Application of Kolmogorov complexity and universal codes to identity testing and nonparametric testing of serial independence for time series. *

Boris Ryabko†, Jaakko Astola ‡, Alex Gammerman ♯

† Institute of Computational Technology of Siberian Branch of Russian Academy of Science.
‡ Tampere University of Technology, Finland.
♯ Department of Computer Science, Royal Holloway, University of London.

Abstract

We show that Kolmogorov complexity and such its estimators as universal codes (or data compression methods) can be applied for hypothesis testing in a framework of classical mathematical statistics. The methods for identity testing and nonparametric testing of serial independence for time series are described.

AMS subject classification: 60G10, 62M07, 68Q30, 68W01, 94A29.
Keywords. *algorithmic complexity, algorithmic information theory, Kolmogorov complexity, universal coding, hypothesis testing, theory of computation, computational complexity.*

1 Introduction.

The Kolmogorov complexity, or algorithmic entropy, was suggested in [14] and was investigated in numerous papers; see for review [17]. Nowadays this notation plays important role in the theory of algorithms, information theory, artificial intelligence and many other fields, and is closely connected with such deep theoretical issues as definition of randomness, logical basis of probability theory, randomness and complexity (see [8, 17, 19, 26, 30, 31, 32, 35]). In this paper we show that Kolmogorov complexity can be applied to hypotheses testing in the framework of mathematical statistics. Moreover, we suggest using universal codes (or methods of data compression), which are estimations of Kolmogorov

*Research was supported by the joint project grant "Efficient randomness testing of random and pseudorandom number generators" of Royal Society, UK (grant ref: 15995) and Russian Foundation for Basic Research (grant no. 03-01-00495.)

complexity, for testing. In other words, in this approach the purpose is to try and apply an ostensibly theoretical theory based on the uncomputable notion of Kolmogorov complexity in the practical domain by replacing the ideal "Kolmogorov compressor" by a real-life compressor. It is important to note that such a replacing was used in [2, 16] and created a new and rapidly growing line of investigations in clustering and classification.

In this paper we consider a stationary and ergodic source (or process), which generates elements from a finite set (or alphabet) A and two problems of statistical testing. The first problem is the identity testing, which is described as follows: a hypotheses H_0^{id} is that the source has a particular distribution π and the alternative hypothesis H_1^{id} that the sequence is generated by a stationary and ergodic source, which differs from the source under H_0^{id}. One particular case where the source alphabet $A = \{0, 1\}$ and the main hypothesis H_0^{id} is that a bit sequence is generated by the Bernoulli source with equal probabilities of 0's and 1's, is applied to the randomness testing of random number and pseudo-random number generators. It is worth noting that this particular case is very close, in spirit, to the problem of randomness definition and the obtained test looks like the Martin-Löf one. The main difference is as follows: in contrast to [17, 19, 30, 35] we consider the alternative hypopiesis that the sequence is generated by a stationary and ergodic source, which, on the one hand, is natural for mathematical statistics and, on the other hand, gives a possibility to obtain explicit, non-asymptotical results.

The second problem is a generalization of the problem of nonparametric testing for independence of time series. More precisely, we consider two following hypotheses: H_0^{ind} is that the source is Markovian, which memory (or connectivity) is not larger than m, $(m \geq 0)$, and the alternative hypothesis H_1^{ind} that the sequence is generated by a stationary and ergodic source, which differs from the source under H_0^{ind}. In particular, if $m = 0$, this is the problem of testing for independence of time series. This problem is well known in mathematical statistics and there is an extensive literature dealing with nonparametric independence testing.

In both cases the testing should be based on a sample $x_1 \dots x_t$ generated by the source.

We suggest statistical tests for identity testing and nonparametric testing of serial independence for time series, which are based on Kolmogorov complexity and such estimates of it as universal codes. It is important that practically used so-called archivers can be used for suggested testing, because they can be considered as methods for estimation of Kolmogorov complexity.

This paper is intended to show that the results of theory of Kolmogorov complexity can be fruitfully applied to classic problems of mathematical statistics, which, at first glance, are far from the theory of algorithms. The applications of this approach to some other problems of mathematical statistics, its extension to the case where the alphabet is a metric space and additional examples of applications will be published in statistical literature [27, 28] (see also [29], where the first such test was described for one particular case).

The outline of the paper is as follows. The next part contains necessary

definitions and some information about universal codes and their applications. The parts three and four are devoted to the identity testing and testing of serial independence, correspondingly. The fifth part contains results of experiments, where the suggested method of identity testing is applied to pseudorandom number generators. All proofs are given in Appendix.

