Available online at www.sciencedirect.com
ScienceDirect
Statistical Methodology 3 (2006) 375–397
Statistical Methodology
Official Journal of the International Indian Statistical Association
www.elsevier.com/locate/stamet

Universal codes as a basis for time series testing

Boris Ryabko[a,*], Jaakko Astola[b]

[a] *Institute of Computational Technology of Siberian Branch of Russian Academy of Science, Novosibirsk, Russian Federation*
[b] *Tampere University of Technology, FI-33101, Tampere, Finland*

Received 13 April 2005; received in revised form 19 August 2005; accepted 7 October 2005

Abstract

We suggest a new approach to hypothesis testing for ergodic and stationary processes. In contrast to standard 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 apply this approach to the following four problems: goodness-of-fit testing (or identity testing), testing for independence, testing of serial independence and homogeneity testing and suggest nonparametric statistical tests for these problems. It is important to note that practically used so-called archivers can be used for suggested testing.

MSC: 60G10; 60J10; 62M02; 62M07; 94A29

Keywords: Universal coding; Data compression; Hypothesis testing; Nonparametric testing; Shannon entropy; Stationary and ergodic source

1. Introduction

Since Claude Shannon published his famous paper "A mathematical theory of communication" [36], the ideas and results of Information Theory have begun to play an important role in cryptography [21,37], mathematical statistics [1,5,20,26], ergodic theory [1,2, 38] and many other fields [3,4,33] which are far from telecommunication. The theory of universal coding, which is a part of Information Theory, also has been efficiently applied to many fields

* Corresponding address: Siberian State University of Telecommunication and Computer Science, Kirov Street, 86, 630102 Novosibirsk, Russian Federation. Tel.: +7 383 2 284938; fax: +7 383 2 669343.
E-mail address: boris@ryabko.net (B. Ryabko).

1572-3127/$ - see front matter
doi:10.1016/j.stamet.2005.10.004

since its discovery in [10,18]. Thus, application of results of universal coding, initiated in [29], created a new approach to prediction [15,23,24].

In this paper we suggest a new approach to hypothesis testing, which is based on ideas of universal coding. We would like to emphasize that, on the one hand, the problem of hypothesis testing is considered in the framework of classical mathematical statistics and, on the other hand, everyday methods of data compression (or archivers) can be used as a tool for testing. It is important to note that the modern archivers are based on deep theoretical results of the source coding theory (see, e.g., [8,16,19,25,35]) and have shown their high efficiency in practice as compressors of texts, DNA sequences and many other types of real data. In fact, universal codes and archivers can find latent regularities of many kinds, that is why they look like a promising tool for hypothesis testing.

1.1. The main idea of the suggested approach

Let us describe the main idea of the suggested approach using one particular problem of hypothesis testing which is conceptually simple and yet is important in practice. Namely, we consider a null hypothesis H_0 that a given bit sequence $x_1 \dots x_t$ is generated by a Bernoulli source with equal probabilities of 0 and 1 and the alternative hypothesis H_1 that the sequence is generated by a stationary and ergodic source, which differs from the source under H_0. This problem is considered in [32] and is a particular case of the goodness-of-fit testing (or identity testing) described below, that is why we give an informal solution only. Let φ be a universal code, $\varphi(x_1 \dots x_t)$ be the encoded sequence, $l_\varphi(x_1 \dots x_t)$ be the length of the word $\varphi(x_1 \dots x_t)$ and α be the required level of significance. Intuition suggests that the sequence cannot be compressed if H_0 is true, and vice versa, if the sequence can be compressed H_0 should be rejected. The corresponding formal test is as follows: if $(t - l_\varphi(x_1 \dots x_t)) > \log(1/\alpha)$, then H_0 should be rejected. (Here and below $\log \equiv \log_2$.) It will be proven below that the Type I error of this test is equal to or less than α for any (uniquely decodable) code φ, whereas the Type II error goes to 0 for any universal code φ, when the sequence length t grows.

Let us look at the described test in more detail. It is well known that the average codeword length of any code is not less than the sequence length t, if H_0 is true. Hence, if we define the codeword length of the best code as $l_{H_0}(x_1 \dots x_t)$, we can see that $l_{H_0}(x_1 \dots x_t) = t$. Now the scheme of the suggested test can be described as follows: *if* $l_{H_0}(x_1 \dots x_t) - l_\varphi(x_1 \dots x_t) \leq \log(1/\alpha)$ *then* H_0, *otherwise* H_1. We will apply this scheme to all considered statistical problems, sometimes replacing the length $l_{H_0}(x_1 \dots x_t)$ with its lower bound (as a rule, such a lower bound will be based on so-called empirical Shannon entropy).

1.2. Description of considered problems

We consider a stationary and ergodic source (or process), which generates elements from some set (or alphabet) A (which can be either finite or infinite) and four problems of statistical testing.

The first problem is the goodness-of-fit testing (or identity testing), which is described as follows: a hypothesis H_0^{id} is that the source has a particular distribution π and the alternative hypothesis H_1^{id} is 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 equals $\{0, 1\}$ 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, was mentioned in Introduction.

The second problem is a generalization of the problem of nonparametric testing for serial independence of time series. More precisely, we consider the two following hypotheses: H_0^{SI} is that the source is Markovian of order not larger than m, $(m \geq 0)$, and the alternative hypothesis H_1^{SI} is that the sequence is generated by a stationary and ergodic source which differs from the source under H_0^{SI}. In particular, if $m = 0$, this is the problem of testing for independence of time series.

