PAPER

Performance of Data Compression in terms of Hausdorff Dimension*

Kouki HOJO[†], **Boris YA. RYABKO**[††], ***Nonmembers, and*** **Joe SUZUKI**[†], ***Member***

SUMMARY Currently, the most popular model in data compression theory is that of stationary ergodic sources. But there do exist sequences each of which is not emitted from any stationary ergodic source but can be compressed sufficiently by a certain algorithm. We estimate the size of the set of such sequences in terms of Hausdorff dimension.

key words: *Hausdorff dimension, stationary ergodic source, typical sequence, Kolmogorov complexity.*

1. Introduction

Let $B = \{0,1\}$, $B^n = \{0,1\}^n$, $B^* = \bigcup_{n\geq 0} B^n$, $B^\infty = \{0,1\}^\infty$. Let ν be a probability measure on B^∞. $x_1^\infty \in B^\infty$ is defined to be typical for ν if for any $u \in B^*$,

$$\lim_{n\to\infty} \frac{\nu_u(x_1^n)}{n-|u|+1} = \nu(u),$$

where $\nu_u(x_1^n)$ is the number of occurrences of u in x_1^n, and

$$M := \{x_1^\infty \in B^\infty : x_1^\infty \text{is typical for some stationary ergodic measure}\}.$$

For $x_1^\infty \in M$, let $H(x_1^\infty)$ be the Shannon entropy of the measure for which x_1^∞ is typical. For $0 \leq \alpha \leq 1$, we define two subsets of B^∞ as follows.

$$M(\alpha) := \{x_1^\infty \in M : H(x_1^\infty) \leq \alpha\},$$

and

$$K(\alpha) := \{x_1^\infty \in B^\infty : \liminf_{n\to\infty} \frac{KC(x_1^n)}{n} \leq \alpha\},$$

where $KC(x_1^n)$ is the Kolmogorov complexity [4] of $x_1^n \in B^n$. Intuitively, $M(\alpha)$ is the set of sequences that should be compressed by universal data compression at a rate less than or equal to α, and $K(\alpha)$ is the set of sequences each of which can be compressed by a certain algorithm at a rate less than or equal to α. (See Section 3 for further discussion.)

Let ν be a nonatomic probability measure over the unit interval $[0,1]$, and $M \subset [0,1]$. Then, a ν-ρ-covering of M is that of M by intervals v_i, $i = 0,1,2,\ldots$, with $\nu(v_i) < \rho$. For $\alpha > 0$, put $\nu_\alpha(M,\rho) = \inf \sum_i \nu(v_i)^\alpha$, where the infimum is taken over the ν-ρ-coverings of $M \subset [0,1]$. Then, the Hausdorff dimension of set M with respect to measure ν is defined as

$$\dim_\nu(M) = \inf\{\alpha : \lim_{\rho\to 0} \nu_\alpha(M,\rho) = 0\}$$

$$= \sup\{\alpha : \lim_{\rho\to 0} \nu_\alpha(M,\rho) = \infty\}.$$

Note that $0 \leq \dim_\nu(M) \leq 1$ for any ν and M. Note also that for the Lebesgue measure L, $\dim_L(M)$ coincides with the Hausdorff dimension in the usual sense [1], which is denoted by $\mathrm{HD}(M)$.

We define the mapping $\tau : B^\infty \to [0,1]$ associating with each $x_1^\infty \in B^\infty$ the real number $\tau(x_1^\infty)$, the binary expansion $0.x_1x_2\cdots$. For each $I \subset B^\infty$, we denote as $\mathrm{HD}(I)$ the Hausdorff dimension of set $\tau(I)$ [3]. Then, the following facts are shown in [6], [7]:

Proposition 1:
1. $M(\alpha) \subset K(\alpha)$;
2. $\mathrm{HD}(M(\alpha)) = \alpha$; and
3. $\mathrm{HD}(K(\alpha)) = \alpha$.

In this paper, we prove the following theorem:

Theorem 1: $\mathrm{HD}(K(\alpha)\backslash M(\alpha)) = \alpha$.

2. Proof of Theorem

First, we state two lemmas needed for the proof.

Lemma 1: ([6], [7])
1. $A_1 \subset A_2 \Rightarrow \dim_\nu(A_1) \leq \dim_\nu(A_2)$;
2. $\nu(A) > 0 \Rightarrow \dim_\nu(A) = 1$; and
3. $\dim_\nu(\bigcup_n A_n) = \sup_n\{\dim_\nu(A_n)\}$.