2 Definitions and Preliminaries.

First, we define stochastic processes (or sources of information). Consider an alphabet $A = \{a_1, \cdots, a_n\}$ with $n \geq 2$ letters and denote by A^t and A^* the set of all words of length t over A and the set of all finite words over A, correspondingly ($A^* = \bigcup_{i=1}^{\infty} A^i$). Let μ be a source which generates letters from A. Formally, μ is a probability distribution on the set of words of infinite length or, more simply, $\mu = (\mu^t)_{t \geq 1}$ is a consistent set of probabilities over the sets A^t; $t \geq 1$. By $M_\infty(A)$ we denote the set of all stationary and ergodic sources, which generate letters from A. Let $M_k(A) \subset M_\infty(A)$ be the set of Markov sources with memory (or connectivity) not greater than k, $k \geq 0$. More precisely, by definition $\mu \in M_k(A)$ if

$$\mu(x_{t+1} = a_{i_1} / x_t = a_{i_2}, x_{t-1} = a_{i_3}, \ldots, x_{t-k+1} = a_{i_{k+1}}, \ldots) = \mu(x_{t+1} = a_{i_1} / x_t = a_{i_2}, x_{t-1} = a_{i_3}, \ldots, x_{t-k+1} = a_{i_{k+1}}) \quad (1)$$

for all $t \geq k$ and $a_{i_1}, a_{i_2}, \ldots \in A$. By definition, $M_0(A)$ is the set of all Bernoulli (or i.i.d.) sources over A and $M^*(A) = \bigcup_{i=0}^{\infty} M_i(A)$ is the set of all finite-memory sources.

Now we define codes and Kolmogorov complexity. Let A^∞ be the set of all infinite words $x_1 x_2 \ldots$ over the alphabet A. A data compression method (or code) φ is defined as a set of mappings φ_n such that $\varphi_n : A^n \to \{0,1\}^*$, $n = 1, 2, \ldots$ and for each pair of different words $x, y \in A^n$ $\varphi_n(x) \neq \varphi_n(y)$. Informally, it means that the code φ can be applied for compression of each message of any length n over alphabet A and the message can be decoded if its code is known. It is also required that each sequence $\varphi_n(u_1)\varphi_n(u_2)...\varphi_n(u_r), r \geq 1$, of encoded words from the set $A^n, n \geq 1$, can be uniquely decoded into $u_1 u_2 ... u_r$. Such codes are called uniquely decodable. For example, let $A = \{a, b\}$, the code $\psi_1(a) = 0, \psi_1(b) = 00$, obviously, is not uniquely decodable. It is well known that if a code φ is uniquely decodable then the lengths of the codewords satisfy the following inequality (Kraft inequality): $\Sigma_{u \in A^n} 2^{-|\varphi_n(u)|} \leq 1$, see, for ex., [6]. (Here and below $|v|$ is the length of v, if v is a word and the number of elements of v if v is a set.) It will be convenient to reformulate this property as follows:

Claim 1. Let φ be a uniquely decodable code over an alphabet A. Then for any integer n there exists a measure μ_φ on A^n such that

$$|\varphi(u)| \geq -\log \mu_\varphi(u) \quad (2)$$

for any u from A^n . (Here and below $\log \equiv \log_2$.)

(Obviously, the claim is true for the measure

$$\mu_\varphi(u) = 2^{-|\varphi(u)|}/\Sigma_{u\in A^n}\, 2^{-|\varphi(u)|}).$$

In this paper we will use the so-called prefix Kolmogorov complexity, whose precise definition can be found in [8, 17]. Its main properties can be described as follows. There exists a uniquely decodable code κ such that i) there is an algorithm of decoding (i.e. there is a Turing machine, which maps $\kappa(u)$ to u for any $u \in A^*$) and ii) for any uniquely decodable code ψ, whose decoding is algorithmically realizable, there exists a constant C_ψ that

$$|\kappa(u)| - |\psi(u)| < C_\psi \tag{3}$$

for any $u \in A^*$. The prefix Kolmogorov complexity $K(u)$ is defined as the length of $\kappa(u)$: $K(u) = |\kappa(u)|$. The code κ is not unique, but the second property means that codelengths of two codes κ_1 and κ_2, for which i) and ii) is true, are equal up to a constant: $|\,|\kappa_1(u)| - |\kappa_2(u)|\,| < C_{1,2}$ for any word u (and the constant $C_{1,2}$ does not depend on u, see (3).) So, $K(u)$ is defined up to a constant.

In what follows we call this value "Kolmogorov complexity" and uniquely decodable codes just "codes".

