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

Boris Ryabko, Jaakko Astola

To cite this version:

HAL Id: hal-01184215
https://hal.inria.fr/hal-01184215
Submitted on 13 Aug 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Application of data compression methods to hypothesis testing for ergodic and stationary processes

Boris Ryabko and Jaakko Astola

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 first problem is the identity testing, which is described as follows: a hypotheses \( H_0^{id} \) is that the source has a particular distribution \( \pi \) 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 \( p(x_{m+1} = (a_{i_1}, \ldots, a_{i_d})/x_1 \ldots x_m) = \prod_{j=1}^d p(x_{m+1} = a_{i_j}/x_1 \ldots x_m) \) for each \( (a_{i_1}, \ldots, a_{i_d}) \in \prod_{i=1}^d A_i \), where \( x_{m+1} = (x_{m+1}^1, \ldots, x_{m+1}^d) \). 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 \( H_0^{ind} \).

1 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 \( x_1 \ldots x_t \) 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 presented. All three problems are well known in mathematical statistics and there is an extensive literature dealing with them — see, for example, [7, 9].

## 2 Definitions and Preliminaries

First, we define stochastic processes (or sources of information). Consider an alphabet \( A \) 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 (\( A^* = \bigcup_{t=1}^{\infty} A^t \)). Let \( \mu \) be a source which generates letters from \( A \). Formally, \( \mu \) 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

\[
\mu(x_{t+1} = a_{i_1} | x_t = a_{i_2}, x_{t-1} = a_{i_3}, ... , x_{t-k+1} = a_{i_k+1}) = \mu(x_{t+1} = a_{i_1} | x_t = a_{i_2}, x_{t-1} = a_{i_3}, ... , x_{t-k+1} = a_{i_k+1})
\]

for all \( t \geq k \) and \( a_{i_1}, a_{i_2}, ... , a_{i_k+1} \in A \). By definition, \( M_0(A) \) is the set of all Bernoulli (or i.i.d.) sources over \( A \) and \( M_\infty(A) \) is the set of all finite-memory sources.

A data compression method (or code) \( \varphi \) is defined as a set of mappings \( \varphi_n : A^n \rightarrow \{0, 1\}^* \), \( n = 1, 2, ... \) 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 \( \varphi \) 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 \( \psi_1(a) = 0, \psi_1(b) = 0 \) obviously, is not uniquely decodable. It is well known that if a code \( \varphi \) is uniquely decodable then the lengths of the codewords satisfy the following inequality (Kraft inequality): \( \sum_{u \in A^n} 2^{-|\varphi_n(u)|} \leq 1 \), 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 \( \varphi \) be a uniquely decodable code over an alphabet \( A \). Then for any integer \( n \) there exists a measure \( \mu_\varphi \) on \( A^n \) such that

\[
|\varphi(u)| \geq -\log \mu_\varphi(u)
\]

for any \( u \) from \( A^n \).

(Here and below \( \log \equiv \log_2 \)) Obviously, Claim 1 is true for the measure

\[
\mu_\varphi(u) = 2^{-|\varphi(u)|} / \sum_{u \in A^n} 2^{-|\varphi(u)|}
\]

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 \( -\log p(x_1 \ldots x_t) \) bits and, on the other hand, for any source \( p \) there is no code \( \psi \) 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...
We will consider sources, which generate letters from an alphabet $A$. Now we consider the problem of the independence testing for Markovian sources. More precisely, in this

3.3 Independence Testing.

First, we give some additional definitions. Let $x$ be an $A$-valued stationary and ergodic process over $\mathbb{N}$. The hypothesis $H_0^{id}$ is accepted if

$$
-h(x_1) \leq -\log \pi(x_1) \leq -\log \alpha.
$$

Otherwise, $H_0^{id}$ is rejected. (Here $\pi$ 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 $\varphi$, the Type I error of the described test $\Gamma^{(n)}_{\pi, \alpha, \varphi}$ is not larger than $\alpha$. ii) If, in addition, $\pi$ is a finite-memory stationary and ergodic process over $A^\infty$ (i.e. $\pi \in M^*(A)$) and $\varphi$ 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 $v = v_1 \ldots v_k$, $k \leq t$, $v_i \in A$. Denote the rate of a word $v$ occurring in the sequence $x_1 x_2 \ldots x_k, x_2 x_3 \ldots x_{k+1}, \ldots, x_{t-k+1} \ldots x_t$ as $\nu^k(v)$. For example, if $x_1, \ldots, x_{10} = 0000100$ and $v = 00$, then $\nu^k(00) = 3$. Now we define for any $0 \leq k < t$ a so-called empirical Shannon entropy of order $k$ as follows:

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

where $\nu^k(v) = \sum_{a \in A} \nu^k(va)$. In particular, if $k = 0$, we obtain $h_0^*(x_1 \ldots x_t) = -\frac{1}{t} \sum_{a \in A} \nu^k(a) \log(\nu^k(a) / t)$.