Lemma 2: ([1]) Let μ and ν be probability measures. Then,

$$M \subset \{x_1^\infty | \liminf_{n\to\infty} \frac{\log \nu(x_1^n)}{\log \mu(x_1^n)} \geq \delta\}$$

implies $\dim_\mu(M) \geq \delta \cdot \dim_\nu(M)$.

Manuscript received June 1, 2000.
Manuscript revised February 8, 2001.

[†]The author is with the Department of Mathematics, Osaka University, Toyonaka, Osaka 560-0043, Japan
[††]The author is with the Siberian State Academy of Telecommunication, Nobosibirsk 630102, Russia

*This work was done when Joe Suzuki visited Novosibirsk, Russia (June 1998), and Boris Ryabko visited Osaka, Japan (July 1998).

(*Proof of theorem 1*)

$\mathrm{HD}(K(\alpha)\backslash M(\alpha)) \le \alpha$ is clear because $\mathrm{HD}(K(\alpha)) = \alpha$ and $K(\alpha)\backslash M(\alpha) \subset K(\alpha)$. Thus we have only to show that $\mathrm{HD}(K(\alpha)\backslash M(\alpha)) \ge \alpha$ for $\alpha > 0$, or even stronger,

$$(*) \quad \mathrm{HD}(K(\alpha)\backslash M) \ge \alpha.$$

In fact, due to $K(\alpha)\backslash M(\alpha) \supset K(\alpha)\backslash M$ and the first item of lemma 1, $(*)$ implies $\mathrm{HD}(K(\alpha)\backslash M(\alpha)) \ge \alpha$. Moreover, we have only to show $(*)$ for rational α in the unit interval $(0,1]$ because for any $\alpha \in (0,1]$, it holds that

$$K(\alpha)\backslash M \supset \bigcup_{\substack{\beta\in \boldsymbol{Q}\\ \beta<\alpha}} (K(\beta)\backslash M),$$

which implies

$$\mathrm{HD}(K(\alpha)\backslash M) \ge \mathrm{HD}(\bigcup_{\substack{\beta\in \boldsymbol{Q}\\ \beta<\alpha}} (K(\beta)\backslash M))$$

$$= \sup_{\substack{\beta\in \boldsymbol{Q}\\ \beta<\alpha}} \{\mathrm{HD}(K(\beta)\backslash M)\}$$

by the first and third items of lemma 1.

(i) Case $0 < \alpha < 1$

We choose an infinite sequence $\{n_i\}_{i=1}^{\infty}$ of positive integers which satisfies the following[†]:

1. $\lim_{i\to\infty} \frac{n_1 + n_2 + \cdots + n_{i-1}}{n_1 + n_2 + \cdots + n_{i-1} + n_i} = 0$, and
2. αn_i and $(1-\alpha)n_i$ are both integers for all i.

We further denote

$$X := \{x_1^\infty \in B^\infty | \lim_{n\to\infty} \frac{\sum_{i=1}^n x_i}{n} = \frac{1}{2}\}.$$

By the law of large numbers, X has Lebesgue measure one. From each $x_1^\infty \in X$, we construct a new infinite binary sequence y_1^∞ as follows: Partition x_1^∞ into blocks of length $\alpha n_1, \alpha n_2, \ldots$, successively and insert $(1-\alpha)n_i$ zeros just after the i-th block. We denote the set of such sequences by Y, that is,

$$Y := \{y_1^\infty = \overbrace{x_1^{\alpha n_1} 0^{(1-\alpha)n_1}}^{n_1} \overbrace{x_{\alpha n_1+1}^{\alpha n_1+\alpha n_2} 0^{(1-\alpha)n_2}}^{n_2}$$

$$\overbrace{x_{\alpha n_1+\alpha n_2+1}^{\alpha n_1+\alpha n_2+\alpha n_3} 0^{(1-\alpha)n_3}}^{n_3} \cdots\cdots : x_1^\infty \in X\},$$

where 0^m denotes the sequence $\underbrace{00\cdots0}_{m}$. We list some basic facts about the set Y.

[†]For example, when $\alpha = \frac{r}{q}$, $(q, r$: integer), it suffices to define $n_1 = q$ and $n_{i+1} = i\sum_{j=1}^{i} n_j$.