We can see from ii) that the code κ is asymptotically (up to the constant) the best method of data compression, but it turns out that there is no algorithm that can calculate the codeword $\kappa(u)$ (and even $K(u)$). That is why the code κ (and Kolmogorov complexity) cannot be used for practical data compression directly. On the other hand, so-called universal codes can be realized and, in a certain sense, can be used instead of the optimal code κ, if they are applied for compression of sequences generated by any stationary and ergodic source. For their description we recall that (as it is known in Information Theory) sequences $x_1...x_t$, generated by a source p, can be "compressed" till the length $-\log p(x_1...x_t)$ bits and, on the other hand, there is no code ψ for which the average codeword length $(\,\Sigma_{x_1...x_t\in A^t}\, p(x_1...x_t)|\psi(x_1...x_t)|\,)$ is less than $-\Sigma_{x_1...x_t\in A^t}\, p(x_1...x_t)\log p(x_1...x_t)$. The universal codes can reach the lower bound $-\log p(x_1...x_t)$ asymptotically for any stationary and ergodic source p with probability 1. The formal definition is as follows: A code φ is universal if for any stationary and ergodic source p

$$\lim_{t\to\infty} t^{-1}(-\log p(x_1...x_t) - |\varphi(x_1...x_t)|) \,=\, 0 \tag{4}$$

with probability 1. So, informally speaking, universal codes estimate the probability characteristics of the source p and use them for efficient "compression".

It will be seen that the universal codes play an important role in the suggested tests, that is why we briefly mention a history of their discovery and applications to mathematical statistics.

It is interesting that the first universal code (for the set of Bernoulli sources) was briefly described by Kolmogorov in the same paper, where he defined the algorithmic complexity [14] (the same code was independently suggested and investigated by Fitingof in [5]). Then the theory of universal codes was developed in numerous papers [4, 12, 23, 24] (see also review in [15]) and now there

are many efficient algorithms of data compression which are based on universal codes. As a matter of fact, the theory of universal coding belongs to Information Theory and, at the same time, mathematical statistics, that is why it is not surprising that results of theory of universal coding have been efficiently applied to problems of prediction [9, 11, 22, 23, 26], classification [7, 33], estimation of the number of states of a finite-state source [34], estimation of the order of a Markov chain [3, 20] and some other problems of mathematical statistics.

We would like to emphasize that, in contrast to all mentioned approaches, we consider the main model of the hypothesis testing where there are two hypotheses and the Type I error is upper bounded (by a small number), see for definition [10] or any other textbook in mathematical statistics. Our approach gives a possibility to use a length of codeword of a real-life compressor as a statistical test in a framework of this main model of the mathematical statistics. In contrast to our approach, the papers [7, 33] develop asymptotical estimations of the statistical errors using different models of hypothesis testing. To our knowledge, the approach, developed in this paper, was not known before the paper [29] was published.

3 Identity Testing.

Now we consider the problem of testing H_0^{id} against H_1^{id}. Let the required level of significance (or a Type I error) be α, $\alpha \in (0,1)$. (By definition, the Type I error occurs if H_0 is true, but the test rejects H_0, and, vice versa, the Type II error occurs if H_1 is true, but the test rejects it.) We describe a statistical test which can be constructed based on any code φ.

The main idea of the suggested test is quite natural: compress a sample sequence $x_1...x_n$ by a code φ. If the length of codeword ($|\varphi(x_1...x_n)|$) is significantly less than the value $-\log \pi(x_1...x_n)$, then H_0^{id} should be rejected. The main observation is that the probability of all rejected sequences is quite small for any φ, that is why the Type I error can be made small. The precise description of the test is as follows: *The hypothesis H_0^{id} is accepted if*

$$-\log \pi(x_1...x_n) - |\varphi(x_1...x_n)| \leq -\log \alpha. \tag{5}$$

Otherwise, H_0^{id} is rejected. We denote this test by $\Gamma^{(n)}_{\pi,\alpha,\varphi}$.

Theorem 1.

i) For each distribution $\pi, \alpha \in (0,1)$ and a code φ, the Type I error of the described test $\Gamma^{(n)}_{\pi,\alpha,\varphi}$ is not larger than α.

ii) If, in addition, π is a finite-memory stationary and ergodic process over A^∞ (i.e. $\pi \in M^*(A)$) and φ is a universal code, then the Type II error of the test $\Gamma^{(n)}_{\pi,\alpha,\varphi}$ goes to 0, when n tends to infinity.

Remarks. The suggested tests is deeply connected with theory of Kolmogorov complexity and its applications.

First, in fact, the described test (5) coincides with the Martin-Löf one. Indeed, the universal π-Martin-Löf test, in a computable approximation based

on the compressor φ inducing a probability mass function $\pi_\varphi(x_1 \dots x_n) = 2^{-|\varphi(x_1 \dots x_n)|}$, is as follows: if

$$\log(\pi_\varphi(x_1 \dots x_n)/\pi(x_1 \dots x_n)) \le -\log \alpha,$$

then H_0, else H_1; see [17, 19]. Obviously, it is the same inequality as (5).

Second, the Kolmogorov complexity can be used instead of the length of a code in the described test (5). Namely, let $K^{(n)}_{\pi,\alpha}$ be the following test: the hypothesis H_0^{id} is accepted if $-\log \pi(x_1 ... x_n) - K(x_1 ... x_n) \le -\log \alpha$, otherwise, H_0^{id} is rejected. Theorem 1 is valid for this test, too.

