



# On the distribution function of the complexity of finite sequences

Janusz Szczepanski<sup>1</sup>

Polish Academy of Sciences, Institute of Fundamental Technological Research, Swietokrzyska 21, 00-049 Warsaw, Poland

## ARTICLE INFO

### Article history:

Received 20 July 2007

Received in revised form 15 November 2007

Accepted 27 December 2008

### Keywords:

Lempel–Ziv complexity

Distribution function

Randomness

## ABSTRACT

Investigations of complexity of sequences lead to important applications such as effective data compression, testing of randomness, discriminating between information sources and many others. In this paper we establish formulae describing the distribution functions of random variables representing the complexity of finite sequences introduced by Lempel and Ziv in 1976. It is known that this quantity can be used as an estimator of entropy. We show that the distribution functions depend affinely on the probabilities of the so-called “exact” sequences.

© 2009 Elsevier Inc. All rights reserved.

## 1. Introduction

The notion of complexity of a given sequence was first introduced in papers by Kolmogorov [13] and Chaitin [7]. Kolmogorov proposed, as a measure for the complexity of that sequence with respect to the given algorithm, the use of the length of the shortest binary program which, when fed into a given algorithm, will cause it to produce a specified sequence. If the length of the program is large we can say that the complexity of the sequence is large.

In 1976 Lempel and Ziv [14] proposed and explored another approach to the problem of the complexity of a specific sequence. They linked the complexity of a specific sequence to the gradual buildup of new patterns along the given sequence. The complexity measure suggested by them is related to the number of distinct phrases and the rate of their occurrence along the sequence. It reflects the behavior of a simple parsing algorithm whose task is to recognize newly encountered phrases during its scanning of a given sequence. In a series of papers, modifications of the Lempel–Ziv parsing algorithm were proposed in response to the needs of various applications. In general, in these algorithms a new phrase is established as the shortest substring which has not occurred previously, where the search for previous occurrences may be restricted or generalized in the modified algorithms in various ways, e.g.: by considering only a fixed number of preceding symbols [20], by considering only complete previously established phrases (Lempel–Ziv Incremental Parsing Algorithm [21]), by allowing a number (not more than a fixed threshold) of previous occurrences of the phrase (Generalized Lempel–Ziv Algorithm [17]), etc.

It turned out that investigations of sequence complexity play an important role in universal data compression schemes and their numerous applications such as efficient transmission of data [1,4–6,9,12,16,18,20,21], test randomness [22], discriminating between information sources [10,22], estimating the statistical model of individual sequences [22] and many others [8,11]. The most recent results [2] show that Lempel–Ziv complexity as defined in 1976 can have a practical meaning as an estimator of entropy. Recently, analytical formula has been derived for its variance estimate [3].

<sup>1</sup> Also at: The Kazimierz Wielki University, ul. Chodkiewicza 30, 85-064 Bydgoszcz, Poland.  
E-mail address: [jszczepa@ippt.gov.pl](mailto:jszczepa@ippt.gov.pl)

In this paper we introduce the concept of “exact” sequences i.e. sequences in which the last phrase of the sequence does not occur in earlier segments of the sequence (precise formulation: Definition 3). We derive formulas describing the distribution function of random variables representing the complexity of finite sequences as defined by Lempel and Ziv in 1976. These formulas turn out to be affine form with respect to the probabilities of exact sequences.

### 2. Lempel–Ziv complexity

In this section we introduce the notation and recall basic definitions [14].

Let  $A$  be a finite alphabet and let  $\alpha = |A|$  denote the size of the alphabet. Let  $A^n$  be the set of all sequences of length  $n$  over  $A$  and let  $S = s_1s_2 \cdots s_n$  be an arbitrary element of  $A^n$ . By  $S(i, j)$  we denote the substrings  $s_i s_{i+1} \cdots s_j$  of  $S$  when  $i \leq j$  and  $S(i, j) = \Lambda$  when  $j < i$ . Symbol  $\Lambda$  denotes the null-sequence, i.e. the “sequence” of length zero.

