2005 International Conference on Analysis of Algorithms DMTCS proc. **AD**, 2005, 399–408

Application of data compression methods to hypothesis testing for ergodic and stationary processes

Boris Ryabko[1†] and Jaakko Astola[2]

[1] *Institute of Computational Technology of Siberian Branch of Russian Academy of Science.* `boris@ryabko.net`
[2] *Tampere University of Technology, Finland.* `jta@cs.tut.fi`

We show that data compression methods (or universal codes) can be applied for hypotheses testing in a framework of classical mathematical statistics. Namely, we describe tests, which are based on data compression methods, for the three following problems: i) identity testing, ii) testing for independence and iii) testing of serial independence for time series. Applying our method of identity testing to pseudorandom number generators, we obtained experimental results which show that the suggested tests are quite efficient.

Keywords: hypothesis testing, data compression, universal coding, Information Theory, universal predictors, Shannon entropy.

1 Introduction

In this paper, we suggest a new approach to testing statistical properties of stationary and ergodic processes. In contrast to known methods, the suggested approach gives a possibility to make tests, based on any lossless data compression method even if the distribution law of the codeword lengths is not known. We describe three statistical tests, which are based on this approach.

We consider a stationary and ergodic source (or process), which generates elements from a finite set (or alphabet) A and three problems of statistical testing. The fist 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 in which 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 randomness testing of random number and pseudorandom number generators. Tests for this particular case were investigated in [20] and the test suggested below can be considered as a generalization of the methods from [20]. We carried out some experiments, where the suggested method of identity testing was applied to pseudorandom number generators. The results show that the suggested methods are quite efficient.

The second problem is a generalization of the problem of nonparametric testing for serial independence of time series. More precisely, we consider the following two hypotheses: H_0^{SI} is that the source is Markovian with memory (or connectivity) not larger than m, $(m \geq 0)$, and the alternative hypothesis H_1^{SI} that the sequence is generated by a stationary and ergodic source which differs from the source under H_0^{SI}. (This problem is considered by the authors in [19].) In particular, if $m = 0$, that is the problem of testing for independence of time series, which is well known in mathematical statistics [7].

The third problem is the independence test. In this case it is assumed that the source is Markovian, whose memory is not larger than m, $(m \geq 0)$, and the source alphabet can be presented as a product of d alphabets $A_1, A_2, \ldots, A_d$ (i.e. $A = \prod_{i=1}^{d} A_i$). The main hypothesis H_0^{ind} is that $p(x_{m+1} = (a_{i_1}, \ldots, a_{i_d})/x_1...x_m) = \prod_{j=1}^{d} p(x_{m+1}^j = a_{i_j}/x_1...x_m)$ for each $(a_{i_1}, \ldots, a_{i_d}) \in \prod_{i=1}^{d} A_i$, where $x_{m+1} = (x_{m+1}^1, ..., x_{m+1}^d)$. The alternative hypothesis H_1^{ind} is that the sequence is generated by a Markovian source with memory not larger than m, $(m \geq 0)$, which differs from the source under H_0^{ind}.

[†]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.)

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

All three problems are well known in mathematical statistics and there is an extensive literature dealing with their nonparametric testing, see, for ex., [7, 9].

We suggest nonparametric statistical tests for these problems. The tests are based on methods of data compression, which are deeply connected with universal codes and universal predictors. It is important to note that practically used so-called archivers can be used for suggested testing. It is no surprise that the results and ideas of universal coding theory can be applied to some classical problems of mathematical statistics. In fact, the methods of universal coding (and a closely connected universal prediction) are intended to extract information from observed data in order to compress (or predict) data efficiently when the source statistics are unknown.

It is important to note that, on the one hand, the universal codes and archivers are based on results of Information Theory, the theory of algorithms and some other branches of mathematics; see, for example, [4, 10, 13, 14, 18]. On the other hand, the archivers have shown high efficiency in practice as compressors of texts, DNA sequences and many other types of real data. In fact, archivers can find many kinds of latent regularities, that is why they look like a promising tool for identity and independence testing; see also [2].

The outline of the paper is as follows. The next section contains definitions and necessary information. Section 3 is devoted to the description of the tests and their properties. In Section 4 the new tests are experimentally compared with methods from [15]. 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) $k,\ k \geq 0$. More precisely, by definition $\mu \in M_k(A)$ if

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

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.

