J_{ij}(\theta),$$

where $\hat{J}(x^n,\theta)$ is the empirical Fisher information:

$$\hat{J}(x^n,\theta) \stackrel{\text{def}}{=} -\frac{1}{n}\sum_{t=1}^n \frac{\partial^2 \log p(x_t|\theta)}{\partial\theta^i\partial\theta^j}.$$

Condition 1 *Let $D = [-b,b]^{d^2}$ for a certain $b > 0$. For a certain $C_0 > 0$, the following holds.*

$$\begin{aligned}&\forall(u,v) \in D^2,\ \forall\theta \in K_0,\\ &E_\theta \exp(v\cdot V(x,\theta) + u\cdot U(x,\theta)) < C_0\end{aligned} \qquad (1)$$

Note that the above implies the following for any $\gamma > 0$.

$$E_\theta(|V(x,\theta)|^\gamma + |U(x,\theta)|^\gamma)e^{v\cdot V(x,\theta)+u\cdot U(x,\theta)} < C(\gamma).$$

We consider the family of sources S' which satisfies the following.

Condition 2 *For a certain C_0, the following holds.*

$$\exists\beta > 0,\ \forall r \in S',\ E_r(|U(x,\theta)|^\beta + |V(x,\theta)|^\beta) < C_0. \quad (2)$$

Under Condition 1, we define an extended family of codes $\bar{C}$:

$$\bar{C} \stackrel{\text{def}}{=} \{p_e(x|\theta,u,v) = \frac{p(x|\theta)e^{u\cdot U(x,\theta)+v\cdot V(x,\theta)}}{\lambda(\theta,u,v)} : \theta \in K_0,\ (u,v) \in D^2\},$$

where $\lambda(\theta,u,v) \stackrel{\text{def}}{=} E_\theta e^{u\cdot U(x,\theta)+v\cdot V(x,\theta)}$. This is the normalization constant. Note that $\bar{C}$ is $(d+2d^2)$-dimensional and the original family C is smoothly embedded in the enlarged family $\bar{C}$.

Let m_{K_t} be Bayes mixture $\int_{K_t} p(\cdot|\theta)w_{K_t}(\theta)d\theta$ with Jeffreys prior: $w_{K_t}(\theta) \stackrel{\text{def}}{=} \sqrt{\det J(\theta)}/C_J(K_t)$.

Let $\xi \stackrel{\text{def}}{=} (\theta,u,v)$. The following is our new code.

$$m_n(x^n) = (1-\epsilon_n)m_{K_n}(x^n) + \epsilon_n \int p_e(x^n|\xi)w(\xi)d\xi,$$

where $w(\xi)$ is some smooth prior with $\inf_\xi w(\xi) > 0$ and $\epsilon_n > 0$ approaches to 0 at polynomial rate.

Remark: If $\mathcal{X}$ is a finite set, the above procedure is drastically reduced to easy one. In that case, we can use the family $\mathcal{P}$, which is finite dimensional, as the enlarged family $\bar{C}$.

We have the following theorem, which claims that the above m_n is asymptotically minimax for relative redundancy.

Theorem 1 *Under Conditions 1 and 2, and certain regularity conditions for the family C, the following holds.*

$$\limsup_{n\to\infty}(R_n(m_n, S', C) - \frac{d}{2}\log\frac{n}{2\pi e}) \le \log C_J(\Theta_c).$$

We note about the lower bound. The inequality $R_n(q, C, C) \le R_n(q, S', C)$ holds, whenever $S' \supseteq C$. Hence, a lower bound for the minimax redundancy is that for the minimax relative redundancy and we can see that the above upper bound matches the lower bound.

Outline of the Proof

We make a case argument about r. Let $a_n = n^{-1/4}$. There are two cases: (i) $|E_r U(x,\tilde{\theta})| + |E_r V(x,\tilde{\theta})| \le 2a_n$ holds. (ii) $|E_r U(x,\tilde{\theta})| \ge a_n$ or $|E_r V(x,\tilde{\theta})| \ge a_n$ hold.