4 Testing of Serial Independence

First, we give some additional definitions. Let v be a word $v = v_1 ... v_k, k \le t, v_i \in A$. Denote the rate of a word v occurring in the sequence $x_1 x_2 \dots x_k$, $x_2 x_3 \dots x_{k+1}$, $x_3 x_4 \dots x_{k+2}$, ..., $x_{t-k+1} \dots x_t$ as $\nu^t(v)$. For example, if $x_1 ... x_t =$ 000100 and $v = 00$, then $\nu^6(00) = 3$. Now we define for any $k \ge 0$ a so-called empirical Shannon entropy of order k as follows:

$$h^*_k(x_1 \dots x_t) = -\frac{1}{(t-k)} \sum_{v \in A^k} \bar{\nu}^t(v) \sum_{a \in A} (\nu^t(va)/\bar{\nu}^t(v)) \log(\nu^t(va)/\bar{\nu}^t(v)) \,, \quad (6)$$

where $k < t$ and $\bar{\nu}^t(v) = \sum_{a \in A} \nu^t(va)$. In particular, if $k = 0$, we obtain $h^*_0(x_1 \dots x_t) = -\frac{1}{t} \sum_{a \in A} \nu^t(a) \log(\nu^t(a)/t)$,

Let, as before, H_0^{ind} be that the source π is Markovian with memory (or connectivity) not grater than m, $(m \ge 0)$, and the alternative hypothesis H_1^{ind} be that the sequence is generated by a stationary and ergodic source, which differs from the source under H_0^{ind}. The suggested test is as follows.

Let ψ be any code. By definition, the hypothesis H_0^{ind} is accepted if

$$(t-m) h^*_m(x_1 ... x_t) - |\psi(x_1 ... x_t)| \le \log(1/\alpha) \,, \quad (7)$$

where $\alpha \in (0,1)$. Otherwise, H_0^{ind} is rejected. We denote this test by $\Upsilon^t_{\alpha,\,\psi,m}$.

Theorem 2. *i) For any distribution π and any code ψ the Type I error of the test $\Upsilon^t_{\alpha,\,\psi,m}$ is less than or equal to $\alpha, \alpha \in (0,1)$.*

ii) *If, in addition, π is a stationary and ergodic process over A^∞ and ψ is a universal code, then the Type II error of the test $\Upsilon^t_{\alpha,\,\psi,m}$ goes to 0, when t tends to infinity.*

Comment. If we use Kolmogorov complexity $K(x_1 ... x_n)$ instead of the length of the code $|\psi(x_1 ... x_t)|$, the obtained test will have the same properties.

5 Experiments

We applied the described method of identity testing to pseudorandom number generators. More precisely, we denote by U a source, which generates equiprobable and independent symbols from the alphabet $\{0,1\}$ and consider the hypothesis H_0^{id} that a sequence is generated by U.

We have taken linear congruent generators (LCG), which are defined by the following equality

$$X_{n+1} = (A * X_n + C) \, mod \, M,$$

where X_n is the n-th generated number [13]. Each such generator we will denote by $LCG(M, A, C, X_0)$, where X_0 is the initial value of the generator. We considered the four following LCG: $L_1 = LCG(10^8 + 1, 23, 0, 47594118)$, $L_2 = LCG(2^{31}, 2^{16} + 3, 0, 1)$, $L_3 = LCG(2^{32}, 134775813, 1, 0)$ and $L_4 = LCG(2^{32}, 69069, 0, 1)$.

In our experiments we extracted an eight-bit word from each generated X_i using the following algorithm. Firstly, the number $\mu = \lfloor M/256 \rfloor$ was calculated and then each X_i was transformed into an 8-bit word $\hat{X}_i$ as follows:

$$\left.\begin{array}{c} \hat{X}_i = \lfloor X_i/256 \rfloor \;\; if X_i < 256\mu \\ \hat{X}_i = empty \; word \;\; if X_i \geq 256\mu \end{array}\right\} \tag{8}$$

Then a sequence was compressed by the archiver *ACE v 1.2b* (see http://www.winace.com/). Experimental data about testing of four linear congruent generators is given in the table.