The third problem is the independence testing. In this case it is assumed that the source is Markovian, whose order is not larger than m, $(m \geq 0)$, and the source alphabet can be presented as a product of $d, d \geq 2$, 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 $p(x_{m+1} = (a_{i_1}, \ldots, a_{i_d}) \mid x_1 \ldots x_m) = \prod_{j=1}^{d} p(x_{m+1}^{(j)} = a_{i_j} \mid x_1 \ldots x_m)$ 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 of order not larger than m, $(m \geq 0)$, which differs from the source under H_0^{ind}.

In all three cases the testing should be based either on one sample $x_1 \ldots x_t$ or on several (l) independent samples $x^1 = x_1^1 \ldots x_{t_1}^1, \ldots x^l = x_1^l \ldots x_{t_l}^l$ generated by the source.[1]

The fourth problem is the homogeneity testing. There are r samples $x_1^1 \ldots x_{t_1}^1$, $x_1^2 \ldots x_{t_2}^2, \ldots, x_1^r \ldots x_{t_r}^r$ and it is assumed that they are generated by Markovian sources, whose orders are not larger than m, $(m \geq 0)$. The main hypothesis H_0^{hom} is that all samples are generated by one source, whereas the alternative hypothesis H_1^{hom} is that at least two samples are generated by different sources.

All four problems are well known in mathematical statistics and there is an extensive literature dealing with their nonparametric testing, see for review, for example, [12,14].

1.3. Main results

We suggest statistical tests for all problems such that the Type I error is less than or equal to a given α and the Type II error goes to zero, when the sample size grows. However, there are some additional restrictions mainly concerned with the case of infinite source alphabet. For this case all tests are described for memoryless (or i.i.d.) sources only. It is important to note that the suggested tests are based on universal codes (and closely connected universal predictors), but the Type I error is less than or equal to a given α for any code and, in particular, it is true for practically used methods of data compression (or archivers), that is why they can be used as a basis for the tests.

1.4. Outline of the paper

The next section contains some necessary facts and definitions. Sections 3 and 4 are devoted to a description of the tests for the cases where alphabets are finite and infinite, respectively. Some experimental results and simulation studies are given in Section 5.

We give a description of one particular universal code in Appendix A, because universal codes play a key role in this paper, but information about them is spread between numerous papers and they are not widely presented in statistical literature (in spite of the fact that universal codes have found different applications to some classical problems of mathematical statistics, see,

[1] For a case of one sample and a finite alphabet A some of these problems were considered by the authors in [31] and reports submitted to conferences.

e.g., [5]). Besides, the universal code described in Appendix A is used for simulation study of serial independence testing in part 5. (On the other hand, this paper focuses on hypothesis testing, that is why the description of the universal codes and ideas behind them are put in the appendix.)

The conclusion is intended to clarify the connection of the suggested approach and briefly describe some possible generalizations of the described tests. All proofs are given in Appendix B.

2. Definitions and auxiliary results

2.1. Stochastic processes and the Shannon entropy

Now we briefly describe stochastic processes (or sources of information). Consider an alphabet A, which can be either finite or infinite, 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 ($A^* = \bigcup_{i=1}^{\infty} A^i$). By $M_\infty(A)$ we denote the set of all stationary and ergodic sources, which generate letters from A; see for definition, e.g., [2,11] and let $M_0(A) \subset M_\infty(A)$ be the set of all i.i.d. processes. Let $M_m(A) \subset M_\infty(A)$ be the set of Markov sources of order (or with memory, or connectivity) not larger than m, $m \geq 0$. In the case of a finite alphabet A Markov processes will play a key role in this paper, that is why we give a formal definition. By definition $\mu \in M_m(A)$ if

$$\begin{aligned}&\mu(x_{t+1} = a_{i_1} \mid x_t = a_{i_2}, x_{t-1} = a_{i_3}, \ldots, x_{t-m+1} = a_{i_{m+1}}, \ldots)\\ &\quad = \mu(x_{t+1} = a_{i_1} \mid x_t = a_{i_2}, x_{t-1} = a_{i_3}, \ldots x_{t-m+1} = a_{i_{m+1}})\end{aligned} \tag{1}$$

for all $t \geq m$ and $a_{i_1}, a_{i_2}, \ldots \in A$. Let $M^*(A) = \bigcup_{i=0}^{\infty} M_i(A)$ be the set of all finite-order sources.

Let τ be a stationary and ergodic source generating letters from a finite alphabet A. The m-order (conditional) Shannon entropy and the limit Shannon entropy are defined as follows:

$$h_m(\tau) = \sum_{v \in A^m} \tau(v) \sum_{a \in A} \tau(a \mid v) \log \tau(a \mid v), \qquad h_\infty(\tau) = \lim_{m \to \infty} h_m(\tau). \tag{2}$$

It is also known that for any m

$$h_\infty(\tau) \leq h_m(\tau), \tag{3}$$

see [2,11]. The well known Shannon–MacMillan–Breiman theorem states that

$$\lim_{t \to \infty} -\log \tau(x_1 \ldots x_t)/t = h_\infty(\tau) \tag{4}$$

with probability 1, see [2,11].