We let $S_n^{(i)}$ and $S_n^{(ii)}$ respectively denote the sets of $r \in S'$ which satisfies the cases (i) and (ii). We use the notation $l_n = (1/n)\nabla \log p(x^n|\tilde{\theta})$.

First, consider the case (i). In this case, we follow the argument used in [5, 6]. We have $|\hat{J}(x^n,\tilde{\theta}) - J(\tilde{\theta})| = |U(x^n,\tilde{\theta})| \le 4a_n = o(1)$ with high probability. Hence the Laplace integration about m_{K_n} works. The following holds with high probability, where we let $\delta\theta = \theta - \tilde{\theta}$.

$$\begin{aligned}\frac{1}{n}\log\frac{p(x^n|\theta)}{p(x^n|\tilde{\theta})} &= \delta\theta^t l_n - \frac{\delta\theta^t \hat{J}(x^n,\theta')\delta\theta}{2} + O(|\delta\theta|^3) \\ &= \delta\theta^t l_n - \frac{\delta\theta^t J(\tilde{\theta})\delta\theta}{2} + O(|\delta\theta^t|^3),\end{aligned}$$

where $\theta' = \lambda\theta + (1-\lambda)\tilde{\theta}$ for a certain $\lambda \in [0,1]$. Let $h \stackrel{\text{def}}{=} (J(\tilde{\theta}))^{-1} l_n$ and $\tilde{J}$ denote $J(\tilde{\theta})$, then we have

$$\begin{aligned}&\frac{1}{n}\log\frac{p(x^n|\theta)}{p(x^n|\tilde{\theta})} \\ =& -\frac{(\delta\theta - h)^t\tilde{J}(\delta\theta - h)}{2} + \frac{h^t\tilde{J}h}{2} + O(|\delta\theta|^3) \\ =& -\frac{(\delta\theta^t - h)^t\tilde{J}(\delta\theta^t - h)}{2} + \frac{l_n^t\tilde{J}^{-1}l_n}{2} + O(|\delta\theta|^3).\end{aligned}$$

Hence we have

$$\frac{p(x^n|\theta)}{p(x^n|\tilde{\theta})} = e^{-n((\theta-\tilde{\theta}-h)^t\tilde{J}(\theta-\tilde{\theta}-h) - l_n^t\tilde{J}^{-1}l_n + O(|\delta\theta|^3))/2}.$$

Note that $|h| \le \log n/n$ with high probability. Evaluating the integration $\int p(x^n|\theta)w_{K_n}(\theta)d\theta/p(x^n|\tilde{\theta})$ with contribution from the neighborhood of $\tilde{\theta} + h$, we have

$$\frac{m_{K_n}(x^n)}{p(x^n|\tilde{\theta})} \sim \frac{e^{nl_n^t\tilde{J}^{-1}l_n/2}(2\pi)^{d/2}}{C_J(K_n)n^{d/2}}$$

with high probability. Hence, we have

$$\frac{m_n(x^n)}{p(x^n|\tilde{\theta})} \gtrsim \frac{(1-\epsilon_n)e^{nl_n^t\tilde{J}^{-1}l_n/2}(2\pi)^{d/2}}{C_J(K_n)n^{d/2}}.$$

Hence, the following holds with high probability.

$$\log\frac{p(x^n|\tilde{\theta})}{m_n(x^n)} \lesssim \frac{d}{2}\log\frac{n}{2\pi} + \log C_J(\Theta_c) - \frac{nl_n^t\tilde{J}^{-1}l_n}{2}.$$

Noting $E_r l_n l_n^t = I(r)/n$, it is possible to show

$$\begin{aligned}&E_r\log\frac{p(x^n|\tilde{\theta})}{m_n(x^n)} \qquad (3)\\ \lesssim& \frac{d}{2}\log\frac{n}{2\pi} + \log C_J(\Theta_c) - \frac{\mathrm{tr}(I(r)J^{-1})}{2}.\end{aligned}$$