Fact 1. $Y \cap M = \emptyset$, i.e., there is no sequence in Y that is typical for some stationary ergodic measure. This is because $\lim_{i\to\infty} \frac{\sum_{j=1}^{m_i} y_j}{m_i} = \frac{1}{2}$ for $m_1 = \alpha n_1$, $m_2 = n_1 + \alpha n_2$, $m_3 = n_1 + n_2 + \alpha n_3, \ldots$, and $\lim_{i\to\infty} \frac{\sum_{j=1}^{l_i} y_j}{l_i} = \frac{\alpha}{2}$ for $l_1 = n_1$, $l_2 = n_1 + n_2$, $l_3 = n_1 + n_2 + n_3, \ldots$

Fact 2. $Y \subset K(\alpha)$, i.e., for any $y_1^\infty \in Y$, $\liminf_{n\to\infty} \frac{KC(y_1^n)}{n} \le \alpha$. This is because we can reconstruct $y_1^{l_i}$ from $x_1^{\alpha l_i}$ in an algorithmic manner, where $l_i = n_1 + \cdots + n_i$, so that $KC(y_1^{l_i}) \le KC(x_1^{\alpha l_i}) + O(1) \le \alpha l_i + O(1)$ holds for $i = 1, 2, \ldots$

Fact 3. For $y_1^\infty \in Y$, let $s(n)$ be the number of bits in y_1^n that stem from x_1^∞, i.e., for $1 \le n \le \alpha n_1$,

$$s(n) = n;$$

for $\alpha n_1 + 1 \le n \le n_1$,

$$s(n) = \alpha n_1;$$

for $n_1 + \cdots + n_k + 1 \le n \le n_1 + \cdots + n_k + \alpha n_{k+1}$,

$$s(n) = n - (1-\alpha)(n_1 + \cdots + n_k);$$

for $n_1 + \cdots + n_k + \alpha n_{k+1} + 1 \le n \le n_1 + \cdots + n_k + n_{k+1}$,

$$s(n) = \alpha(n_1 + \cdots + n_k + n_{k+1}).$$

Then, $\liminf_{n\to\infty} \frac{s(n)}{n} = \alpha$.

By Facts 1 and 2, we get $Y \subset K(\alpha)\backslash M$. Therefore it is enough to show $\mathrm{HD}(Y) \ge \alpha$. In what follows, we construct a measure ν such that

(ν-1) $\nu(Y) = 1$; and

(ν-2) $Y \subset \{y_1^\infty | \liminf_{n\to\infty} \frac{\log \nu(y_1^n)}{\log L(y_1^n)} \ge \alpha\}$,

where L is the Lebesgue measure. If we find such a ν, we can derive $\mathrm{HD}(Y) = \dim_L(Y) \ge \alpha \cdot \dim_\nu(Y) = \alpha$, where the second inequality follows from (ν-2) and lemma 2, and the last equality follows from (ν-1) and the second item of lemma 1.

<Construction of ν>

We define ν as follows: for $1 \le n \le \alpha n_1$,

$$\nu(y_1^n) = \frac{1}{2^n}, y_1^n \in B^n;$$

for $\alpha n_1 + 1 \le n \le n_1$, $y_1^{\alpha n_1} \in B^{\alpha n_1}$,

$$\nu(y_{\alpha n_1+1}^n | y_1^{\alpha n_1}) = \begin{cases} 1, & y_{\alpha n_1+1}^n = 0^{n-\alpha n_1} \\ 0, & \text{otherwise} \end{cases} ;$$

for $n_1 + 1 \le n \le n_1 + \alpha n_2$, $y_1^{n_1} = x_1^{\alpha n_1} 0^{(1-\alpha)n_1}$, $x_i \in B$

$$\nu(y_{n_1+1}^n | y_1^{n_1}) = \frac{1}{2^{n-n_1}},\ y_{n_1+1}^n \in B^{n-n_1};$$

for $n_1 + \alpha n_2 + 1 \leq n \leq n_1 + n_2$, $y_1^{n_1+\alpha n_2} = x_1^{\alpha n_1} 0^{(1-\alpha)n_1} x_{\alpha n_1+1}^{\alpha n_1+\alpha n_2}$, $x_i \in B$,

$$\nu(y_{n_1+\alpha n_2+1}^n | y_1^{n_1+\alpha n_2})$$
$$= \begin{cases} 1, & y_{n_1+\alpha n_2+1}^n = 0^{n-n_1-\alpha n_2} \\ 0, & \text{otherwise} \end{cases};$$