Let, as before, $H_0^{SI}$ be that the source $\pi$ is Markovian with memory (or connectivity) not greater than $m$, ($m \geq 0$), and the alternative hypothesis $H_0^{SI}$ 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 $\psi$ be any code. By definition, the hypothesis $H_0^{SI}$ is accepted if

$$(t-m) h_0^*(x_1 \ldots x_t) - \log(1 / \alpha),$$

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

Theorem 2. i) For any distribution $\pi$ and any code $\psi$ the Type I error of the test $\Upsilon_{\alpha, \psi, m}$ is less than or equal to $\alpha$, $\alpha \in (0, 1)$ and, ii) if, in addition, $\pi$ is a stationary and ergodic process over $A^\infty$ and $\psi$ is a universal code, then the Type II error of the test $\Upsilon_{\alpha, \psi, m}$ 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 $m, m \geq 0$. We will consider sources, which generate letters from an alphabet $A = \prod_{i=1}^d A_i$, $d \geq 2$, and present each
generated letter \( x_i \) as the following string: \( x_i = (x^1_i, \ldots, x^d_i) \), where \( x^j_i \in A_j \). The hypothesis \( H^{\text{ind}}_0 \) is that a sequence \( x_1, \ldots, x_n \) 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 \( x_1, \ldots, x_m \in A^m \) the following equality is valid:

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

(7)

where, by definition,

\[
\mu^i(x^i_{m+1} = a_i/x_1, \ldots, x_m) = \sum_{b_{i-1} \in \prod_{j=1}^{i-1} A_j} \sum_{b_i \in \prod_{j=i+1}^d A_j} \mu(b_1, \ldots, b_{i-1}, a_i, b_{i+1}, \ldots, b_d)/x_1, \ldots, x_m).
\]

The hypothesis \( H^{\text{ind}}_1 \) 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 \( x_1, \ldots, x_m \in A^m \).

Let us describe a test for hypotheses \( H^{\text{ind}}_0 \) and \( H^{\text{ind}}_1 \). Let \( \varphi \) be any code. By definition, the hypothesis \( H^{\text{ind}}_0 \) is accepted if

\[
\sum_{i=1}^d (t - m)h^*_m(x^i_1, x^i_2, \ldots, x^i_d) - |\varphi(x_1, \ldots, x_t)| \leq \log(1/\alpha),
\]

(9)

where \( (x_1, \ldots, x_t) = (x^1_1, x^1_2, \ldots, x^1_d), (x^2_1, x^2_2, \ldots, x^2_d), \ldots, (x^t_1, x^t_2, \ldots, x^t_d) \) and \( \alpha \in (0, 1) \). Otherwise, \( H^{\text{ind}}_0 \) 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 \( h^*_m(x_1, \ldots, x_t) \) is its estimate. So, if \( H^{\text{ind}}_0 \) is true, the sum \( \sum_{i=1}^d (t - m)h^*_m(x^i_1, x^i_2, \ldots, x^i_d) \) is, on average, close to lower bound. Hence, if the length of a codeword of some code \( \varphi \) 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 \( \varphi \) 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, \( \varphi \) 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 \( H^{\text{ind}}_0 \) 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 \( 5 \times 10^4 \) 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 \( \Gamma^{(l)}_{\text{uniform}, \alpha, \varphi} \) 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 (\( H^{\text{ind}}_0 \)) 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 \( H^{\text{ind}}_0 \) 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.

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

## 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 \( \theta \) be a measure from \( M_m(A), m \geq 0 \), and \( x_1 \ldots x_t \in A^t \). Then

\[
\theta(x_1 \ldots x_t) \leq \prod_{u \in A^m} \prod_{a \in A} \left( \frac{\nu^t(ua)/\nu^t(u)}{\nu^t(ua)/\nu^t(u)} \right)^{\theta^t(ua)} = 2^{-(t-m)\sum_{u \in A^m} \log \frac{\nu^t(ua)/\nu^t(u)}{\nu^t(ua)/\nu^t(u)}} \tag{10}
\]

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

\[
\theta^*(x_1 \ldots x_t) = \prod_{a \in A} (\theta^*(a))^{\nu^t(a)} \leq \prod_{a \in A} \left( \frac{\nu^t(a)/t}{\theta^*(a)} \right)^{\nu^t(a)} \tag{11}
\]