Let $v = v_1 \ldots v_k$ and $x = x_1 x_2 \ldots x_t$ be words from A^*. Denote the rate of a word v occurring in the sequence $x = x_1 x_2 \ldots x_k$, $x_2 x_3 \ldots x_{k+1}$, $x_3 x_4 \ldots x_{k+2}$, $\ldots$, $x_{t-k+1} \ldots x_t$ as $\nu_x(v)$. For example, if $x = 000100$ and $v = 00$, then $\nu_x(00) = 3$. For any $0 \leq k < t$ the empirical Shannon entropy of order k is defined as follows:

$$h_k^*(x) = -\sum_{v \in A^k} \frac{\bar{\nu}_x(v)}{(t-k)} \sum_{a \in A} \frac{\nu_x(va)}{\bar{\nu}_x(v)} \log \frac{\nu_x(va)}{\bar{\nu}_x(v)}, \tag{5}$$

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

We extend these definitions to a case where a sample is presented as several (independent) sequences $x^1 = x_1^1 \ldots x_{t_1}^1$, $x^2 = x_1^2 \ldots x_{t_2}^2$, $\ldots$, $x^r = x_1^r \ldots x_{t_r}^r$ generated by a source. (The point

is that we cannot simply combine all samples into one, if the source is not i.i.d.) We denote this sample by $x^1 \diamond x^2 \diamond \cdots \diamond x^r$ and define $t = \sum_{i=1}^r t_i$, $\nu_{x^1 \diamond x^2 \diamond \cdots \diamond x^r}(v) = \sum_{i=1}^r \nu_{x^i}(v)$. For example, if $x^1 = 0010$, $x^2 = 011$, then $\nu_{x^1 \diamond x^2}(00) = 1$. Analogously to (5),

$$h_k^*(x^1 \diamond x^2 \diamond \cdots \diamond x^r) = -\sum_{v \in A^k} \frac{\bar{\nu}_{x^1 \diamond \cdots \diamond x^r}(v)}{(t - kr)} \sum_{a \in A} \frac{\nu_{x^1 \diamond \cdots \diamond x^r}(va)}{\bar{\nu}_{x^1 \diamond \cdots \diamond x^r}(v)} \log \frac{\nu_{x^1 \diamond \cdots \diamond x^r}(va)}{\bar{\nu}_{x^1 \diamond \cdots \diamond x^r}(v)}, \tag{6}$$

where $\bar{\nu}_{x^1 \diamond \cdots \diamond x^r}(v) = \sum_{a \in A} \nu_{x^1 \diamond \cdots \diamond x^r}(va)$.

For any sequence of words $x^1 = x_1^1 \ldots x_{t_1}^1$, $x^2 = x_1^2 \ldots x_{t_2}^2, \ldots, x^r = x_1^r \ldots x_{t_r}^r$ from A^* and any measure θ we define $\theta(x^1 \diamond x^2 \diamond \cdots \diamond x^r) = \prod_{i=1}^r \theta(x^i)$.

We will use the following well known inequality, whose proof can be found in [11]:

For any two probability distributions p and q over some alphabet B the following inequality

$$\sum_{b \in B} p(b) \log \frac{p(b)}{q(b)} \geq 0 \tag{7}$$

is valid with equality if and only if $p = q$.

The value $\sum_{b \in B} p(b) \log \frac{p(b)}{q(b)}$ is often called Kullback–Leibler divergence.

The following property of the empirical Shannon entropy will be used later.

Lemma. *Let θ be a measure from $M_m(A)$, $m \geq 0$, and $x^1, \ldots, x^r$ be words from A^*, whose lengths are not less than m. Then*

$$\theta(x^1 \diamond \cdots \diamond x^r) \leq 2^{-(t - rm) h_m^*(x^1 \diamond \cdots \diamond x^r)}. \tag{8}$$

2.2. Codes

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 $x, y \in A^n$ $\varphi_n(x) \neq \varphi_n(y)$. It is also required that each sequence $\varphi_n(u_1)\varphi_n(u_2) \ldots \varphi_n(u_r)$, $r \geq 1$, of encoded words from the set A^n, $n \geq 1$, could be uniquely decoded into $u_1 u_2 \ldots u_r$. 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): $\sum_{u \in A^n} 2^{-|\varphi_n(u)|} \leq 1$, see, e.g., [11]. (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:

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

$$-\log \mu_\varphi(u) \leq |\varphi(u)| \tag{9}$$

for any u from A^n.

It is easy to see that it 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".

We suppose that any code is defined for each sequence of words $x^1 \diamond x^2 \diamond \cdots \diamond x^l$. (For example, any code φ can be extended to this case as follows: $\varphi(x^1 \diamond x^2 \diamond \cdots \diamond x^l) = \varphi(x^1)\varphi(x^2) \ldots \varphi(x^l)$.)

There exist so-called universal codes. To introduce these codes we first recall that (as it is known in Information Theory) sequences $x_1 \ldots x_t$, generated by a source p, can be "compressed" up to the length $-\log p(x_1 \ldots x_t)$ bits; on the other hand, for any source p there is no code ψ

for which the average codeword length $\sum_{u\in A^t} p(u)|\psi(u)|$ is less than $-\sum_{u\in A^t} p(u)\log p(u)$. Universal codes can reach the lower bound $-\log p(x_1\dots x_t)$ asymptotically for any stationary and ergodic source p with probability 1.

A formal definition is as follows: a code φ is universal if for any stationary and ergodic source p

$$\lim_{t\to\infty} t^{-1}(-\log p(x_1\dots x_t) - |\varphi(x_1\dots x_t)|) = 0 \tag{10}$$

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 [28], see also [29], and now there are many efficient universal codes and universal predictors connected with them, see [13,15,24,25,30,35].

