Hausdorff Dimension as a New Dimension in Source Coding and Predicting

Boris Ryabko  
Siberian University  
ofNovosibirsk, Russia  
(ryabko@niec.nsk.su)  

Joe Suzuki  
Osaka University  
Osaka, Japan.  
(suzuki@math.sci.osaka-u.ac.jp)  

Flemming Topsoe  
University of Copenhagen  
Denmark  
(topsoe@math.ku.dk)

Abstract
— It is generally accepted that investigations in universal coding and predicting are based on the model of stationary ergodic sources. In this report we show that this model does not give possibilities to investigate large important classes of source codes and to distinguish asymptotic performances of popular universal codes. A new approach suggested here is to consider a set of all infinite sequences (over a given alphabet) and estimate the size of sets of compressible sequences with the help of Hausdorff dimension. This approach enables us, first, to show that there exist large sets of well compressible (and predictable) sequences which have got zero measure for every stationary and ergodic measure, and second, to distinguish an asymptotic efficiency of LZ codes and codes that are based on the technique of model weighting.

I. INTRODUCTION
Nowadays source coding has got many important applications to telecommunications and computer engineering that are based on Shannon theory, Lempel-Ziv codes and many other elegant constructions and theories. Besides, it is well known that the problem of source coding is close to the prediction problem if the precision is estimated by the Kullback-Leibler divergence (see [1-2]), and that makes both theories more powerful and attractive.

The majority of investigations in source coding and source predicting are based on the model in which messages are generated by an ergodic and stationary source (ESS model). This approach has led to many important results, however it seems there are several problems that cannot be solved in a framework of the model of stationary and ergodic source.

The first problem may be formulated as follows: the ESS model does not distinguish the main classes of universal codes. More exactly, it is known that LZ codes [3], two universal codes [4], CTW code [5] and other ones that are based on the model weighting technique have got the entropy rate of stationary ergodic source as a limit code length. So all the codes are asymptotically equal if we use the ESS model. This problem is known in source coding and researchers look for other models which can distinguish different codes. Thus in [6] an individual sequence of letters has been constructed in order to compare the performances of the finite state encoder and the context-tree weighting method. It has been shown in [6] that there exist sequences that are incompressible by a finite-state encoder but compressible by the extended context-tree weighting method.

The next questions seem natural, namely, how many sequences have got this property and how to estimate the size of sets of such sequences. It is important to note that it is impossible to use an ergodic and stationary measure because for such a measure the set of sequences with different limit performance for different codes has got the measure zero.

In order to consider the second problem we give some definitions. Let a source generate letters from a finite alphabet $A$. Define the set of all infinite words $x_1 x_2 \ldots, x_i \in A$, as $A^\infty$. It is known that there exists a set $B$ of sequences from $A^\infty$ such that every $x \in B$ can be compressed and predicted quite well but for every stationary and ergodic measure $\mu$ on $A^\infty \mu(B) = 0$. Informally, it means that we cannot see the set $B$ if we use the ESS model. It is natural to consider the same question: how large $B$ is and how to estimate the size of the set $B$. We will show that the size of the set $B$ is quite big. So if somebody uses the ESS model he cannot see the large set of sequences that can be compressed and predicted quite well.

The purpose of the present contribution is to show a possible alternative approach which may be formulated as follows: consider sets of sequences and use Hausdorff dimension in order to measure the size of the sets. It is important to note that using of Hausdorff dimension for this purpose is quite natural because this measure is closely related with information theory [7,8]. This approach enables us to solve the above mentioned problems. Namely, we shall show that there exist large sets of sequences which can be compressed and predicted quite well but the measure of the sets is zero for every ergodic and stationary measure. (So the set is invisible if we use the ESS model). We also show that there exist a large number of sequences that are well compressible by Lempel-Ziv code but are not compressible by the model weighting technique.

