Redundancy of universal coding, Kolmogorov complexity, and Hausdorff dimension (Abstract)

Hayato Takahashi [*]

January 28, 2003

In [1, 4, 5], under a suitable condition, it is shown that asymptotic code-lengths of sequences generated by a parametric model P_θ is given as follows;

$$-\log P_{\hat{\theta}} + \frac{k}{2}\log n + o(\log n),\ \ P_\theta - a.e., \tag{1}$$

where $\hat{\theta}$ is the maximum-likelihood estimator, k is the dimension of parameter space, n is the sample size, and the base of log is 2.

In view of the proof of Rissanen [4], the second term of (1) is the description of the maximum likelihood estimator $\hat{\theta}$ with $(\log n)/2$ bit accuracy, therefore, it is natural to study a universal coding obtained by compressing the description of the maximum likelihood estimator. In fact, Vovk [6] studied a universal coding for Bernoulli model with code-length

$$\inf_\theta -\log P_\theta + K(\theta|n), \tag{2}$$

where θ ranges over computable real, and K is the prefix Kolmogorov complexity [2, 3].

In order to study the code (2), we study asymptotic expansion of Bayes mixture $\int P_\theta dm(\theta)$ with two kind of priors. One is a prior that is singular with respect to Lebesgue measure, and another is a priori probability on Euclidean space.

By considering prior of Bayes mixture to be a priori probability on Euclidean space, we extend the universal coding (2) to multidimensional parameter space, and show a universal coding whose code-length is

$$-\log P_{\hat{\theta}} + \sum_{j=1}^{k} K(\text{description of } \theta^j \text{ up-to } (\log n)/2 \text{ bit } |n) + O(\log\log n),\ P_\theta - a.e., \tag{3}$$

where $\theta = (\theta^1, \cdots, \theta^k)$. On the other hand Rissanen [5] showed that the code-length (1) is optimal up to $O(\log n)$ term except for parameters in a set of

---

[*] The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo, 106-8569, Japan, tel: 81-3-3446-1501(ext. 9701), fax: 81-3-5421-8750, E-mail: takahasi@ism.ac.jp

Lebesgue measure 0. Therefore we characterize the parameter set such that

$$\frac{\sum_{j=1}^{k} K(\text{description of } \theta^j \text{ up-to } (\log n)/2 \text{ bit } |n)}{\frac{k}{2}\log n} < 1$$

$(n \to \infty)$ with Hausdorff dimension. Consequently we show a universal coding having the following property: For each real numbers $h_j,\ 0 \leq h_j \leq 1,\ 1 \leq j \leq k$, there are subsets of parameter space, $H_1 \times \cdots \times H_k$ such that if $\theta^j \in H_j$, the code-length is

$$-\log P_{\hat{\theta}} + \frac{(\sum_{j=1}^{k} \dim H_j)}{2} \log n + o(\log n),\ P_\theta - a.e., \tag{4}$$

where $\dim H$ is the Hausdorff dimension of H. Also, we show that the code-length optimal up-to $O(\log n)$ term when the parameter space is unit interval.

Since the code-length of the universal coding (2) and (3) involves Kolmogorov complexity, we can not construct such the code effectively. To avoid this difficulty, we approximate Kolmogorov complexity in (2) and (3), by considering Bayes mixture with singular prior with respect to Lebesgue measure. Then we show a universal coding, which is constructive, such that the code-length is

$$-\log P_{\hat{\theta}} + \frac{h}{2} \log n + o(\log n),\ P_\theta - a.e., \tag{5}$$

where $h = -p\log p - (1-p)\log(1-p)$, and p is the relative frequency of 1 in the dyadic expansion of $\hat{\theta}$. Note that $h < 1 \Leftrightarrow p \neq 1/2$, i.e. the relative frequency of 1 in the dyadic expansion of $\hat{\theta}$ is biased then code-length (5) is asymptotically less than that of MDL coding. Also we show that the code (5) is optimal up-to $O(\log n)$ term for almost every θ with respect to the prior.

Finally we remark that the code-lengths shown in this paper give a non-trivial upper and lower bound of Kolmogorov complexity when the source is not a computable measure.

References

[1] A. R. Barron. *Logically smooth density estimation.* Ph.D. dissertation, Dept. Elec. Eng., Stanford Univ., Stanford, CA, Sept. 1985.

[2] G. J. Chaitin. A theory of program size formally identical to information theory. *J. ACM*, 22:329–340, 1975.

[3] L. A. Levin. Laws of information conservation (nongrowth) and aspects of the foundation of probability theory. *Prob. Inf. Transm.*, 10:206–210, 1974.

[4] J. Rissanen. Universal coding, information, prediction, and estimation. *IEEE Trans. Inform. Theory*, IT-30(4):629–636, 1984.

[5] J. Rissanen. Stochastic complexity and modeling. *Ann. Statist.*, 14(3):1080–1100, 1986.

[6] V. G. Vovk. Learning about the parameter of the Bernoulli model. *J. Comput. System Sci.*, 55:96–104, 1997.