3. Tests for a finite alphabet

3.1. Goodness-of-fit testing or identity testing

Now we consider the problem of testing the hypothesis H_0^{id} that the source has a particular distribution π, $\pi \in M_\infty(A)$, against H_1^{id} that the source is stationary and ergodic and differs from π. Let the required level of significance (or the Type I error) be α, $\alpha \in (0,1)$. 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 $\bar{x}$ by a code φ. If the length of the codeword $|\varphi(\bar{x})|$ is significantly less than the value $-\log\pi(\bar{x})$, then H_0^{id} should be rejected. The key observation is that the probability of all rejected samples is quite small for any φ, that is why the Type I error can be made small. The formal description of the test is as follows:

Let there be a sample $\bar{x}$ presented by sequences $x^1 = x_1^1\dots x_{t_1}^1, \dots, x^l = x_1^l\dots x_{t_l}^l$, generated independently by a source. The hypothesis H_0^{id} is accepted if

$$-\log\pi(\bar{x}) - |\varphi(\bar{x})| \le -\log\alpha. \tag{11}$$

Otherwise, H_0^{id} is rejected. We denote this test by $T_\varphi^{id}(A,\alpha)$.

Theorem 1. (i) *For each distribution π, $\alpha \in (0,1)$ and a code φ, the Type I error of the described test $T_\varphi^{id}(A,\alpha)$ is not larger than α and* (ii) *if, in addition, π is a finite-order stationary and ergodic process over A^∞ (i.e. $\pi \in M^*(A)$), φ is a universal code then the Type II error of the test $T_\varphi^{id}(A,\alpha)$ goes to 0 as the sample size t ($t = \sum_{i=1}^l t_i$) tends to infinity.*

3.2. Testing of serial independence

Let there be a sample $\bar{x}$ presented by sequences $x^1 = x_1^1\dots x_{t_1}^1, \dots, x^l = x_1^l\dots x_{t_l}^l$, generated independently by a (unknown) source and let $t = \sum_{i=1}^l t_i$. The main hypothesis H_0^{SI} is that the source is Markovian, whose order is not greater than m, ($m \ge 0$), and the alternative hypothesis H_1^{SI} is that the sample $\bar{x}$ is generated by a stationary and ergodic source whose order is greater than m (i.e. the source belongs to $M_\infty(A) \setminus M_m(A)$). The suggested test is as follows.

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

$$(t - ml)h_m^*(\bar{x}) - |\varphi(\bar{x})| \le \log(1/\alpha), \tag{12}$$

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

Theorem 2. (i) *For any code φ the Type I error of the test $T^{SI}_{\varphi}(A,\alpha)$ is less than or equal to $\alpha, \alpha \in (0,1)$ and,* (ii) *if, in addition, φ is a universal code and the sample size t tends to infinity, then the Type II error goes to* 0.

3.3. Independence testing

Now we consider the problem of the independence testing for Markovian sources. It is supposed that the source alphabet A is the Cartesian product of d alphabets $A_1, \dots, A_d$, i.e. $A = \prod_{i=1}^{d} A_i$, $d \geq 2$ and it is known a priori that the source belongs to $M_m(A)$ for some known m, $m \geq 0$. We present each letter x as $x = (x^{(1)}, \dots, x^{(d)})$, where $x^{(j)} \in A_j$. The hypothesis H_0^{ind} is that $\mu \in M_m(A)$ is such a source that for each $a = (a^{(1)}, \dots, a^{(d)}) \in \prod_{i=1}^{d} A_i$ and each $x_1 \dots x_m \in A^m$ the following equality is valid:

$$\mu(x_{m+1} = (a^{(1)}, \dots, a^{(d)}) | x_1 \dots x_m) = \prod_{i=1}^{d} \mu^{(i)}(x^{(i)}_{m+1} = a^{(i)} | x_1 \dots x_m), \tag{13}$$

where, by definition,

$$\mu^{(i)}(x^{(i)}_{m+1} = a | x_1 \dots x_m) = \sum_{b_1,\dots,b_{i-1} \in \prod_{j=1}^{i-1} A_j} \; \sum_{b_{i+1},\dots,b_d \in \prod_{j=i+1}^{d} A_j} \mu(x_{m+1} = (b_1, \dots, b_{i-1}, a, b_{i+1}, \dots, b_d) | x_1 \dots x_m). \tag{14}$$

The hypothesis H_1^{ind} is that Eq. (13) is not valid at least for one $(a^{(1)}, \dots, a^{(d)}) \in \prod_{i=1}^{d} A_i$ and $x_1 \dots x_m \in A^m$.

Let us describe the test for hypotheses H_0^{ind} and H_1^{ind}. Suppose that there is a sample $\bar{x}$ presented as sequences $x^1 = x^1_1 \dots x^1_{t_1}, \dots, x^l = x^l_1 \dots x^l_{t_l}$, generated independently by a source, where, in turn, any $x^j_i = (x^{j(1)}_i, \dots, x^{j(d)}_i)$. We define $t = \sum_{i=1}^{l} t_i$ and $\bar{x}^{(k)} = x^{1(k)}_1 \dots x^{1(k)}_{t_1} \diamond \dots \diamond x^{l(k)}_1 \dots x^{l(k)}_{t_l}$ for $k = 1, 2, \dots, d$.