II. THE INVISIBLE SET
We give some definitions. Let $A$ be $\{0, 1\}$ -alphabet, and $A^n, A^*$ and $A^\infty$ the sets of all words of the length $n, n \geq 1$, all finite words and one-side-infinite words, respectively, in the alphabet $A$. (The result in this report can be extended to the case $|A| < \infty$ without any difficulty). A code $\varphi_n, n \geq 1$, is given by a mapping $\varphi_n : A^n \to A^*$ such that the set $\{\varphi_n(x), x \in A^n\}$ satisfies the prefix condition. The set of mappings $\{\varphi_n, n = 1, 2, \ldots\}$ we also call the code $\varphi$. The cost of the code $\varphi$ on a word $x \in A^n$ is defined to be

$$c(\varphi, x) = \frac{1}{n} |\varphi_n(x)|$$

where $|u|$ is the length of a word $u$.

Let $K(z)$ be the Kolmogorov complexity of the word $x, x \in A^*$. Let us define $K(\alpha) = \{z : x \in A^\infty \lim_{n \to \infty} \inf K(z)/n \leq \alpha\}$ for every $\alpha \in (0,1)$. Informally, $K(\alpha)$ is the set of infinite binary sequences that can be compressed by any algorithm at the rate not more than $\alpha$. It is shown in [7] that

$$DH(K(\alpha)) = \alpha$$

(1)
where DH denotes Hausdorff dimension. The theorem below shows that there exists a large set of sequences that can be compressed quite well but the set is invisible in the framework of the ESS model.

**Theorem 1.** For every \( \alpha \in (0, 1) \), there exists the set \( U_\alpha \) such that
i) there exists the code \( \varphi \) for which
\[
\lim_{n \to \infty} \frac{1}{n} |\varphi(x^n)| \leq \alpha
\]
for every \( x \in U_\alpha \).

ii) \( DH(U_\alpha) = \alpha \).

iii) for every stationary and ergodic measure \( \mu \)
\[
\mu(U_\alpha) = 0
\]

Comment. From i) and ii) we can see that \( U_\alpha \) contains the sequences that can be compressed (and predicted) quite well and this set is quite large. On the other hand the measure of the set \( U_\alpha \) zero for every stationary and ergodic source.

We shall not give the complete proof but construct the set \( U_\alpha \) and explain why \( U_\alpha \) is invisible. Let us give some definitions. Let \( \mu \) be a probability measure on \( A^\infty \). A sequence \( x \in A^\infty \) is defined to be \( \mu \) typical if for any \( u \in A^* \)
\[
\lim_{n \to \infty} \frac{\nu_u(x^n)}{n - |u| + 1} = \mu(u)
\]
where \( \nu_u(x^n) \) is the number of occurrences of \( u \) in \( x^n \).

For every \( \alpha \in (0, 1) \) there exists \( \pi \) for which
\[
- (\pi \log \pi + (1 - \pi) \log (1 - \pi)) = \alpha
\]

We consider two Bernoulli sources \( \mu_1 \) and \( \mu_2 \) such that
\[
\mu_1(0) = \pi, \quad \mu_1(1) = 1 - \pi
\]
\[
\mu_2(0) = 1 - \pi, \quad \mu_2(1) = \pi
\]
and let \( B_1 \) and \( B_2 \) be the sets of all \( \mu_1 \) typical sequences and \( \mu_2 \) ones respectively. In order to define the set \( U_\alpha \) we take any sequences \( v = v_1 v_2 \ldots \in B_1 \) and \( w = w_1 w_2 \ldots \in B_2 \) and define the sequence \( u \in U_\alpha \) as follows
\[
u = v_1 w_1 v_2 w_2 v_3 w_3 \ldots v_{1,2} w_{1,2} \ldots
\]

It is known that Hausdorff dimension of the set of all typical sequences of ergodic stationary source is equal to the Shannon entropy of the source [8]. So \( DH(B_1) = DH(B_2) = - (\pi \log \pi + (1 - \pi) \log (1 - \pi)) = \alpha \), see (4). Intuitively, \( U_\alpha \) may be transformed into \( B_1 \) or \( B_2 \) without compression and expansion. That is why all three sets have got the same Hausdorff dimension.