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 $A^n, n \geq 1,$ could 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, Claim 1 is true for the measure

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

In what follows we call uniquely decodable codes just ”codes”.

There exist so-called universal codes. For their description we recall that (as it is known in Information Theory) sequences $x_1 \ldots x_t$, generated by a source p, can be ”compressed” till the length $-\log p(x_1 ... x_t)$ bits and, on the other hand, for any source p there is no code ψ for which the average codeword length ($\Sigma_{u \in A^t} p(u) |\psi(u)|$) is less than $-\Sigma_{u \in A^t} p(u) \log p(u)$. 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

$$\lim_{t\to\infty} t^{-1}(-\log p(x_1...x_t) - |\varphi(x_1...x_t)|) = 0 \tag{3}$$

with probability 1. So, informally speaking, universal codes estimate the probability characteristics of the source p and use them for efficient "compression". One of the first universal codes was described in [16], see also [17]. Now there are many efficient universal codes (and universal predictors connected with them), which are described in numerous papers, see [8, 10, 12, 13, 14, 18].

3 The tests.

3.1 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 $\alpha,\ \alpha \in (0,1)$. (By definition, the Type I error occurs if H_0 is true, but the test rejects H_0.) 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 the 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*

$$-\log \pi(x_1...x_n) - |\varphi(x_1...x_n)| \le -\log \alpha. \tag{4}$$

Otherwise, H_0^{id} is rejected. (Here π is a given distribution and $\alpha \in (0,1)$.) 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 and 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.*

3.2 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 $x_1x_2\dots x_k$, $x_2x_3\dots x_{k+1}$, $x_3x_4\dots x_{k+2}$, ..., $x_{t-k+1}\dots x_t$ 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 $0 \le k < t$ a so- called empirical Shannon entropy of order k as follows:

$$h_k^*(x_1\dots x_t) = -\frac{1}{(t-k)}\sum_{v\in A^k}\bar\nu^t(v)\sum_{a\in A}(\nu^t(va)/\bar\nu^t(v))\log(\nu^t(va)/\bar\nu^t(v)), \tag{5}$$

where $\bar\nu^t(v) = \sum_{a\in A}\nu^t(va)$. 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^{SI} be that the source π is Markovian with memory (or connectivity) not greater than $m,\ (m \ge 0)$, and the alternative hypothesis H_1^{SI} be that the sequence is generated by a stationary and ergodic source, which differs from the source under H_0^{SI}. The suggested test is as follows.

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

$$(t-m)h_m^*(x_1...x_t) - |\psi(x_1...x_t)| \le \log(1/\alpha), \tag{6}$$

where $\alpha \in (0,1)$. Otherwise, H_0^{SI} 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 $\alpha, \alpha \in (0,1)$ and, 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.*

3.3 Independence Testing.

Now we consider the problem of the independence testing for Markovian sources. More precisely, in this subsection we suppose that it is known a priori that a source belongs to $M_m(A)$ for some known $m, m \ge 0$. We will consider sources, which generate letters from an alphabet $A = \prod_{i=1}^d A_i,\ d \ge 2,$ and present each generated letter x_i as the following string: $x_i = (x_i^1, \ldots, x_i^d)$, where $x_i^j \in A_j$. The hypothesis H_0^{ind} is that a sequence $x_1...x_t$ is generated by such a source $\mu \in M_k(A)$ that for each $a = (a_1, \ldots, a_d) \in \prod_{i=1}^d A_i$ and each $x_1...x_m \in A^m$ the following equality is valid:

$$\mu(x_{m+1} = (a_1, \ldots, a_d)/x_1...x_m) = \prod_{i=1}^{d} \mu^i(x^i_{m+1} = a_i/x_1...x_m), \tag{7}$$

where, by definition,

$$\mu^i(x^i_{m+1} = a_i/x_1...x_m) = \sum_{b_1,\ldots,b_{i-1} \in \prod_{j=1}^{i-1} A_j} \; \sum_{b_{i+1},\ldots,b_d \in \prod_{j=i+1}^{d} A_j} \mu(x_{m+1} = (b_1, \ldots, b_{i-1}, a_i, b_{i+1}, \ldots, b_d)/x_1...x_m). \tag{8}$$

The hypothesis H_1^{ind} is that the source belongs to $M_m(A)$ and the equation (7) is not valid at least for one $(a_1, \ldots, a_d) \in \prod_{i=1}^d A_i$ and $x_1...x_m \in A^m$.