Table 1: Results of experiments

generator / length (bits)	400 000	8 000 000
L_1	390 240	7635936
L_2	extended	7797984
L_3	extended	extended
L_4	extended	extended

So, we can see from the first line of the table that the 400000−bit sequence generated by L_1 and transformed according to (8), was compressed to a 390240−bit sequence. (Here 400000 is the length of the sequence after transformation.) If we take the level of significance, say, 0.001 ($\alpha = 0.001$) and take into account that $0.001 \geq 2^{-9760}$ and apply the test $\Gamma^{(400000)}_{U,\alpha,\varphi}$,($\varphi = ACE \; v \; 1.2b$), the hypothesis H_0 should be rejected, see Theorem 1 and (5). Analogously, the second line of the table shows that the 8000000−bit sequence generated by L_2 cannot be considered as random. (Indeed, H_0^{id} should be rejected because the level of significance 0.001 is greater than $2^{-202016}$.) On the other hand, the suggested test accepts H_0^{id} for the sequences generated by the two latter generators, because the lengths of the "compressed" sequences increased.

The obtained information corresponds to the known data about the generators mentioned above. Thus, it is shown in [13] that L_1 and L_2 are bad, whereas L_3 and L_4 were investigated in [21] and [18], correspondingly, and are regarded as good. So, we can see that the suggested testing is quite efficient.

Some other examples of application of the identity testing and serial independence testing are described in [28, 29] and show that the suggested method can be useful in practice.

6 Appendix.

The following well known inequality, whose proof can be found in [6], will be used in proofs of both theorems.

Lemma. Let p and q be two probability distributions over some alphabet B. Then $\sum_{b\in B} p(b)\log(p(b)/q(b)) \geq 0$ with equality if and only if $p = q$.

Proof of Theorem 1. Let C_α be a critical set of the test $\Gamma^{(n)}_{\pi,\alpha,\varphi}$, i.e., by definition, $C_\alpha = \{u : u \in A^t \ \& \ -\log\pi(u) - |\varphi(u)| > -\log\alpha\}$. Let μ_φ be a measure for which the claim 1 is true. We define an axillary set

$$\hat{C}_\alpha = \{u : -\log\pi(u) - (-\log\mu_\varphi(u)) > -\log\alpha\}.$$

We have

$$1 \geq \sum_{u\in\hat{C}_\alpha} \mu_\varphi(u) \geq \sum_{u\in\hat{C}_\alpha} \pi(u)/\alpha = (1/\alpha)\pi(\hat{C}_\alpha).$$

(Here the second inequality follows from the definition of $\hat{C}_\alpha$, whereas all others are obvious.) So, we obtain that $\pi(\hat{C}_\alpha) \leq \alpha$. From definitions of $C_\alpha, \hat{C}_\alpha$ and (2) we immediately obtain that $\hat{C}_\alpha \supset C_\alpha$. Thus, $\pi(C_\alpha) \leq \alpha$. By definition, $\pi(C_\alpha)$ is the value of the Type I error. The first statement of the theorem 1 is proven.

Let us prove the second statement of the theorem. Suppose that the hypothesis H_1^{id} is true. That is, the sequence $x_1 \dots x_t$ is generated by some stationary and ergodic source τ and $\tau \neq \pi$. Our strategy is to show that

$$\lim_{t\to\infty} -\log\pi(x_1\dots x_t) - |\varphi(x_1\dots x_t)| = \infty \tag{9}$$

with probability 1 (according to the measure τ). First we represent (9) as

$$-\log\pi(x_1\dots x_t) - |\varphi(x_1\dots x_t)| \\ = t(\frac{1}{t}\log\frac{\tau(x_1\dots x_t)}{\pi(x_1\dots x_t)} + \frac{1}{t}(-\log\tau(x_1\dots x_t) - |\varphi(x_1\dots x_t)|)).$$

From this equality and the property of a universal code (4) we obtain

$$-\log\pi(x_1\dots x_t) - |\varphi(x_1\dots x_t)| = t\,(\frac{1}{t}\log\frac{\tau(x_1\dots x_t)}{\pi(x_1\dots x_t)} + o(1)). \tag{10}$$

Now we use some results of the ergodic theory and the information theory, which can be found, for ex., in [1]. First, according to the Shannon-MacMillan-Breiman theorem, there exists the limit $\lim_{t\to\infty} -\log\tau(x_1\dots x_t)/t$ (with probability 1) and this limit is equal to so-called limit Shannon entropy, which we denote as $h_\infty(\tau)$. Second, it is known that for any integer k the following inequality is true: $h_\infty(\tau) \leq -\sum_{v\in A^k}\tau(v)\sum_{a\in A}\tau(a/v)\log\tau(a/v)$. (Here the right hand value is called $m-$ order conditional entropy). It will be convenient to represent both statements as follows:

$$\lim_{t\to\infty} -\log\tau(x_1\dots x_t)/t \leq -\sum_{v\in A^k}\tau(v)\sum_{a\in A}\tau(a/v)\log\tau(a/v) \tag{11}$$

for any $k \geq 0$ (with probability 1). It is supposed that the process π has a finite memory, i.e. belongs to $M_s(A)$ for some s. Having taken into account the definition of $M_s(A)$ (1), we obtain the following representation:

$$-\log \pi(x_1 \dots x_t)/t = -t^{-1} \sum_{i=1}^{t} \log \pi(x_i/x_1 \dots x_{i-1})$$
$$= -t^{-1}(\sum_{i=1}^{k} \log \pi(x_i/x_1 \dots x_{i-1}) + \sum_{i=k+1}^{t} \log \pi(x_i/x_{i-k} \dots x_{i-1}))$$

for any $k \geq s$. According to the ergodic theorem there exists a limit

$$\lim_{t\to\infty} t^{-1} \sum_{i=k+1}^{t} \log \pi(x_i/x_{i-k} \dots x_{i-1}),$$

which is equal to $-\sum_{v\in A^k} \tau(v) \sum_{a\in A} \tau(a/v) \log \pi(a/v)$, see [1, 6]. So, from the two latter equalities we can see that

$$\lim_{t\to\infty} (-\log \pi(x_1 \dots x_t))/t = -\sum_{v\in A^k} \tau(v) \sum_{a\in A} \tau(a/v) \log \pi(a/v).$$

Taking into account this equality, (11) and (10), we can see that

$$-\log \pi(x_1 \dots x_t) - |\varphi(x_1 \dots x_t)| \geq t\,(\sum_{v\in A^k} \tau(v) \sum_{a\in A} \tau(a/v) \log(\tau(a/v)/\pi(a/v))) + o(t)$$

for any $k \geq s$. From this inequality and the Lemma we can obtain that $-\log \pi(x_1 \dots x_t) - |\varphi(x_1 \dots x_t)| \geq c\,t + o(t)$, where c is a positive constant, $t \to \infty$. Hence, (9) is true and the theorem is proven.

Proof of Theorem 2. First we show that for any source $\theta^* \in M_0(A)$ and any word $x_1 \dots x_t \in A^t, t > 1$, the following inequality is valid:

$$\theta^*(x_1 \dots x_t) = \prod_{a\in A} (\theta^*(a))^{\nu^t(a)} \leq \prod_{a\in A} (\nu^t(a)/t)^{\nu^t(a)} \qquad (12)$$

Here the equality holds, because $\theta^* \in M_0(A)$. The inequality follows from the Lemma. Indeed, if $p(a) = \nu^t(a)/t$ and $q(a) = \theta^*(a)$, then $\sum_{a\in A} \frac{\nu^t(a)}{t} \log \frac{(\nu^t(a)/t)}{\theta^*(a)} \geq 0$. From the latter inequality we obtain (12).

Let now θ belong to $M_m(A), m > 0$. We will prove that for any $x_1 \dots x_t$

$$\theta(x_1 \dots x_t) \leq \prod_{u\in A^m} \prod_{a\in A} (\nu^t(ua)/\bar{\nu}^t(u))^{\nu^t(ua)} . \qquad (13)$$

Indeed, we can present $\theta(x_1 \dots x_t)$ as

$$\theta(x_1 \dots x_t) = \theta(x_1 \dots x_m) \prod_{u\in A^m} \prod_{a\in A} \theta(a/u)^{\nu^t(ua)} ,$$

where $\theta(x_1 \dots x_m)$ is the limit probability of the word $x_1 \dots x_m$. Hence, $\theta(x_1 \dots x_t)$ $\leq \prod_{u \in A^m} \prod_{a \in A} \theta(a/u)^{\nu^t(ua)}$. Taking into account the inequality (12), we obtain

$$\prod_{a \in A} \theta(a/u)^{\nu^t(ua)} \leq \prod_{a \in A} (\nu^t(ua)/\bar{\nu}^t(u))^{\nu^t(ua)}$$

for any word u. So, from the last two inequalities we obtain (13).

It will be convenient to define two auxiliary measures on A^t as follows:

$$\pi_m(x_1 ... x_t) = \Delta \, 2^{-t \, h^*_m(x_1 ... x_t)} \, , \; \sigma(x_1 ... x_t) = 2^{-|\psi(x_1 ... x_t)|} \tag{14}$$

where $x_1 ... x_t \in A^t$ and $\Delta = (\sum_{x_1 ... x_t \in A^t} 2^{-t \, h^*_m(x_1 ... x_t)})^{-1}$. If we take into account that $2^{-(t-m) \, h^*_m(x_1 ... x_t)} = \prod_{u \in A^m} \prod_{a \in A} (\nu^t(ua)/\bar{\nu}^t(u))^{\nu^t(ua)}$, we can see from (13) and (14) that, for any measure $\theta \in M_m(A)$ and any $x_1 \dots x_t \in A^t$,

$$\theta(x_1 \dots x_t) \leq \pi_m(x_1 ... x_t)/\Delta \, . \tag{15}$$

Let us denote the critical set of the test $\Upsilon^t_{\alpha, \sigma, m}$ as C_α, i.e., by definition, $C_\alpha = \{x_1 \dots x_t : \; (t-m) h^*_m(x_1 \dots x_t) - |\psi(x_1 ... x_t)|) > \log(1/\alpha)\}$. From (14) we obtain

$$C_\alpha = \{x_1 \dots x_t : \; (t-m) \, h^*_m(x_1 \dots x_t) - (-\log \sigma(x_1 ... x_t)) \,) > \log(1/\alpha)\}. \tag{16}$$