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

$$\sum_{k=1}^{d} (t - ml) h^*_m(\bar{x}^{(k)}) - |\varphi(\bar{x})| \leq \log(1/\alpha), \tag{15}$$

$\alpha \in (0,1)$. *Otherwise, H_0^{ind} is rejected*. We denote this test by $T^{ind}_{\varphi}(A,\alpha)$. 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 $h^*_m(\bar{x}^{(k)})$ is its estimate. So, if H_0^{ind} is true, the sum $\sum_{k=1}^{d} (t - ml) h^*_m(\bar{x}^{(k)})$ is, on average, close to the 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 code φ the Type I error of the test $T^{ind}_{\varphi}(A,\alpha)$ is less than or equal to $\alpha, \alpha \in (0,1)$, and* (ii) *if, in addition, φ is a universal code and t tends to infinity, then the Type II error of the test $T^{ind}_{\varphi}(A,\alpha)$ goes to* 0.

3.4. Homogeneity testing

Let there be r samples $x^1 = x_1^1 \dots x_{t_1}^1$, $x^2 = x_1^2 \dots x_{t_2}^2, \dots, x^r = x_1^r \dots x_{t_r}^r$, $(r \geq 2)$, and it is assumed that they are generated by Markovian sources, whose orders are not larger than m, $(m \geq 0)$ and m is known a priori (i.e. the sources belong to $M_m(A)$). The null hypothesis H_0^{hom} is that all samples are generated by one source, whereas the alternative hypothesis H_1^{hom} is that at least two samples are generated by different sources.

Let us describe the test for hypotheses H_0^{hom} and H_1^{hom}. *Let φ be any code, $t = \sum_{i=1}^r t_i$ and $\alpha \in (0, 1)$. By definition, the hypothesis H_0^{hom} is accepted if*

$$(t - mr)h_m^*(x^1 \diamond x^2 \diamond \dots \diamond x^r) - \sum_{i=1}^{r} |\varphi(x^i)| \leq \log(1/\alpha). \tag{16}$$

Otherwise, H_0^{hom} is rejected. We denote this test by $T_\varphi^{hom}(A, \alpha)$.

Theorem 4. (i) *For any code φ the Type I error of the test $T_\varphi^{hom}(A, \alpha)$ is less than or equal to $\alpha, \alpha \in (0, 1)$ and* (ii) *if, in addition, φ is a universal code and the sample size t goes to infinity in such a way that there exists a positive constant c for which*

$$c < t_j/t \tag{17}$$

for each j, then the Type II error of the test $T_\varphi^{hom}(A, \alpha)$ goes to 0.

Let us give some comments concerning the constant c. In fact, the existence of such a constant means that all samples are present and grow. Otherwise, some samples could have a negligible length, say, 1 letter and, obviously, it would be difficult to build a reasonable test for such a case.

The suggested test can be extended for a case where it is known beforehand that some sequences (from $x^1, x^2, \dots, x^r$) were generated by the same source. In this case the same test can be applied, but condition (ii) can be weakened as follows: for each source the inequality (17) must be valid for at least one sample.

4. Infinite alphabet

In this part we consider the case where the source alphabet A is infinite, say, a part of R^n. Our strategy is to use finite partitions of A and to consider hypotheses corresponding to the partitions. The main problem of this approach is as follows: if someone combines letters (or states) of a Markov chain, the chain order (or memory) can increase. For example, if an alphabet contains three letters, there exists a Markov chain of order one, such that combining two letters into one subset transfers the chain into a process with infinite memory. On the other hand, the main part of the results described above is valid for finite-order processes. That is why in this part we will consider i.i.d. processes only (i.e. processes from $M_0(A)$).

In order to avoid numerous repetitions, we will consider a general scheme, which can be applied to all tests using notations $H_0^\aleph$, $H_1^\aleph$ and $T_\varphi^\aleph(A, \alpha)$, where $\aleph$ is an abbreviation of one of the described tests (i.e. *id*, *SI*, *ind* and *hom*).

Let us give some definitions. Let $\Lambda = \lambda_1, \dots, \lambda_s$ be a finite (measurable) partition of A and let $\Lambda(x)$ be an element of the partition Λ, which contains $x \in A$. For any process π we define a process π_Λ over a new alphabet Λ by equation

$$\pi_\Lambda(\lambda_{i_1} \dots \lambda_{i_k}) = \pi(x_1 \in \lambda_{i_1}, \dots, x_k \in \lambda_{i_k}),$$

where $x_1 \dots x_k \in A^k$. (Such partitions are widely used in information theory; see, e.g., [6,7,11] for a detailed description.)

We will consider an infinite sequence of partitions $\hat{\Lambda} = \Lambda_1, \Lambda_2, \dots$ and say that such a sequence discriminates between a pair of hypotheses $H_0^{\aleph}(A), H_1^{\aleph}(A)$ about processes from $M_0(A)$, if for each process ϱ, for which $H_1^{\aleph}(A)$ is true, there exists a partition Λ_j for which $H_1^{\aleph}(\Lambda_j)$ is true for the process ϱ_{Λ_j}. We also define a probability distribution $\{\omega = \omega_1, \omega_2, \dots\}$ on integers $\{1, 2, \dots\}$ by

$$\omega_1 = 1 - 1/\log 3, \dots, \omega_i = 1/\log(i+1) - 1/\log(i+2), \dots. \tag{18}$$

(In what follows we will use this distribution, but the theorem described below is obviously true for any distribution with nonzero probabilities.)