for $n_1 + n_2 + 1 \leq n \leq n_1 + n_2 + \alpha n_3$, $y_1^{n_1+n_2} = x_1^{\alpha n_1} 0^{(1-\alpha)n_1} x_{\alpha n_1+1}^{\alpha n_1+\alpha n_2} 0^{(1-\alpha)n_2}$, $x_i \in B$,

$$\nu(y_{n_1+n_2+1}^n | y_1^{n_1+n_2}) = \frac{1}{2^{n-n_1-n_2}},$$

$$y_{n_1+n_2+1}^n \in B^{n-n_1-n_2}$$

and so on.

Then, (ν-1) holds because $L(X) = 1$. (ν-2) follows since for any $y_1^\infty \in Y$

$$\liminf_{n\to\infty} \frac{\log \nu(y_1^n)}{\log L(y_1^n)} = \liminf_{n\to\infty} \frac{\log(1\backslash 2)^{s(n)}}{\log(1\backslash 2)^n}$$

$$= \liminf_{n\to\infty} \frac{s(n)}{n} = \alpha,$$

where the last equality follows from Fact 3.

(ii) Case $\alpha = 1$

We choose $\{\alpha_n\}_{n=1}^\infty$ such that $0 < \alpha_1 < \alpha_2 < \cdots < \alpha_n < \cdots < \alpha = 1$ and $\alpha_i \in \boldsymbol{Q}$. Then,

$$K(\alpha_1)\backslash M \subset K(\alpha_2)\backslash M \subset \cdots \subset K(\alpha_n)\backslash M \subset \cdots$$
$$\subset K(1)\backslash M.$$

So,

$$K(1)\backslash M \supset \bigcup_{i=n}^{\infty} K(\alpha_n)\backslash M. \tag{1}$$

Then,

$$\mathrm{HD}(K(1)\backslash M) \geq \mathrm{HD}(\bigcup_{n=1}^{\infty} K(\alpha_n)\backslash M)$$

$$= \sup_n \{\mathrm{HD}(K(\alpha_n)\backslash M)\} \geq 1,$$

where the first inequality follows from (1) and the first item of lemma 1, the second equality follows from the third item of lemma 1, the last inequality follows from the fact $\mathrm{HD}(K(\alpha_n)\backslash M) \geq \alpha_n$ in the case $\alpha_n < 1$.

This completes the proof of ($*$).

3. Concluding Remarks

Let T be a Turing machine (TM) having an infinite working tape, an infinite input tape, and an infinite output tape. We assume that the writing head can move to the right only. We denote as $T(x_1^\infty) = y_1^\infty$ if the TM T produces $y_1^\infty \in B^\infty$ given x_1^∞ as an input. For $I \subset B^\infty$, a one to one map $\varphi : I \to B^\infty$ is called *T-code for I* if there exist a pair of TMs T_φ^c (encoder) and T_φ^d (decoder) such that for all $x_1^\infty \in I$, $\varphi(x_1^\infty) = T_\varphi^c(x_1^\infty)$, and that for all $x_1^\infty \in I$, $T_\varphi^d(T_\varphi^c(x_1^\infty)) = x_1^\infty$.

Let φ be a T-code for I realized by T_φ^c and T_φ^d, and let $x_1^\infty \in I$ and $y_1^\infty := T_\varphi^c(x_1^\infty)$. Further let $s(i)$ =(the first step at which T_φ^d reads y_i), and $t(j)$ =(the first step at which T_φ^d writes x_j). Then, there exist natural numbers $i_1, i_2, \ldots$, and $j_1, j_2, \ldots$, such that $s(1) < s(2) < \cdots < s(i_1) < t(1) < t(2) < \cdots < t(j_1) < s(i_1+1) < s(i_1+2) < \cdots < s(i_2) < t(j_1+1) < t(j_1+2) < \cdots < t(j_2) < \cdots$. We define the *T-cost of φ for a single word x_1^∞* by

$$c(\varphi, x_1^\infty) = \limsup_{k\to\infty} \frac{i_k}{j_k},$$

and define the *T-cost for $I \subset B^\infty$* by

$$c(I) = \inf_\varphi \sup_{x_1^\infty \in I} c(\varphi, x_1^\infty),$$

where the infimum is taken over all T-codes φ. The notion of T-cost is a measure of compression rate based on the decoding procedure, and is related to the quantity appeared in the definition of $K(\alpha)$ as follows.