Let us describe a test for hypotheses H_0^{ind} and H_1^{ind}. *Let φ be any code. By definition, the hypothesis H_0^{ind} is accepted if*

$$\sum_{i=1}^{d} (t-m) h^*_m(x_1^i...x_t^i) - |\varphi(x_1...x_t)| \le \log(1/\alpha)\,, \tag{9}$$

where $(x_1, ..., x_t) = (x_1^1, x_1^2, ...x_1^d), (x_2^1, x_2^2, ...x_2^d), \ldots, (x_t^1, x_t^2, ...x_t^d)$ and $\alpha \in (0,1)$. Otherwise, H_0^{ind} is rejected. We denote this test by $\Phi^t_{\alpha,\,\varphi,m}$. First we give an informal explanation of the main idea of the test. The Shannon entropy is the lower bound of the compression ratio and the empirical entropy $h^*_m(x_1^i...x_t^i)$ is its estimate. So, if H_0^{ind} is true, the sum $\sum_{i=1}^d (t-m) h^*_m(x_1^i...x_t^i)$ is, on average, close to lower bound. Hence, if the length of a codeword of some code φ is significantly less than the sum of the empirical entropies, it means that there is some dependence between components, which is used for some additional compression. The following theorem describes the properties of the suggested test.

Theorem 3. *i) For any distribution $\mu \in M_m(A)$ and any code φ the Type I error of the test $\Phi^t_{\alpha,\,\varphi,m}$ is less than or equal to $\alpha, \alpha \in (0,1)$ and ii) if, in addition, φ is a universal code, then the Type II error of the test $\Upsilon^t_{\alpha,\,\varphi,m}$ goes to 0, when t tends to infinity.*

4 Experiments

In this section we describe some experiments carried out to compare new tests with known ones. We consider a problem of the randomness testing, i.e. a particular case of the identity testing, where the source alphabet is $A = \{0,1\}$ 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.

We have compared tests which are based on archivers RAR and ARJ, and tests from [15]. The point is that the tests from [15] are selected basing on comprehensive theoretical and experimental analysis and can be considered as the state-of-the-art in randomness testing.

The behavior of the tests was investigated for files of various lengths generated by the pseudo random generator RANDU, whose description can be found in [5]. We generated 100 different files of each length and applied each test from [15] to each file with level of significance 0.01. So, if a test is applied to a truly random bit sequence, on average 1 file from 100 should be rejected. All results are given in the table, where integers in the cells are the numbers of rejected files (from 100). For example, the first number of the fourth row of the table 1 is 2. It means that there were 100 files of the length $5\;10^4$ bits generated by PRNG RANDU. When the Frequency test from [15] was applied, the hypothesis H_0 was rejected 2 times from 100 (and, correspondingly, H_0 was accepted 98 times.) If a number of rejections is not given for a certain length and test, it means that the test cannot be applied for files of such length.

When we used archivers RAR and ARJ, we applied each method to a file and first estimated the length of compressed data. Then we used the test $\Gamma^{(t)}_{uniform,\alpha,\varphi}$ with the critical value $1/256$ as follows. The length of a file (in bits) is equal to $8n$ (before compression), where n is the length in bytes. So, taking $\alpha = 1/256$, we see that the hypothesis about randomness (H_0^{id}) should be rejected, if the length of compressed file is less than or equal to $8n - 8$ bits. Taking into account that the length of computer files is measured in bytes, we use the very simple rule : if the $n-$byte file is really compressed (i.e. the length of the encoded file is $n-1$ bytes or less), this file is not random (and H_0^{id} is rejected). So, the following table contains numbers of cases, where files were really compressed.

Let us now give some comments about parameters of the methods from [15]. The point is that there are some tests from [15], where parameters can be chosen from a certain interval. In such cases we repeated all calculations three times, taking the minimal possible value of the parameter, the maximal one and the average one. Then the data for the case when the number of rejections of the hypothesis H_0 is maximal, was taken into the table.

We can see from the table that the new tests, which are based on data compression methods, can detect non-randomness quite efficiently.

Tab. 1: Number of files generated by PRNG RANDU and recognized as non-random for different tests.