Let $H_0^{\aleph}(A), H_1^{\aleph}(A)$ be a pair of hypotheses, $\hat{\Lambda} = \Lambda_1, \Lambda_2, \dots$ be a sequence of partitions, α be from $(0, 1)$ and φ be a code. The scheme for all the tests is as follows:

The hypothesis $H_0^{\aleph}(A)$ is accepted if for all $i = 1, 2, 3, \dots$ the test $T_{\varphi}^{\aleph}(\Lambda_i, (\alpha\omega_i))$ accepts the hypothesis $H_0^{\aleph}(\Lambda_i)$. Otherwise, $H_0^{\aleph}$ is rejected. We denote this test as $\mathbf{T}_{\alpha,\varphi}^{\aleph}(\hat{\Lambda})$.

Comment. It is important to note that one does not need to check an infinite number of inequalities when one applies this test. The point is that the hypothesis $H_0^{\aleph}(A)$ has to be accepted if the left part in (11), (12), (15) and (16), correspondingly, is less than $-\log(\alpha\omega_i)$. Obviously, $-\log(\alpha\omega_i)$ goes to infinity if i increases. That is why there are many cases, where it is enough to check a finite number of hypotheses $H_0^{\aleph}(\Lambda_i)$.

Theorem 5. (i) *For each $\alpha \in (0, 1)$, sequence of partitions $\hat{\Lambda}$ and a code φ, the Type I error of the described test $\mathbf{T}_{\alpha,\varphi}^{\aleph}(\hat{\Lambda})$ is not larger than α and* (ii) *if, in addition, φ is a universal code and $\hat{\Lambda}$ discriminates between $H_0^{\aleph}(A), H_1(A)^{\aleph}$, then the Type II error of the test $\mathbf{T}_{\alpha,\varphi}^{\aleph}(\hat{\Lambda})$ goes to* 0, *when the sample size tends to infinity (in the case of the homogeneity testing, in addition, the inequality* (17) *should be valid).*

5. The experiments

In this part we describe results of some experiments and a simulation study carried out to estimate an efficiency of the suggested tests. The obtained results show that the described tests as well as the suggested approach in general can be used in applications.

5.1. Randomness testing

First we consider the problem of randomness testing, which is a particular case of goodness-of-fit testing. Namely, we will consider a null hypothesis H_0^{rt} that a given bit sequence is generated by Bernoulli source with equal probabilities of 0 and 1 and the alternative hypothesis H_1^{rt} that the sequence is generated by a stationary and ergodic source which differs from the source under H_0^{rt}. This problem is important for random number (RNG) and pseudorandom number generators (PRNG) testing and there are many methods for randomness testing suggested in the literature. Thus, recently National Institute of Standards and Technology (NIST, USA) suggested "A statistical test suite for random and pseudorandom number generators for cryptographic applications", see [27].

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

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

Table 1
Pseudorandom number generators testing

Parameters/length (bits)	400 000	8000 000
M, A, C, X_0		
$10^8+1, 23, 0, 47\,594\,118$	390 240	7635 936
$2^{31}, 2^{16}+3, 0, 1$	Extended	7797 984
$2^{32}, 134\,775\,813, 1, 0$	Extended	Extended

where X_n is the n-th generated number [18]. Each such generator we will denote by LCG(M, A, C, X_0), where X_0 is the initial value of the generator. Such generators are well studied and many of them are used in practice, see [17].

In our experiments we extract 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:

$$\left.\begin{array}{l} \hat{X}_i = \lfloor X_i/256 \rfloor \text{ if } X_i < 256\mu \\ \hat{X}_i = \text{ empty word if } X_i \geq 256\mu \end{array}\right\}. \tag{19}$$

Then a sequence was compressed by the archiver *ACE v 1.2b* (see http://www.winace.com/). Experimental data about testing of three linear congruent generators is given in Table 1.

So, we can see from the first line of the table that the 400 000-bit sequence generated by the LCG($10^8+1, 23, 0, 47\,594\,118$) and transformed according to (19) was compressed to a 390 240-bit sequence. (Here 400 000 is the length of the sequence after transformation.) If we take the level of significance $\alpha \geq 2^{-9760}$ and apply the test $T_\varphi^{id}(\{0,1\}, \alpha)$, ($\varphi = ACE\ v\ 1.2b$), the hypothesis H_0^{rt} should be rejected, see Theorem 1 and (11). Analogously, the second line of the table shows that the 8000 000-bit sequence generated by LCG($2^{31}, 2^{16}+3, 0, 1$) cannot be considered random (H_0^{rt} should be rejected if the level of significance α is greater than $2^{-202\,016}$). On the other hand, the suggested test accepts H_0^{rt} for the sequences generated by the third generator, because the lengths of the "compressed" sequences increased.

The obtained information corresponds to the known data about the considered generators. Thus, it is shown in [17] that the first two generators are bad whereas the third generator was investigated in [22] and is regarded as good. So, we can see that the suggested testing is quite efficient.

In a recently published paper [32] the described method was applied for testing random number and pseudorandom number generators and its efficiency was compared with the mentioned methods from "A statistical test suite for random and pseudorandom number generators for cryptographic applications" [27]. The point is that the tests from [27] are selected basing on comprehensive theoretical and experimental analysis and can be considered as the state-of-the-art in randomness testing. It turned out that the suggested tests, which were based on archivers RAR and ARJ, were more powerful than many methods recommended by NIST in [27]; see [32] for details.

