Comparing the Multilevel Pattern Matching Code and the Lempel-Ziv Codes

Boris Ya. Ryabko
Siberian State University of Telecommunication and Computer Science, Novosibirsk, Russia
e-mail: ryabko@neic.nsk.su

Joe Suzuki
Department of Mathematics,Osaka University,Osaka, Japan
e-mail: suzuki@mathsun01.math.sci.osaka-u.ac.jp

Abstract — **Asymptotic performances of the multilevel pattern matching (MPM) code and a Lempel-Ziv'77 (LZ) code are compared. It is known that both codes have got Shannon entropy as asymptotic performance if they are employed for stationary ergodic sources, but the redundancy of MPM is less than that of LZ. In this paper, it is shown that there exists a large set of sequences that can be compressed well by LZ but cannot be compressed by MPM.**

I. Introduction

Recently, Kieffer and Yang [1] proposed a new class of lossless universal codes that includes all known universal codes such as the class of Lempel-Ziv codes and the class of universal codes based on model weighting (MW) technique (equivalently, the twice universal coding method [4] and the context tree weighting method [7]). The new class (grammar-based codes [1]) includes many new promising coding methods as well. One of them is the multilevel pattern matching (MPM) coding method which has been proposed by Kieffer et al [2].

It turns out that the asymptotical compression ratio of those codes achieves the Shannon entropy as an asymptotical limit when they are applied to stationary ergodic sources. So, it is not enough to consider the model of stationary ergodic sources because it is extreamly difficult to consider that all the real data sequences have been generated by them although all universal codes are supposed to be equally good in that they can compress data sequences emitted from a stationary ergodic source as low as its entropy.

In order to get over this difficulty, several non-probabilistic approaches have been considered thus far. Thus, Shields [6] suggested to compare the performance of universal codes using individual sequences. In this paper, we develop this approach in such a way that we evaluate the cardinality of the set of sequences such that a code outperforms another code.

The main result of this paper may be formulated as follows: there exists a large set of infinite sequences which can be compressed well by LZ77 with unbounded window and buffer sizes [5] and at the same time cannot be compressed by MPM. This does not immediately mean that LZ77 is better than MPM because, in the first place, to the contrary, there might exist a large set of such sequences which are better compressed by MPM than by LZ77 (this question seems to be an open problem). Secondly, there are other many important issues of codes such as time and space complexities, and asymptotical compression ratio does not seem to tell the whole story.

II. Main Result

Theorem 1. *Let $LZ(z)$ and $MPM(z)$ be the codelength of z realized by LZ77 and the MPM coding method. Then, for all $\gamma \in (0,1)$, there exist $n_1 < n_2 < \cdots$ and $W^{(i)}$ $i = 1, 2, \cdots$, such that for $\forall w \in W^{(i)}$*

$$\frac{MPM(w)}{n_i} = 1+o(1), \frac{LZ(w)}{n_i} = \gamma+o(1), \frac{\log |W^{(i)}|}{n_i} = \gamma+o(1)$$

as $n \to \infty$.

Remark Theorem 1 shows that there exist infinitely many sets $W^{(i)}$ such that each element in $W^{(i)}$ cannot be compressed by the MPM code but can be compressed well by the LZ77 code Moreover, the sets $W^{(i)}$ have got the maximal size. (Indeed, it is obvious that for every code over $\{0,1\}^{n_i}$, the set of codewords cannot include more than $2^{\gamma n_i}$ words of length no more than γn_i.) We give the main idea of the proof. For simplicity, we consider the case $\gamma = 1/2$. For each set $W^{(i)}$, we choose elements of the form $w = vu \in B^{n_i}$ such that $u \in B^{n_i/2}$ is a cyclic shift of $v \in B^{n_i/2}$, and that vu cannot be compressed well by MPM. On the other hand, vu can be compressed almost as low as a half by the LZ77 code because all the bits in u can be obtained by twicely implementing the LZ77 parsing procedure. Hence $LZ(vu) \leq |v| + O(\log |v|)$ whereas $MPM(vu) = |vu|(1 + o(1)) = 2|v|(1 + o(1))$.

Now we compare the performances of MPM and LZ77 using the notion of combinatorial sources. In this case, there is a natural way of estimating the sizes of sets that is based on Hausdorff dimension.

Theorem 2. *For $\forall\gamma \in (0,1)$, there exists a set of infinite sequences, X_γ such that for $\forall x \in X_\gamma$ $\liminf_{t\to\infty} \frac{LZ(x_1^t)}{t} = \gamma$, $\liminf_{t\to\infty} \frac{MPM(x_1^t)}{t} = 1$, $HD(X_\gamma) > \gamma$.*

References

[1] J. Kieffer and E. Yang, "Grammer Based Codes: A new class of universal lossless source codes", *IEEE Trans. on Inform. Theory*, vol. IT-46, no. 3, 2000.

[2] J. Kieffer, E. Yang, G. Nelson, and P. Cosman, "Universal lossless compression via multilevel pattern matching", *IEEE Trans. on Inform. Theory*), 2000.

[3] J. Ziv and A. Lempel, "A Universal algorithm for data compression", *IEEE Trans. Inform. Theory*, vol. IT-23, pp. 337-343, 1977.

[4] B. Y. Ryabko, "Twice-universal coding", *Probl. Peredachi Inform.*, vol. 20, No. 3, pp. 173-177 (1984).

[5] S. Savari, "Redundancy of the Lempel-Ziv String Matching Code",*IEEE Trans. on Inform. Theory*, vol. 44, pp. 787-791 (1998).

[6] P. Shields, "Performance of LZ algorithms on individual sequencs", *IEEE Trans. on Inform. Theory* ,vol. 45, pp. 1283-1288 (1999).

[7] F. M. J. Willems, Y. M. Shtarkov, and T. J. Yjalkens, "The context tree weighting method: Basic properties", *IEEE Trans. on Inform. Theory*, vol. IT-41, no. 3, pp. 653-664 (1995).