When $r \in S_n^{(i)}$, we have $|I(r) - J(\tilde{\theta})| = |E_r V(x^n,\tilde{\theta})| \le 2a_n$. Hence,

$$\sup_{r\in S_n^{(i)}}(-\mathrm{tr}(I(r)\tilde{J}^{-1})/2) \sim -d/2. \qquad (4)$$

Therefore, we have

$$\sup_{r\in S_n^{(i)}} E_r\log\frac{p(x^n|\tilde{\theta})}{m_n(x^n)} \lesssim \frac{d}{2}\log\frac{n}{2\pi e} + \log C_J(\Theta_c). \qquad (5)$$

Now, we consider the case (ii). We have $|U(x^n,\tilde{\theta})| + |V(x^n,\tilde{\theta})| \ge a_n/2$ with high probability. Let

$$(\tilde{u},\tilde{v}) = \frac{\alpha a_n(U(x^n,\tilde{\theta}), V(x^n,\tilde{\theta}))}{\sqrt{|U(x^n,\tilde{\theta})|^2 + |V(x^n,\tilde{\theta})|^2}},$$

where α is a certain small positive number. Then,

$$\frac{p_e(x^n|\tilde{\theta},\tilde{u},\tilde{v})}{p(x^n|\tilde{\theta})} = \frac{e^{n(\tilde{u}\cdot U(x^n,\tilde{\theta}) + \tilde{v}\cdot V(x^n,\tilde{\theta}))}}{(\Lambda(\tilde{\theta},\tilde{u},\tilde{v}))^n} \ge e^{C_1 n a_n^2} \qquad (6)$$

holds with high probability. We can easily show that

$$\frac{\int p_e(x^n|\xi)w(\xi)d\xi}{p_e(x^n|\tilde{\theta},\tilde{u},\tilde{v})} \ge \frac{C_2}{n^{d+2d^2}}.$$

Therefore,

$$\begin{aligned}&\frac{\epsilon_n\int p_e(x^n|\xi)w(\xi)d\xi}{p(x^n|\tilde{\theta})} \\ =& \frac{\epsilon_n\int p_e(x^n|\xi)w(\xi)d\xi}{p_e(x^n|\tilde{\theta},\tilde{u},\tilde{v})}\frac{p_e(x^n|\tilde{\theta},\tilde{u},\tilde{v})}{p(x^n|\tilde{\theta})} \\ \ge& \frac{C_2\epsilon_n\exp(C_1 n a_n^2)}{n^{d+2d^2}} = \frac{C_2\epsilon_n\exp(C_1\sqrt{n})}{n^{d+2d^2}}\end{aligned}$$

holds with high probability. Hence for $r \in S_n^{(ii)}$,

$$E_r\log\frac{m_n(x^n)}{p(x^n|\tilde{\theta})} \ge E_r\log\frac{\epsilon_n\int p(x^n|\xi)w(\xi)d\xi}{p(x^n|\tilde{\theta})} \to \infty$$

holds. Together with (5), this implies

$$\sup_{r\in S'} E_r\log\frac{p(x^n|\tilde{\theta})}{m_n(x^n)} \le \frac{d}{2}\log\frac{n}{2\pi e} + \log C_J(\Theta_c) + o(1).$$

4 Discussion

4.1 Semi Universality

The minimax codes have the property of 'semi universality' ([7]). Let us take the code based on the Bayes mixture with the Jeffreys prior m_{K_n}, which is asymptotically minimax for the redundancy $R_n(q, C, C)$. The expectation of its code length per data symbol approaches to the entropy rate of the true source, when the true source r belongs to the family C. This dose not hold, when r is not an element of C. For m_{K_n}, (3) holds as an (approximated) equality rather than an inequality, i.e. we have

$$E_r \log \frac{p(x^n|\tilde{\theta})}{m_{K_n}(x^n)} = \frac{d}{2} \log \frac{n}{2\pi} + O(1).$$