5.2. Simulation study of serial independence testing

A selection of the simulation results concerning independence tests is presented in this part. We generated binary sequences by the first order Markov source with different probabilities (see Table 2) and applied the test $T_\varphi^{SI}(\{0,1\}, \alpha)$ to test the hypothesis H_0^{SI} that a given bit sequence is

Table 2
Serial independence testing for Markov source of order 6 ("rej" means rejected, "acc" — accepted. In all cases $p(x_{i+1} = 0|x_i = 1) = 0.5$)

Probabilities/length (bits)	2^9	2^{14}	2^{16}	2^{18}	2^{23}
$p(x_{i+1} = 0\|x_i = 0) = 0.8$	rej	rej	rej	rej	rej
$p(x_{i+1} = 0\|x_i = 0) = 0.6$	acc	rej	rej	rej	rej
$p(x_{i+1} = 0\|x_i = 0) = 0.55$	acc	acc	rej	rej	rej
$p(x_{i+1} = 0\|x_i = 0) = 0.525$	acc	acc	acc	rej	rej
$p(x_{i+1} = 0\|x_i = 0) = 0.505$	acc	acc	acc	acc	rej

Table 3
Serial independence testing for Markov source of order 6 (in all cases $p(x_{i+1} = 0|(\sum_{j=i-6}^{i} x_i) \bmod 2 = 1) = 0.5$)

Probabilities/length (bits)	2^{14}	2^{18}	2^{20}	2^{23}	2^{28}
$p(x_{i+1} = 0\|(\sum_{j=i-6}^{i} x_j) \bmod 2 = 0) = 0.8$	rej	rej	rej	rej	rej
$p(x_{i+1} = 0\|(\sum_{j=i-6}^{i} x_j) \bmod 2 = 0) = 0.6$	acc	rej	rej	rej	rej
$p(x_{i+1} = 0\|(\sum_{j=i-6}^{i} x_j) \bmod 2 = 0) = 0.55$	acc	acc	rej	rej	rej
$p(x_{i+1} = 0\|(\sum_{j=i-6}^{i} x_j) \bmod 2 = 0) = 0.525$	acc	acc	acc	rej	rej
$p(x_{i+1} = 0\|(\sum_{j=i-6}^{i} x_j) \bmod 2 = 0) = 0.505$	acc	acc	acc	acc	rej

generated by Bernoulli source 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}.

We tried several different archivers and the universal code R described in Appendix B. It turned out that the power of the code R is larger than the power of the tried archivers, that is why we present results for the test $T_R^{SI}(\{0, 1\}, \alpha)$, which is based on this code, for $\alpha = 0.01$. Table 2 contains results of calculations.

We know that the source is Markovian and, hence, the hypothesis H_0^{SI} (that a sequence is generated by Bernoulli source) is not true. Table 2 shows how the value of the Type II error depends on the sample size and the source probabilities.

The similar calculations were carried out for the Markov source of order 6. We applied the test $T_\varphi^{SI}(\{0, 1\}, \alpha)$, $\alpha = 0.01$, for checking the hypothesis H_0^{SI} that a given bit sequence is generated by Markov source of order at most 5 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}. Again, we know that H_0^{SI} is not true and Table 3 shows how the value of the Type II error depends on the sample size and the source probabilities.

6. Conclusion

In this part we point out some generalizations of the suggested approach as well as clarify the connection with some statistical methods.

Having taken into account the Kraft inequality (9), we can rewrite the goodness-of-fit test (11) as follows:

$$\text{if } \pi(\bar{x})/\mu_\varphi(\bar{x}) \geq \alpha \text{ then } H_0, \qquad \text{otherwise } H_1, \tag{20}$$

where, as before, $\mu_\varphi(\bar{x}) = 2^{-|\varphi(\bar{x})|}/\sum_{u\in A^t} 2^{-|\varphi(u)|}$, t is the sample size. Clearly, (20) looks like the likelihood ratio test, which is one of the main statistical tools. Moreover, all other tests can be

presented in the same manner. Thus, if we denote $2^{-(t-lm)h^*_m(\bar{x})}$ from (12) by π, we can rewrite the serial independence test (12) in the same form as (20). The same is true for the independence testing (15) and homogeneity testing (16), if we denote by π the values $2^{-\sum_{k=1}^{d}(t-lm)h^*_m(\bar{x}^{(k)})}$ and $2^{-(t-mr)h^*_m(x^1\diamond x^2\diamond\cdots\diamond x^r)}$, correspondingly.

Now we use the representation (20) in order to extend the suggested tests to the following more general case. Let there be several codes (or archivers) $\varphi_1, \varphi_2, \ldots, \varphi_l$ and we want to build a test, which is based on all of them. In order to get such a test, we define the "mixture" probability distribution and the mixture distribution of codeword lengths by equalities

$$\mu_{\mathrm{mix}}(\bar{x}) = (2^{-|\varphi_1(\bar{x})|} + 2^{-|\varphi_2(\bar{x})|} + \cdots + 2^{-|\varphi_l(\bar{x})|})/l, \qquad |\varphi_{\mathrm{mix}}(\bar{x})| = -\log \mu_{\mathrm{mix}}(\bar{x}),$$

correspondingly. Obviously, the Kraft inequality (9) is valid for $|\varphi_{\mathrm{mix}}|$ and, therefore, $|\varphi_{\mathrm{mix}}|$ can be used in all suggested tests instead of $|\varphi|$. In the case when the set of codes $\varphi_1, \varphi_2, \ldots$ is infinite, we can use some probability distribution τ on the set $1, 2, 3, \ldots$ and define μ_{mix} and $|\varphi_{\mathrm{mix}}|$ as follows:

$$\mu_{\mathrm{mix}}(\bar{x}) = \sum_{i=1}^{\infty} \tau_i 2^{-|\varphi_i(\bar{x})|}, \qquad |\varphi_{\mathrm{mix}}(\bar{x})| = -\log \mu_{\mathrm{mix}}(\bar{x}). \tag{21}$$

(For example, the distribution ω (18) can be used here as the distribution τ.)