Name of test / Length of file (in bits)	50 000	100 000	500 000	1 000 000
RAR	0	0	100	100
ARJ	0	0	99	100
Frequency	2	1	1	2
Block Frequency	1	2	1	1
Cumulative Sums	2	1	2	1
Runs	0	2	1	1
Longest Run of Ones	0	1	0	0
Rank	0	1	1	0
Discrete Fourier Transform	0	0	0	1
NonOverlapping Templates	–	–	–	2
Overlapping Templates	–	–	–	2
Universal Statistical	–	–	1	1
Approximate Entropy	1	2	2	7
Random Excursions	–	–	–	2
Random Excursions Variant	–	–	–	2
Serial	0	1	2	2
Lempel-Ziv Complexity	–	–	–	1
Linear Complexity	–	–	–	3

5 Appendix

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

Claim 2. *Let p and q be two probability distributions over some alphabet B. Then $\sum_{b\in B} p(b)\log\frac{p(b)}{q(b)} \geq 0$ with equality if and only if $p = q$.*

The following property of the empirical Shannon entropy will be used in proofs of the Theorem 2 and Theorem 3.

Lemma. *Let θ be a measure from $M_m(A), m \geq 0$, and $x_1 \dots x_t \in A^t$. Then*

$$\theta(x_1 \dots x_t) \leq \prod_{u\in A^m}\prod_{a\in A}(\nu^t(ua)/\bar{\nu}^t(u))^{\nu^t(ua)} \;=\; 2^{-(t-m)\,h^*_m(x_1\dots x_t)} \tag{10}$$

Proof of the Lemma. First we show that for any source $\theta^* \in M_0(A)$ and any word $x_1 \dots x_t \in A^t, t > 1$,

$$\theta^*(x_1 \dots x_t) = \prod_{a\in A}(\theta^*(a))^{\nu^t(a)} \leq \prod_{a\in A}(\nu^t(a)/t)^{\nu^t(a)} \tag{11}$$

Here the equality holds, because $\theta^* \in M_0(A)$. The inequality follows from the Claim 2. 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 (11). Now we 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,

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

Taking into account the inequality (11), we obtain

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

for any word u. So, from the last two inequalities we obtain the inequality (10). The equality in (10) follows from (5).

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 auxiliary set $\hat{C}_\alpha = \{u : -\log \pi(u) - (-\log \mu_\varphi(u)) > -\log \alpha\}$. We have $1 \ge \sum_{u \in \hat{C}_\alpha} \mu_\varphi(u) \ge \sum_{u \in \hat{C}_\alpha} \pi(u)/\alpha$ $= (1/\alpha)\pi(\hat{C}_\alpha)$. (Here the second inequality follows from the definition of $\hat{C}_\alpha$, whereas all others are obvious.) So, we obtain that $\pi(\hat{C}_\alpha) \le \alpha$. From definitions of $C_\alpha, \hat{C}_\alpha$ and (2) we immediately obtain that $\hat{C}_\alpha \supset C_\alpha$. Thus, $\pi(C_\alpha) \le \alpha$. 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 $\tau \ne \pi$. Our strategy is to show that

$$\lim_{t \to \infty} -\log \pi(x_1 \dots x_t) - |\varphi(x_1 \dots x_t)| = \infty \tag{12}$$

with probability 1 (according to the measure τ). First we represent (12) 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 (3) we obtain

$$-\log \pi(x_1 \dots x_t) - |\varphi(x_1 \dots x_t)| = t\,(\frac{1}{t} \log \frac{\tau(x_1 \dots x_t)}{\pi(x_1 \dots x_t)} + o(1)). \tag{13}$$

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, $\lim_{t\to\infty} -\log \tau(x_1 \dots x_t)/t$ exists (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:

$$h_\infty(\tau) \le -\sum_{v \in A^k} \tau(v) \sum_{a \in A} \tau(a/v) \log \tau(a/v).$$

(Here the right hand value is called $m-$ order conditional entropy). It will be convenient to represent both statements as follows:

$$\lim_{t \to \infty} -\log \tau(x_1 \dots x_t)/t \le -\sum_{v \in A^k} \tau(v) \sum_{a \in A} \tau(a/v) \log \tau(a/v) \tag{14}$$

for any $k \ge 0$ (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 $k \ge s$. 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, (14) and (13), we can see that

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

for any $k \geq s$. From this inequality and the Claim 2 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, (12) is true and the theorem is proven.