Proposition 2: For any subset $I \subset B^\infty$,

$$c(I) = \sup_{x_1^\infty \in I} \liminf_{n\to\infty} \frac{KC(x_1^n)}{n}$$

For the proof, see [7].

By Proposition 2, we get $K(\alpha) = \{x_1^\infty \in B^\infty : c(\{x_1^\infty\}) \leq \alpha\}$. This means that $K(\alpha)$ is the set of sequences each of which can be compressed, in the sense of T-cost, at a rate less than or equal to α.

We briefly mention the meaning of our main Theorem 1 in the context of data compression. Currently, the most popular model in data compression theory is that of stationary ergodic sources. The set $K(\alpha)\backslash M(\alpha)$ can be regarded as sequences each of which is not emitted from any stationary ergodic source but can be compressed by a certain algorithm. While the existence of such sequences is obvious, it is not obvious how large large $K(\alpha)\backslash M(\alpha)$ is or how to estimate the size of it. Thoerem 1 says that the size of $K(\alpha)\backslash M(\alpha)$ is as large as $K(\alpha)$ and $M(\alpha)$ in terms of Hausdorff dimension, and suggests for investigating investigate a new approach for data compression method.

References

[1] P. Billingsley, "Ergodic Theory and Information," John Wiley & Sons, 1965.

[2] T.M. Cover and S.K. Leung, "Some equivalences between Shannon entropy and Kolmogorov complexity," IEEE Trans. Inf. Theory, vol.24, pp.331–338, 1978.

[3] C.M. Colebrook, "The Hausdorff dimension of certain sets of nonnormal numbers," Michigan Math. J., vol.17, pp.103–116, 1970.

[4] A.N. Kolmogorov, "Three approaches to the definition of quantity of information," Probl. Peredachi Inf., vol. 1, pp.3–7, 1965.
[5] A. Lempel and J. Ziv, "A universal algorithm for sequential data compression," IEEE Trans. Inf. Theory, vol.23, pp.337–343, 1977.
[6] B.Ya. Ryabko,"Coding of combinatorial sources and Hausdorff dimension," Soviet Math. Doklady., vol.30, pp.219–222, 1984.
[7] B.Ya. Ryabko,"Noiseless coding of combinatorial sources, Hausdorff dimension, and Kolmogorov complexity," Problems of Information Transmission, vol.22, pp.170–179, 1986.
[8] A.K. Zvonkin and L.A. Levin, "The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms," Russian Math. Surveys, vol.25, pp.83–124, 1970.

Kouki Hojo was born in Osaka, Japan, 1974. He received the B.E. and M.E. in mathematical course from Osaka University, Japan, in 1997 and 1999, respectively. He is in the second year of doctor course in Department Mathematics Graduate School of Science, Osaka University.

Boris Ya. Ryabko was bom in Novosibirsk, Russia, on January 14, 1949. He graduated from the Department of Mathematics, Novosibirsk State University in 1971, received the Ph.D. degree from the Institute of Mathematics,Novosibirsk, in 1981, and the Dr.Sci. degree from the Institute for Problems of Information Transmission, Moscow,in 1981. He is currently a Professor and Head of the Department of Applied Mathematics and Cybernetics at the Siberian State University of Telecommunications and Information Sciences, Novosibirsk, and Head of Laboratory of data protection, Institute of Computational Technology of Siberian Branch of Russian Academy of Science. His research interests include data compression, source coding, prediction, cryptography, mathematical biology, and other applied (discrete) mathematics. Dr. Ryabko was awarded prizes by the Russian Acadely of Sciences for the best paper in 1996 and 1997. Since 1997, he has held the Russian Federation President stipend for outstanding scientific achievements.

Joe Suzuki was born in Tokyo, Japan, 1960. He received the B.E., M.E, Dr. of Engineering degrees from Waseda University, Tokyo, Japan, in 1984, 1986, 1993, respectively. In 1994, he joined a faculty member of the Department of Mathematics, Osaka University, Osaka, Japan. He was Visiting Assistant Professor of Stanford University (1995-1997) and Visiting Scientist of Brown University (1998). Currently, he is Associate Professor of Osaka University. His research interests are mainly on mathematical analyses of computer science such as universal data compression, Kolmogorov complexity, statistical model selection, Bayesian belief networks, genetic algorithms, elliptic curve cryptosystems, algebraic geometry codes.