The partition

$$H(S) = S(1, h_1)S(h_1 + 1, h_2) \cdots S(h_{m-1} + 1, n) \tag{1}$$

of  $S$  such that for every  $i$ ,  $S(h_{i-1} + 1, h_i - 1)$  is a substring of  $S(1, h_i - 2)$  is called a history of  $S$  and the  $m$  strings  $H_i(S) = S(h_{i-1} + 1, h_i)$ ,  $i = 1, 2, \dots, m$  where  $h_0 = 0$  and  $h_m = n$ , are called the components of the history (note that  $h_1 = 1$ ). Let  $c(H(S))$  denote the number of components in a history  $H(S)$  of  $S$ .

Let  $\mathcal{H}(S)$  denotes the set of all histories for the sequence  $S$ .

**Definition 1.** The complexity  $c(S)$  of the sequence  $S$  is the number

$$c(S) = \min_{H \in \mathcal{H}(S)} \{c(H(S))\}. \tag{2}$$

**Definition 2.** The component  $H_i(S) = S(h_{i-1} + 1, h_i)$  is called *exhaustive* if this string does not appear in the string  $S(1, h_i - 1)$ . A history of  $S$  is called *exhaustive* if each of its components, except possibly the last one, is exhaustive.

It is easy to see [14] that every sequence has a unique exhaustive history, denoted by  $H_E(S)$ . For instance, the exhaustive history of the sequence  $S = 001101110110110$  is given by the following parsing of  $S$ : 0, 01, 10, 111, 0110110 where successive components are separated by commas.

**Remark 1.** It was proven in [14] that  $c(S) = c(H_E(S))$ , where  $c(H_E(S))$  is the number of components in  $H_E(S)$ . Thus, below we shall use  $c(H_E(S))$  as the definition of complexity.

**Definition 3.** The sequence  $S = s_1s_2 \cdots s_n$  is called *exact* if the last string  $S(h_{m-1} + 1, n)$  in its exhaustive history  $H_E(S) = S(1, h_1)S(h_1 + 1, h_2) \cdots S(h_{m-1} + 1, n)$  does not occur as a substring  $S(i, j)$  (where  $1 \leq i \leq j \leq n - 1$ ) in the sequence  $S(1, n - 1) = s_1s_2 \cdots s_{n-1}$ .

From now on we shall assume that for a fixed  $n$  any element of  $A^n$  is equiprobable, i.e. assign the same probability  $\alpha^{-n}$  to each element of  $A^n$  and

$$P_n : 2^{A^n} \rightarrow [0, 1] \tag{3}$$

denotes the probability in this sense. By  $P_n(k)$  we denote the probability of the event consisting of all sequences of length  $n$  and complexity  $k$  while  $P_n^{(e)}(k)$  is the probability of the event consisting of all exact sequences of length  $n$  and complexity  $k$ .

Under the above assumptions for every  $n \in N$  we define the random variable  $C_n : A^n \rightarrow N$  representing the complexity:

$$C_n(S) := c(H_E(S)) \tag{4}$$

for every sequence  $S \in A^n$ .

### 3. The distribution function of $C_n$

In this section we describe the distribution function of  $C_n$ ,  $n \in N$ . We prove the following

**Theorem.** Using the above notation,

$$P_{n+1}(C_{n+1} \leq k) = 1 - \sum_{r=1}^n P_r^{(e)}(k) \tag{5}$$

for every  $n, k \in N$ .

**Proof.** We first express  $P_{n+1}(k + 1)$  in terms of  $P_n$ . Taking into account that the number of all sequences of length  $n$  is  $\alpha^n$ , by definitions of  $P_n$  and  $P_n^{(e)}$  we find that:

- the number of sequences with complexity  $k + 1$  and length  $n$  is  $\alpha^n P_n(k + 1)$ ,
- the number of exact sequences with complexity  $k + 1$  and length  $n$  is  $\alpha^n P_n^{(e)}(k + 1)$ ,
- the number of exact sequences with complexity  $k$  and length  $n$  is  $\alpha^n P_n^{(e)}(k)$ .

Taking into account the definitions of complexity and exact sequences we conclude that every sequence with complexity  $k + 1$  and length  $n + 1$  can be obtained from a sequence of length  $n$  in one of the following two ways only:

- by adding a symbol to a sequence with complexity  $k + 1$  which is not exact,
- by adding a symbol to an exact sequence with complexity  $k$ .