From (15) and (16) we can see that for any measure $\theta \in M_m(A)$

$$\theta(C_\alpha) \leq \pi_m(C_\alpha)/\Delta \, . \tag{17}$$

From (16) and (14) we obtain

$$\begin{aligned} C_\alpha &= \{x_1 \dots x_t : \; 2^{\,(t-m) \, h^*_m(x_1 ... x_t)} > (\alpha \; \sigma(x_1 \dots x_t))^{-1}\} \\ &= \{x_1 \dots x_t : \; (\pi_m(x_1 \dots x_t)/\Delta)^{-1} > (\alpha \; \sigma(x_1 \dots x_t))^{-1}\} \, . \end{aligned}$$

Finally,

$$C_\alpha = \{x_1 \dots x_t : \; \sigma(x_1 \dots x_t) > \pi_m(x_1 \dots x_t)/(\alpha \, \Delta)\}. \tag{18}$$

The following chain of inequalities and equalities is valid:

$$\begin{aligned} 1 \geq \sum_{x_1 ... x_t \in C_\alpha} \sigma(x_1 \dots x_t) &\geq \sum_{x_1 ... x_t \in C_\alpha} \pi_m(x_1 \dots x_t)/(\alpha \, \Delta) \\ &= \pi_m(C_\alpha)/(\alpha \, \Delta) \geq \theta(C_\alpha)\Delta/(\alpha \, \Delta) = \theta(C_\alpha)/\alpha. \end{aligned}$$

(Here both equalities and the first inequality are obvious, the second and the third inequalities follow from (18) and (17), correspondingly.) So, we obtain that $\theta(C_\alpha) \leq \alpha$ for any measure $\theta \in M_m(A)$. Taking into account that C_α is the critical set of the test, we can see that the probability of the Type I error is not greater than α. The first claim of the theorem is proven.

The proof of the second statement of the theorem will be based on some results of Information Theory. The $t-$ order conditional Shannon entropy is defined as follows:

$$h_t(p) = - \sum_{x_1 ... x_t \in A^t} p(x_1 ... x_t) \sum_{a \in A} p(a/x_1 ... x_t) \log p(a/x_1 ... x_t), \tag{19}$$

where $p \in M_\infty(A)$. It is known that for any $p \in M_\infty(A)$ firstly, $\log|A| \geq h_0(p) \geq h_1(p) \geq ...$, secondly, there exists limit Shannon entropy $h_\infty(p) = \lim_{t\to\infty} h_t(p)$, thirdly, $\lim_{t\to\infty} -t^{-1}\log p(x_1...x_t) = h_\infty(p)$ with probability 1 and, finally, $h_m(p)$ is strictly greater than $h_\infty(p)$, if the memory of p is grater than m, (i.e. $p \in M_\infty(A) \setminus M_m(A)$), see, for example, [1, 6].

Taking into account the definition of the universal code (4), we obtain from the above described properties of the entropy that

$$\lim_{t\to\infty} t^{-1}|\psi(x_1...x_t)| = h_\infty(p) \tag{20}$$

with probability 1. It can be seen from (6) that h^*_m is an estimate for the $m-$order Shannon entropy (19). Applying the ergodic theorem we obtain $\lim_{t\to\infty} h^*_m(x_1\dots x_t) = h_m(p)$ with probability 1; see [1, 6]. Having taken into account that $h_m(p) > h_\infty(p)$ and (20) we obtain from the last equality that $\lim_{t\to\infty}((t-m)\, h^*_m(x_1\dots x_t) - |\psi(x_1...x_t)|) = \infty$. This proves the second statement of the theorem.

References

[1] P. Billingsley , Ergodic theory and information. John Wiley & Sons, (1965).

[2] R. Cilibrasi, P.M.B. Vitanyi, Clustering by compression, IEEE Trans. Information Theory, 51(4) (2005), pp. 1523- 1545.

[3] I. Csiszár, P. Shields, The consistency of the BIC Markov order estimation, Annals of Statistics, 6(2000), pp. 1601-1619.

[4] M. Effros, K. Visweswariah, S.R. Kulkarni, S. Verdu, Universal lossless source coding with the Burrows Wheeler transform, IEEE Trans. Inform. Theory 48(5) (2002) pp. 1061 - 1081.

[5] B.M. Fitingof, Optimal encoding for unknown and changing statistica of messages, Problems of Information Transmission, 2(2)(1966) pp. 3-11.