Let us prove the last property. It is obvious that the sequences from \( U(\alpha) \) are not typical for every ESS measure because
\[
\lim_{n \to \infty} \frac{\nu_0(x^n)}{n}, \quad \lim_{n \to \infty} \frac{\nu_1(x^n)}{n}
\]
don’t exist. That is why \( \mu(U_\alpha) = 0 \) for every ergodic and stationary measure \( \mu \).

### III. The Comparison of Universal Codes

We shall compare LZ codes and CTW ones. Let us define
\[
LZ(\alpha) = \{ x \in A^\infty : \lim inf_{n \to \infty} \frac{1}{n} |LZ(x^n)| \leq \alpha \}
\]
\[
CTW(\alpha) = \{ x \in A^\infty : \lim inf_{n \to \infty} \frac{1}{n} |CTW(x^n)| \leq \alpha \}
\]
for \( \alpha \in (0, 1) \). It is easy to prove that
\[
DH(LZ(\alpha)) = DH(CTW(\alpha)) = \alpha
\]

**Theorem 2.**
\[
DH(LZ(\alpha)) = DH(CTW(\alpha)) = \alpha
\]

As it may be seen from (6) and (7) \( LZ(\alpha) \) is much larger than \( CTW(\alpha) \). Moreover, the set \( LZ(\alpha) \cap CTW(\alpha) \) is as large as \( LZ(\alpha) \). Let us define Markov sources \( \mathcal{M}_1^\alpha, \mathcal{M}_2^\alpha, \mathcal{M}_1^\beta, \mathcal{M}_2^\beta, \mathcal{M}_3^\beta, \mathcal{M}_4^\beta, \ldots \) as follows:
\[
P_0^\alpha(0) = \pi, \quad P_0^\alpha(0) = 1 - \pi
\]
\[
P_1^\alpha(0) = 1 - \pi, \quad P_1^\alpha(0) = \pi
\]
\[
P_0^\alpha(00) = \pi, \quad P_1^\alpha(00) = 1 - \pi
\]
\[
P_1^\alpha(11) = \pi, \quad P_1^\alpha(11) = 1 - \pi
\]
\[
P_0^\beta(00) = \pi, \quad P_0^\beta(00) = 1 - \pi
\]
\[
P_1^\beta(11) = \pi, \quad P_1^\beta(11) = 1 - \pi
\]
\[
P_0^\beta(01) = \pi, \quad P_0^\beta(01) = 1 - \pi
\]
\[
P_1^\beta(10) = \pi, \quad P_1^\beta(10) = 1 - \pi
\]
and so on. (Here \( \mathcal{M}_1^\alpha, \mathcal{M}_0^\beta \) are Bernoulli sources, \( \mathcal{M}_1^\beta, \mathcal{M}_2^\beta \) are the first order Markov sources, \( \mathcal{M}_2^\alpha, \mathcal{M}_3^\beta \) are the second order ones and so on). Let \( x_\alpha^\beta \) be the set of all typical sequences that are generated by the source \( \mathcal{M}_\alpha^\beta \), \( \alpha = 0, 1, \ldots, \beta = 0, 1 \).

For the sake of simplicity we give only informal definition of the set \( D_\alpha \) and explain the main idea of construction only. Let a sequence \( x_\alpha^\beta \) belongs to \( x_\alpha^\beta \). We define a sequence \( Z \in D_\alpha \) as follows:
\[
Z = x_0^\alpha x_1^\beta x_2^\alpha x_3^\beta x_4^\alpha x_5^\beta \ldots
\]

As it may be seen from (8) the CTW encoder has got a wrong sample when it encodes every letter of the sequence \( Z \). That is why the CTW code can not compress the sequence up to \( \alpha \). Note that the entropy of the source \( \mathcal{M}_\beta^\beta \) is equal to \(- (\pi \log \pi + (1 - \pi) \log (1 - \pi)) = \alpha \) for \( i = 0, 1, \ldots, j = 0, 1 \).

On the other hand the LZ encoder uses as a sample that part of an encoded word which gives the best compression.

The proof of (8) is more complicated. In short, the length of the code word of the CTW code is closed to so called empirical entropy. On the other hand the empirical entropy is the upper bound for the code length of the LZ code.
REFERENCES