We also see that all sequences obtained from exact sequences of length  $n$  and complexity  $k + 1$  by adding a symbol from  $A$  will increase their complexity to  $k + 2$  and the number of such sequences is  $\alpha \cdot \alpha^n \cdot P_n^{(e)}(k + 1)$ . From the definition of  $P_{n+1}(k + 1)$  and the above observations we conclude that

$$P_{n+1}(k + 1) = \frac{\alpha \cdot \alpha^n \cdot P_n(k + 1) - \alpha \cdot \alpha^n \cdot P_n^{(e)}(k + 1) + \alpha \cdot \alpha^n \cdot P_n^{(e)}(k)}{\alpha^{n+1}} \tag{6}$$

and thus

$$P_{n+1}(k + 1) = P_n(k + 1) + P_n^{(e)}(k) - P_n^{(e)}(k + 1) \tag{7}$$

for every  $n, k \in N$ . Adding side by side the above formula  $P_{r+1}(k + 1) = P_r(k + 1) + P_r^{(e)}(k) - P_r^{(e)}(k + 1)$  for  $r = 1, 2, \dots, n$  we arrive at

$$P_{n+1}(k + 1) = P_1(k + 1) + \sum_{r=1}^n [P_r^{(e)}(k) - P_r^{(e)}(k + 1)]. \tag{8}$$

Since  $P_1(k + 1) = 0$  for  $k \geq 1$  we have

$$P_{n+1}(k + 1) = \sum_{r=1}^n [P_r^{(e)}(k) - P_r^{(e)}(k + 1)] \tag{9}$$

for every  $n, k \in N$ . Next, adding side by side formula (9) for  $k: = k, k + 1, k + 2, k + 3, \dots, k + (n - k) - 1$  and taking into account the fact that  $\sum_{r=1}^n P_r^{(e)}(n) = 0$  for  $n \geq 2$  we have

$$\sum_{s=1}^{n-k} P_{n+1}(k + s) = \sum_{r=1}^n P_r^{(e)}(k). \tag{10}$$

One can easily see that  $P_{n+1}(k + s) = 0$  for  $s > n - k$ , where  $n > k \geq 1$ . Thus, we obtain the following expression for the distribution function of  $C_{n+1}$ :

$$P_{n+1}(C_{n+1} \leq k) = 1 - \sum_{r=1}^n P_r^{(e)}(k), \tag{11}$$

which completes the proof.  $\square$

**Corollary 1.** For every  $n$  and  $k$

$$P_{n+1}(C_{n+1} \leq k) = P_n(C_n \leq k) - P_n^{(e)}(k). \tag{12}$$

**Proof.** From (11) we have

$$1 - \sum_{r=1}^{n-1} P_r^{(e)}(k) = P_n(C_n \leq k). \tag{13}$$

Adding (11) and (13) we obtain (12).  $\square$

**Remark 2.** It follows from the above corollary that  $P_{n+1}(C_{n+1} \leq k) \leq P_n(C_n \leq k)$ .

**Corollary 2.** From (11) and the fact that [14]

$$\lim_{n \rightarrow \infty} P_n(C_n \leq k) = 0 \tag{14}$$

we deduce that

$$\sum_{r=1}^{\infty} P_r^{(e)}(k) = 1. \quad (15)$$

#### 4. Conclusions

The complexity of sequences was suggested as a statistical test of randomness of a random number generators and block ciphers [13,15] or entropy estimator [2,3,14]. It was proven in [14] that  $\lim_{n \rightarrow \infty} P_n(C_n \leq \frac{n}{\log_2 n}) = 0$ . Therefore, the sets  $K_{n,k} = \{S \in A^n : C_n(S) \leq k\}$  are good candidates for critical sets (usually  $k$  is assumed [19] to be  $\frac{n}{\log_2 n}$ ). This means, in fact, that for an arbitrarily chosen probability  $p$  close to 0 there is  $n_0$  such that for  $n > n_0$ , for a given randomly chosen sequence  $S$  the inequality  $C_n(S) \leq \frac{n}{\log_2 n}$  holds with probability less than  $p$ . Thus, it is essential to estimate  $P_n(K_{n,k}) = \sum_{s=1}^k P_n(s)$ , i.e. the levels of significance for  $K_{n,k}$ . In practice, for a fixed  $n$  these sums are computed numerically by finding all terms. Formula (12) makes it possible to find the probability  $P_{n+1}(K_{n+1,k})$  for sequences of length  $n+1$  from the probabilities  $P_n(K_{n,k})$  and  $P_n^{(e)}(k)$  for sequences of length  $n$  (the letter two can be calculated simultaneously). This reduces the computational time.