[6] R.G. Gallager, Information Theory and Reliable Communication, John Wiley & Sons, New York (1968).

[7] M.Gutman, Asymptotically Optimal Classification for Multiple Tests with Empirically Observed Statistics, IEEE Trans. Inform. Theory, 35(2) (1989) pp. 401-408.

[8] M. Hutter. Universal Artificial Intelligence. Sequential Decisions based on algorithmic probability, Springer-Verlag (2005).

[9] P. Jacquet, W. Szpankowski, L.Apostol, Universal predictor based on pattern matching, IEEE Trans. Inform. Theory, 48 (2002), pp. 1462-1472.

[10] M.G. Kendall, A. Stuart, The advanced theory of statistics; Vol.2: Inference and relationship, London, (1961).

[11] J. Kieffer, Prediction and Information Theory, Preprint, (1998), (available at ftp://oz.ee.umn.edu/users/kieffer/papers/prediction.pdf/)

[12] J.C. Kieffer, En-Hui Yang, Grammar-based codes: a new class of universal lossless source codes, IEEE Transactions on Information Theory, 46(3)(2000) pp. 737 - 754.

[13] D.E.Knuth, The art of computer programming. Vol.2. Addison Wesley, (1981).

[14] A.N.Kolmogorov, Three approaches to the quantitative definition of information, Problems of Inform. Transmission, 1 (1965), pp.3-11.

[15] R. Krichevsky, Universal Compression and Retrival, Kluver Academic Publishers, (1993).

[16] M. Li, X. Chen, X. Li, B. Ma, P.M.B. Vitanyi, The similarity metric, IEEE Trans. Inform. Th., 50(12) (2004), pp. 3250- 3264.

[17] M. Li , P. Vitanyi, An Introduction to Kolmogorov Complexity and Its Applications, Springer-Verlag, New York, 2nd Edition (1997).

[18] G. Marsaglia and A. Zaman. Monkey tests for random number generators, Computers Math. Applic., 26 (1993), pp.1-10.

[19] P. Martin-Löf, The definition of random sequences, Information and Control, 9 (1966), pp.602-619.

[20] N.Merhav, M.Gutman, J.Ziv, On the estimation of the order of a Markov chain and universal data compression, IEEE Trans. Inform. Theory, 35(5) (1989), pp. 1014-1019.

[21] O. Moeschlin, E. Grycko, C. Pohl, and F. Steinert, Experimental Stochastics, Springer-Verlag, Berlin Heidelberg, (1998).

[22] A.B. Nobel, On optimal sequential prediction, IEEE Trans. Inform. Theory, 49(1), (2003), pp. 83-98.

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

[24] B.Ya. Ryabko, Twice-universal coding, Problems of Information Transmission, 20(3) (1984), pp. 173-177.

[25] B.Ya. Ryabko, Prediction of random sequences and universal coding, Problems of Inform. Transmission 24(2) (1988), pp. 87-96.

[26] B.Ya. Ryabko, The complexity and effectiveness of prediction algorithms, J. of Complexity, 10 (1994), pp.281-295.

[27] B.Ryabko, J. Astola, Universal Codes as a Basis for Nonparametric Testing of Serial Independence for Time Series, Journal of Statistical Planning and Inference, accepted.

[28] B. Ryabko, J.Astola Universal Codes as a Basis for Time Series Testing, Statistical Methodology, submitted.

[29] B. Ryabko, V. Monarev, Using Information Theory Approach to Randomness Testing, Journal of Statistical Planning and Inference, 133(1)(2005), pp.95-110. (The preliminary version was published in: Cryptology ePrint Archive: Report 2003/127, http://eprint.iacr.org/2003/127 , 21 Jun 2003.)

[30] V.A.Uspenskii, A.L.Semenov, A.K.Shen Can an individual sequence of zeros and ones be random?, Russian Mathematical Surveys, 45, (1990).

[31] N.Vereshchagin, P.M.B. Vitanyi, Kolmogorov's structure functions with application to the foundations of model selections. In: Proc. 43th Symposium on Foundations of Computer Science, (2002), pp. 751- 760.

[32] P.M.B. Vitanyi, M.Li Minimum description length induction, Bayesianism, and Kolmogorov complexity, IEEE Trans. Inform. Theory, 46(2) (2000), pp. 446-464.

[33] J.Ziv, On Classification with Empirically Observed Statistics and Universal Data Compression, IEEE Trans. Inform. Theory, 34(2) (1978), pp. 278- 286.

[34] J.Ziv, N.Merhav, Estimating the Number of States of a Finite-State Source, IEEE Trans. Inform. Theory, 38(1) (1992), pp. 61-65.

[35] A.K.Zvonkin, L.A.Levin, The complexity of finite objects and concepts of information and randomness through the algorithm theory, Uspehi Math. Nauk 25(6) (1970).