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 \( \theta(x_1 \ldots x_t) \) as

\[
\theta(x_1 \ldots x_t) = \theta(x_1 \ldots x_m) \prod_{u \in A^m} \prod_{a \in A} \theta(a/u)^{\nu^t(ua)},
\]
where $\theta(x_1 \ldots x_m)$ is the limit probability of the word $x_1 \ldots x_m$. Hence,

$$\theta(x_1 \ldots x_I) \leq \prod_{u \in A^m} \prod_{a \in A} \theta(a/u)^{\nu'(ua)}.$$

Taking into account the inequality (11), we obtain

$$\prod_{a \in A} \theta(a/u)^{\nu'(ua)} \leq \prod_{a \in A} \left(\nu'(ua)/\nu'(u)\right)^{\nu'(ua)}$$

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 $C_\alpha$ be a critical set of the test $\Gamma^{(n)}_{x,\alpha,\phi}$, i.e., by definition, $C_\alpha = \{u : u \in A^t \& - \log \pi(u) - |\varphi(u)| > - \log \alpha \}$. Let $\mu_\phi$ be a measure for which the claim 1 is true. We define an auxiliary set $\hat{C}_\alpha = \{u : - \log \pi(u) - ( - \log \mu_\phi(u)) > - \log \alpha \}$. We have $1 \geq \sum_{u \in C_\alpha} \mu_\phi(u) \geq \sum_{u \in \hat{C}_\alpha} \pi(u)/\alpha = (1/\alpha)\pi(\hat{C}_\alpha)$. (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 $C_\alpha \supset \hat{C}_\alpha$. 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 $H_{id}^1$ is true. That is, the sequence $x_1 \ldots x_t$ is generated by some stationary and ergodic source $\tau$ and $\tau \neq \pi$. Our strategy is to show that

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

(12)

with probability 1 (according to the measure $\tau$). First we represent (12) as

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

From this equality and the property of a universal code (3) we obtain

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

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

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

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

$$\lim_{t \to \infty} - \log \tau(x_1 \ldots x_I)/t \leq - \sum_{v \in A^k} \tau(v) \sum_{a \in A} \tau(a/v) \log \tau(a/v)$$

(14)

for any $k \geq 0$ (with probability 1). It is supposed that the process $\pi$ 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:

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

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

$$\lim_{t \to \infty} t^{-1} \sum_{i=k+1}^t \log \pi(x_i/x_{i-k} \ldots x_1),$$

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

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

Taking into account this equality, (14) and (13), we can see that

$$- \log \pi(x_1 \ldots x_I) - |\varphi(x_1 \ldots x_I)| \geq t \left(\sum_{v \in A^k} \tau(v) \sum_{a \in A} \tau(a/v) \log(\tau(a/v)/\pi(a/v))\right) + o(t)$$