#### Acknowledgements

The author would like to thank Prof. Melvin Slater for careful reading of the manuscript. This work has been partially supported by Polish Committee for Scientific Research: Grant No. 4T07A00130.

#### References

- [1] J. Adiego, G. Navarro, P. de la Fuente, Lempel–Ziv compression of highly structured documents, *Journal of the American Society for Information Science and Technology* 58 (4) (2007) 461–478.
- [2] J.M. Amigo, J. Szczepanski, E. Wajnryb, M.V. Sanchez-Vives, Estimating the entropy rate of spike trains via Lempel–Ziv complexity, *Neural Computation* 16 (4) (2004) 717–736.
- [3] J.M. Amigo, M.B. Kennel, Variance estimators for the Lempel–Ziv entropy estimator, *Chaos* 16 (2006) 043102.
- [4] J. Bentley, D. McIlroy, Data compression with long repeated strings, *Information Sciences* 135 (1–2) (2001) 1–11.
- [5] N.J. Brittain, M.R. El-Sakka, Grayscale two-dimensional Lempel–Ziv encoding, *Image Analysis and Recognition, LNCS 3656* (2005) 328–334.
- [6] V. Castelli, L.A. Lastras-Montano, Bounds on expansion in LZ77-like coding, *IEEE Transactions on Information Theory* 52 (5) (2006) 1974–1989.
- [7] G. Chaitin, Information-theoretic limitations of formal systems, *Journal of the Association for Computing Machinery* 21 (1974) 403–424.
- [8] S. Constantinescu, L. Ilie, The Lempel–Ziv complexity of fixed points of morphisms, *SIAM Journal on Discrete Mathematics* 21 (2) (2007) 466–481.
- [9] M. Factor, D. Sheinwald, Compression in the presence of shared data, *Information Sciences* 135 (1–2) (2001) 29–41.
- [10] E. Gilbert, T. Kadota, The Lempel–Ziv algorithm and message complexity, *IEEE Transactions on Information Theory* 38 (1992) 1839–1842.
- [11] L. Ilie, S. Yu, K.Z. Zang, Word complexity and repetitions in words, *International Journal of Foundations of Computer Science* 15 (1) (2004) 41–55.
- [12] S.T. Klien, Y. Wiseman, Parallel Lempel–Ziv coding, *Discrete Applied Mathematics* 146 (2) (2005) 180–191.
- [13] A.N. Kolmogorov, Three approaches to the qualitative definition of information, *Problems of Information Transmission* 1 (1965) 1–7.
- [14] A. Lempel, J. Ziv, On the complexity of finite sequences, *IEEE Transactions on Information Theory* IT-22 (1) (1976) 75–81.
- [15] A.K. Leung, S.E. Tavares, Sequence complexity as a test for cryptographic systems, in: G.R. Blakley, D. Chaum (Eds.), *Advances in Cryptology, Crypto'84*, LNCS, vol. 196, Springer-Verlag, 1985, pp. 468–474.
- [16] T. Linder, R. Zamir, Casual coding of stationary sources and individual sequences with high resolution, *IEEE Transactions on Information Theory* 52 (2) (2006) 662–680.
- [17] G. Louchard, W. Szpankowski, J. Tang, Average profile of the generalized digital search tree and the generalized Lempel–Ziv algorithm, *SIAM Journal on Computation* 28 (3) (1999) 904–934.
- [18] Y.A. Reznik, W. Szpankowski, On the average redundancy rate of the Lempel–Ziv code with the  $k$ -error protocol, *Information Sciences* 135 (1–2) (2001) 57–70.
- [19] G. Wignarajah, Complexity tests for statistical independence, M.S. Thesis, University Toledo, 1985.
- [20] J. Ziv, A. Lempel, A universal algorithm for sequential data compression, *IEEE Transactions on Information Theory* 23 (1977) 337–343.
- [21] J. Ziv, A. Lempel, Compression of individual sequences via variable rate coding, *IEEE Transactions on Information Theory* 24 (1978) 530–536.
- [22] J. Ziv, Compression, tests for randomness and estimating the statistical model of individual sequences, in: R. Capocelli (Ed.), *SEQUENCES*, Springer-Verlag, New York, 1990, pp. 366–373.