Hence, we have

$$\lim_{n\to\infty} \frac{1}{n} E_r \log \frac{1}{m_{K_n}(x^n)} = \lim_{n\to\infty} \frac{1}{n} E_r \log \frac{1}{p(x^n|\tilde{\theta})}$$
$$= E_r \log \frac{1}{p(x|\tilde{\theta})} = H(r) + D(r|p(\cdot|\tilde{\theta})),$$

where H denotes differential entropy rate and D denotes the Kullback Leibler divergence. Hence, the expected code length by m_{K_n} per source symbol does not converge to the entropy rate. This property is called as 'semi universality'.

Concerning m_n, when $r \notin C$, the relative redundancy is negative. Hence, the code length by m_n is shorter than that by $p(\cdot|\tilde{\theta})$, but the expected code length per source symbol does not converge to the entropy rate.

4.2 Necessity of Enlarging Model

We have succeeded to construct the asymptotically minimax codes for the relative redundancy by enlarging the family of codes. Here, we consider why this enlargement is needed. Two information matrices, $\hat{J}(x^n, \tilde{\theta})$ and $I(r)$ are important. When the true source r belongs to the family of codes C, then both $I(r)$ and the expectation of $\hat{J}(x^n, \tilde{\theta})$ equal the Fisher information $J(\tilde{\theta})$. Then, the asymptotics (3) and (4) hold. However, if r is displaced from C, then I and $\hat{J}$ is different from J with high probability. This spoils (3) and (4). However in that case, the contribution from enlarged family works well, utilizing $\hat{I} - J$ or $\hat{J} - J$. When we treat not the relative redundancy but the regret, then we do not have to care the asymptotic (4). Therefore, the enlargement for the regret uses $\hat{J} - J$ alone.

Finally, the author would like to briefly note about the differential geometrical interpretation (see [1, 2]). The quantity $\hat{J} - J$ is related to exponential curvature of the family C. When $\hat{J}(x^n, \hat{\theta}) - J(\hat{\theta})$ always equals zero, then C is an exponential family. Also, The quantity $\hat{I} - \hat{J}$ is related to the mixture curvature of the family C. When $\hat{I}(x^n, \hat{\theta}) - \hat{J}(x^n, \hat{\theta})$ always equals zero, C is a finite mixture model. Since our enlargement is spanned by $\hat{J} - J$ and $\hat{I} - J$, this is equivalent to the enlargement spanned by $\hat{J} - J$ and $\hat{I} - \hat{J}$. Therefore, our enlargement is to the direction of both exponential and mixture curvature.

Acknowledgements: The one of the authors (Takeuchi) would like to express his sincere gratitude to Prof. Tsutomu Kawabata of Univ. of Electro-Communications for his helpful advice.

References

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

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

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

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

[5] B. Clarke & A. R. Barron, "Information-theoretic asymptotics of Bayes methods," *IEEE trans. 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," *JSPI,* 41:37-60, 1994.

[7] T. S. Han & K. Kobayashi, *Mathematics of information and coding* (in Japanese), Iwanami Publisher, 1994.

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

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

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

[11] J. Takeuchi & A. R. Barron, "Asymptotically minimax regret for exponential families," *Proc. of SITA'97,* pp. 665-668, 1997.

[12] J. Takeuchi & A. R. Barron, "Asymptotically minimax regret by Bayes mixtures," *in preparation,* an abstract appeared in *Proc. of 1998 IEEE ISIT,* 1998.

[13] Q. Xie & A. R. Barron, "Minimax redundancy for the class of memoryless sources", *IEEE trans. IT,* vol. 43, no. 2, pp. 646-657, 1997.

[14] Q. Xie & A. R. Barron, "Asymptotic minimax regret for data compression, gambling and prediction," to appear, *IEEE trans. IT,* 1998.

[15] K. Yamanishi, "A decision-theoretic extension of stochastic complexity and its application to learning," *IEEE trans. IT,* vol. 44, no. 4, 1998.