for any word $u$. So, from the last two inequalities we obtain the inequality (10). The equality in (10) follows from (5).
for any \( k \geq s \). From this inequality and the Claim 2 we can obtain that
\[-\log \pi(x_1 \ldots x_t) - |\varphi(x_1 \ldots x_t)| \geq c t + o(t), \]
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:
\[
\pi_m(x_1 \ldots x_t) = \Delta 2^{-(t-m)} h_m^*(x_1 \ldots x_t),
\]
where \( x_1 \ldots x_t \in A^t \) and \( \Delta = (\sum_{x_1 \ldots x_t \in A^t} 2^{-(t-m)} h_m^*(x_1 \ldots x_t))^{-1} \). From this definition and Lemma we can see that for any measure \( \theta \in M_m(A) \) and any \( x_1 \ldots x_t \in A^t \),
\[
\theta(x_1 \ldots x_t) \leq \pi_m(x_1 \ldots x_t)/\Delta.
\]
Let us denote the critical set of the test \( \Theta_m^1 \) as \( C_\alpha \), i.e., by definition, \( C_\alpha = \{x_1 \ldots x_t : (t-m) h_m^*(x_1 \ldots x_t) - |\varphi(x_1 \ldots x_t)| > \log(1/\alpha)\} \). From the Claim 1 we can see that there exists such a measure \( \mu_\varphi \) that
\[-\log \mu_\varphi(x_1 \ldots x_t) \leq |\varphi(x_1 \ldots x_t)|. \]
We also define
\[
\hat{C}_\alpha = \{x_1 \ldots x_t : (t-m) h_m^*(x_1 \ldots x_t) - (-\log \mu_\varphi(x_1 \ldots x_t)) > \log(1/\alpha)\}.
\]
From the definition of \( C_\alpha \) 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 \( \theta \in M_m(A) \)
\[
\theta(C_\alpha) \leq \pi_m(C_\alpha)/\Delta.
\]
From (17) and (15) we obtain
\[
\hat{C}_\alpha = \{x_1 \ldots x_t : 2^{(t-m)} h_m^*(x_1 \ldots x_t) > (\alpha \mu_\varphi(x_1 \ldots x_t))^{-1}\} = \{x_1 \ldots x_t : (\pi_m(x_1 \ldots x_t)/\Delta)^{-1} > (\alpha \mu_\varphi(x_1 \ldots x_t))^{-1}\}.
\]
Finally,
\[
\hat{C}_\alpha = \{x_1 \ldots x_t : \mu_\varphi(x_1 \ldots x_t) > \pi_m(x_1 \ldots x_t)/(\alpha \Delta)\}.
\]
The following chain of inequalities and equalities is valid:
\[
1 \geq \sum_{x_1 \ldots x_t \in \hat{C}_\alpha} \mu_\varphi(x_1 \ldots x_t) \geq \sum_{x_1 \ldots x_t \in \hat{C}_\alpha} \pi_m(x_1 \ldots x_t)/(\alpha \Delta) = \pi_m(\hat{C}_\alpha)/(\alpha \Delta) = \theta(\hat{C}_\alpha)/(\alpha \Delta) = \theta(C_\alpha)/\alpha.
\]
(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 \( \theta(C_\alpha) \leq \alpha \) for any measure \( \theta \in M_m(A) \). Taking into account that \( \hat{C}_\alpha \supset C_\alpha \), where \( C_\alpha \) is the critical set of the test, we can see that the probability of the First Type error is not greater than \( \alpha \). 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:
\[
h_t(p) = \sum_{x_1 \ldots x_t \in A^t} p(x_1 \ldots x_t) \sum_{a \in A} p(a|x_1 \ldots x_t) \log p(a|x_1 \ldots x_t),
\]
where \( p \in M_{\infty}(A) \). It is known that for any \( p \in M_{\infty}(A) \) first, \( \log |A| \geq h_0(p) \geq h_1(p) \geq \ldots \), second, there exists limit Shannon entropy \( h_\infty(p) = \lim_{t \to \infty} h_t(p) \), third, \( \lim_{t \to \infty} -t^{-1} \log p(x_1 \ldots x_t) = h_\infty(p) \) 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
\[
\lim_{t \to \infty} -t^{-1}|\varphi(x_1 \ldots x_t)| = h_\infty(p)
\]
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
\[
\lim_{t \to \infty} h_m^*(x_1 \ldots x_t) = h_m(p) \text{ with probability 1; see [1, 6]}
\]
Having taken into account that \( h_m(p) > h_\infty(p) \) and (21) we obtain from the last equality that
\[
\lim_{t \to \infty} ((t-m) h_m^*(x_1 \ldots x_t) - |\varphi(x_1 \ldots x_t)|) = \infty.
\]
This proves the second statement of the theorem.
Proof of Theorem 3. Let $C_\alpha$ be a critical set of the test, i.e., by definition, $C_\alpha = \{(x_1, ..., x_t) : (x_1, ..., x_t) = (x_1^i, x_1^2, ..., x_t^i), (x_2^i, x_2^2, ..., x_t^i), ..., (x_t^i, x_t^2, ..., x_t^i) \} \subset \{h_m(x_1^1, ..., x_t^1) > \log(1/\alpha)\}$. According to the Claim 1, there exists a measure $\mu_\varphi$, for which (2) is valid. Hence, $C_\alpha \subset C_\alpha^* \equiv \{(x_1, ..., x_t) : \sum_{i=1}^d (t-m)h_m(x_1^i, ..., x_t^i) - \log(1/\mu_\varphi(x_1, ..., x_t) > \log(1/\alpha)\}$. (22)

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

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

Having taken into account Lemma, we obtain

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

It is supposed that $H_0^{ind}$ is true and, hence, (7) is valid. So, from the latter inequalities we can see that $1 \geq \mu_\varphi(C_\alpha^*) \geq \sum_{x_1, ..., x_t \in C_\alpha^*} \mu_\varphi(x_1, ..., x_t) = \mu(C_\alpha^*)$ 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 $\theta_m(\mu)$ (Informally, it uses dependence for the better compression.) That is why the difference $t(\theta_m(\mu) - \sum_{i=1}^d h_m(\mu^i))$ goes to infinity, when $t$ increases and, hence, $H_0^{ind}$ will be rejected.

References


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