Proof of Theorem 2. It will be convenient to define two auxiliary measures on A^t as follows:

$$\pi_m(x_1...x_t) = \Delta\, 2^{-(\,t-m\,)\, h^*_m(x_1...x_t)}\,, \tag{15}$$

where $x_1...x_t \in A^t$ and $\Delta = (\sum_{x_1...x_t \in A^t} 2^{-t\, h^*_m(x_1...x_t)}\,)^{-1}$. From this definition and Lemma we can see that for any measure $\theta \in M_m(A)$ and any $x_1 \dots x_t \in A^t$,

$$\theta(x_1 \dots x_t) \leq \pi_m(x_1...x_t)/\Delta\,. \tag{16}$$

Let us denote the critical set of the test $\Upsilon^t_{\alpha,\,\psi,m}$ 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 the Claim 1 we can see that there exists such a measure μ_ψ that $-\log \mu_\psi(x_1...x_t) \leq |\psi(x_1...x_t)|$. We also define

$$\hat{C}_\alpha = \{x_1 \dots x_t :\ (t-m)\, h^*_m(x_1 \dots x_t) - (-\log \mu_\psi(x_1...x_t))\,) > \log(1/\alpha)\}. \tag{17}$$

From the definition of C_α and and the latest inequality we can see that $\hat{C}_\alpha \supset C_\alpha$.

From (16) and (17) we can see that for any measure $\theta \in M_m(A)$

$$\theta(C_\alpha) \leq \pi_m(C_\alpha)/\Delta\,. \tag{18}$$

From (17) and (15) we obtain

$$\hat{C}_\alpha = \{x_1 \dots x_t :\ 2^{(t-m)\, h^*_m(x_1...x_t)} > (\alpha\ \mu_\psi(x_1 \dots x_t))^{-1}\}$$
$$= \{x_1 \dots x_t :\ (\pi_m(x_1 \dots x_t)/\Delta)^{-1} > (\alpha\ \mu_\psi(x_1 \dots x_t))^{-1}\}\,.$$

Finally,

$$\hat{C}_\alpha = \{x_1 \dots x_t :\ \mu_\psi(x_1 \dots x_t) > \pi_m(x_1 \dots x_t)/(\alpha\, \Delta)\}. \tag{19}$$

The following chain of inequalities and equalities is valid:

$$1 \geq \sum_{x_1...x_t \in \hat{C}_\alpha} \mu_\psi(x_1 \dots x_t) \geq \sum_{x_1...x_t \in \hat{C}_\alpha} \pi_m(x_1 \dots x_t)/(\alpha\, \Delta) = \pi_m(\hat{C}_\alpha)/(\alpha\, \Delta) \geq \theta(\hat{C}_\alpha)\Delta/(\alpha\, \Delta) = \theta(C_\alpha)/\alpha.$$

(Here both equalities and the first inequality are obvious, the second and the third inequalities follow from (19) and (18), correspondingly.) So, we obtain that $\theta(\hat{C}_\alpha) \leq \alpha$ for any measure $\theta \in M_m(A)$. Taking into account that $\hat{C}_\alpha \supset C_\alpha$, where C_α is the critical set of the test, we can see that the probability of the First Type error is not greater than α. The first statement 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:

$$h_t(p) = -\sum_{x_1...x_t \in A^t} p(x_1...x_t) \sum_{a \in A} p(a/x_1...x_t) \log p(a/x_1...x_t), \tag{20}$$

where $p \in M_\infty(A)$. It is known that for any $p \in M_\infty(A)$ first, $\log|A| \geq h_0(p) \geq h_1(p) \geq ...$, second, there exists limit Shannon entropy $h_\infty(p) = \lim_{t\to\infty} h_t(p)$, third, $\lim_{t\to\infty} -t^{-1} \log p(x_1...x_t) = h_\infty(p)$ with probability 1 and, fourth, $h_m(p)$ is strictly greater than $h_\infty(p)$, if the memory of p is greater 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 (3), we obtain from the above described properties of the entropy that

$$\lim_{t\to\infty} t^{-1}|\psi(x_1...x_t)| = h_\infty(p) \tag{21}$$

with probability 1. It can be seen from (5) that h^*_m is an estimate for the $m-$order Shannon entropy (20). 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 (21) 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.

Proof of Theorem 3. Let C_α be a critical set of the test, i.e., by definition, $C_\alpha = \{(x_1,...,x_t) : (x_1,...,x_t) = (x_1^1, x_1^2, ...x_1^d), (x_2^1, x_2^2, ...x_2^d), \ldots, (x_t^1, x_t^2, ...x_t^d)$ & $\sum_{i=1}^d (t-m) h_m^*(x_1^i ... x_t^i) - |\varphi(x_1...x_t)| > \log(1/\alpha)\}$. According to the Claim 1, there exists a measure μ_φ, for which (2) is valid. Hence,

$$C_\alpha \subset C_\alpha^* \equiv \{(x_1,...,x_t) : \sum_{i=1}^d (t-m) h_m^*(x_1^i...x_t^i) - \log(1/\mu_\varphi(x_1,...,x_t) > \log(1/\alpha)\}. \quad (22)$$

Let θ be any measure from $M_m(A)$. Then, the following chain of inequalities and equalities is valid:

$$1 \geq \mu_\varphi(C_\alpha^*) \geq \alpha^{-1} \sum_{x_1,\ldots,x_t \in C_\alpha^*} \prod_{i=1}^d 2^{-(t-m)h_m^*(x_1^i...x_t^i)}.$$

Having taken into account Lemma, we obtain

$$1 \geq \mu_\varphi(C_\alpha^*) \geq \sum_{x_1,\ldots,x_t \in C_\alpha^*} \prod_{i=1}^d \mu^i(x_1^i...x_t^i).$$

It is supposed that H_0^{ind} is true and, hence, (7) is valid. So, from the latter inequalities we can see that $1 \geq \mu_\varphi(C_\alpha^*) \geq \sum_{x_1,\ldots,x_t \in C_\alpha^*} \mu(x_1,...,x_t)$. Taking into account that $\sum_{x_1,\ldots,x_t \in C_\alpha^*} \mu(x_1,...,x_t) = \mu(C_\alpha^*)$ and (22), we obtain that $\mu(C_\alpha) \leq \alpha$. So, the first statement of the theorem is proven.

We give a short scheme of the proof of the second statement of the theorem, because it is based on well-known facts of Information Theory. It is known that $h_m(\mu) - \sum_{i=1}^d h_m(\mu^i) = 0$ if H_0^{ind} is true and this difference is negative under H_1^{ind}. A universal code compresses a sequence till $t h_m(\mu)$ (Informally, it uses dependence for the better compression.) That is why the difference $t\ (h_m(\mu) - \sum_{i=1}^d h_m(\mu^i))$ goes to infinity, when t increases and, hence, H_0^{ind} will be rejected.

References

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

[2] Cilibrasi R., Vitanyi P.M.B. Clustering by Compression. *IEEE Transactions on Information Theory,* **51**(4) (2005),

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

[4] M. Drmota, H. Hwang, W. Szpankowski. Precise Average Redundancy of an Idealized Arithmetic Coding, In: *Data Compression Conference,* Snowbirds, (2002), 222-231.

[5] Dudewicz E.J. and Ralley T.G. *The Handbook of Random Number Generation and Testing With TESTRAND Computer Code,* v. 4 of American Series in Mathematical and Management Sciences. American Sciences Press, Inc., Columbus, Ohio, 1981.

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

[7] Ghoudi K., Kulperger R.J., Remillard B., A Nonparametric Test of Serial Independence for Time Series and Residuals. *Journal of Multivariate Analysis,* **79**(2), (2001), 191-218.

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

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

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

[11] Knuth D.E. *The art of computer programming.* Vol.2. Addison Wesley, 1981.

[12] Morvai G. , Yakowitz S.J., Algoet P.H. , Weakly convergent nonparametric forecasting of stationary time series. *IEEE Trans. Inform. Theory,* **43** (1997), 483 - 498.

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

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

[15] Rukhin A. and others. *A statistical test suite for random and pseudorandom number generators for cryptographic applications.* NIST Special Publication 800-22 (with revision dated May,15,2001). http://csrc.nist.gov/rng/SP800-22b.pdf

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

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

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

[19] Ryabko B., Astola J. Universal Codes as a Basis for Nonparametric Testing of Serial Independence for Time Series . *Journal of Statistical Planning and Inference.*(Submitted)

[20] Ryabko B. Ya., Monarev V.A. Using information theory approach to randomness testing. *Journal of Statistical Planning and Inference,* **133**(1)(2005), 95-110