It can be easily seen from the descriptions of the tests that their power is greater, if the length of the codeword $|\varphi(\bar{x})|$ is less. That is why it is natural to look for a code φ_i whose length is minimal. First of all we can find such a code φ_δ that

$$-\log(\tau_\delta 2^{-|\varphi_\delta(\bar{x})|}) = \min_i(-\log(\tau_i 2^{-|\varphi_i(\bar{x})|})). \tag{22}$$

Having taken into account (21), we can see that

$$-\log(\tau_\delta 2^{-|\varphi_\delta(\bar{x})|}) \le |\varphi_{\mathrm{mix}}(\bar{x})|.$$

If we denote by φ_{mm} the code, whose codeword length $|\varphi_{mm}(\bar{x})| = -\log(\tau_\delta 2^{-|\varphi_\delta(\bar{x})|})$ for each $\bar{x}$, the later inequality shows that $|\varphi_{mm}(\bar{x})| \le |\varphi_{\mathrm{mix}}(\bar{x})|$ for any sample $\bar{x}$, and, hence, the power of the tests based on the code φ_{mm} is not less than the power of the tests based on the code φ_{mix}.

It is worth noting that codes φ_{mix} and φ_{mm} (and corresponding distributions, which are based on the Kraft inequality (9)), were applied for constructing optimal universal codes and predictors in [28,29] and later both constructions were used in mathematical statistics and related fields under different names (aggregating strategy, weighted majority algorithms, etc.).

One of the reason of a popularity of both constructions is their asymptotical optimality. Thus, in the case of hypothesis testing, the codes φ_{mix} and φ_{mm} give, in a certain sense, the most powerful (asymptotically) tests. Indeed, if we suppose that the family of codes $\varphi_1, \varphi_2, \ldots$ contains a code φ_{opt}, whose codeword length ($|\varphi_{opt}(\bar{x})|$) is minimal (say, with probability 1, when the sample size increases), we can see from the definitions φ_{mix} and φ_{mm} that $|\varphi_{\mathrm{mix}}(\bar{x})| \le |\varphi_{opt}(\bar{x})| + const$ and $|\varphi_{mm}(\bar{x})| \le |\varphi_{opt}(\bar{x})| + const$, where $const = -\log \tau_{opt}$. On the other hand, for any processes (whose entropy is larger than zero), the codeword length $|\varphi_{opt}(\bar{x})|$ goes to infinity, if the sample size ($|\bar{x}|$) increases and, hence, the impact of *const* decreases.

Acknowledgments

The authors wish to thank Andrey Gruzin and Viktor Monarev who carried out all experiments described in part 5. The second author's 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).

Appendix A. Predictors and universal codes

Let a source generate a message $x_1 \dots x_{t-1} x_t \dots$, $x_i \in A$ for all i. After the first t letters $x_1, \dots, x_{t-1}, x_t$ have been processed the following letter x_{t+1} needs to be predicted. By definition, the prediction is the set of nonnegative numbers $\gamma(a_1|x_1 \dots x_t), \dots, \gamma(a_n|x_1 \dots x_t)$ which are estimates of the unknown conditional probabilities $p(a_1|x_1 \dots x_t), \dots, p(a_n|x_1 \dots x_t)$, i.e. of the probabilities $p(x_{t+1} = a_i|x_1 \dots x_t)$; $i = 1, \dots, n$.

Laplace suggested the following predictor:

$$L_0(a|x_1 \dots x_t) = (\nu_{x_1 \dots x_t}(a) + 1)/(t + |A|), \tag{23}$$

see [9]. (We use L_0 here in order to show that it is intended to predict sources from $M_0(A)$. Later this predictor will be extended to $M_i(A)$, $i > 0$.) For example, if $A = \{0, 1\}$, $x_1 \dots x_5 = 01010$, then the Laplace prediction is as follows: $L_0(x_6 = 0|01010) = (3 + 1)/(5 + 2) = 4/7$, $L_0(x_6 = 1|01010) = (2 + 1)/(5 + 2) = 3/7$.

It is natural to estimate the error of prediction by the Kullback–Leibler (K–L) divergence between a distribution p and its estimation. Consider a source p and a predictor γ. The *error* is characterized by the divergence

$$\rho_{\gamma,p}(x_1 \dots x_t) = \sum_{a \in A} p(a|x_1 \dots x_t) \log \frac{p(a|x_1 \dots x_t)}{\gamma(a|x_1 \dots x_t)}. \tag{24}$$

As we mentioned above, for any distributions p and γ the K–L divergence is nonnegative and equals 0 if and only if $p(x) = \gamma(x)$ for all x. For fixed t, $r_{\gamma,p}$ is a random variable, because $x_1, x_2, \dots, x_t$ are random variables. We define the average error at time t by

$$\rho^t(p\|\gamma) = E(r_{\gamma,p}(\cdot)) = \sum_{x_1 \dots x_t \in A^t} p(x_1 \dots x_t) \rho_{\gamma,p}(x_1 \dots x_t). \tag{25}$$

It is shown in [30] that the error of Laplace predictor goes to 0 for any i.i.d. source p. More precisely, it is proven that

$$r^t(p\|L_0) < (|A| - 1)/(t + 1) \tag{26}$$

for any source p (see also [34]).

For any predictor γ we define the corresponding probability measure by

$$\gamma(x_1 \dots x_t) = \prod_{i=1}^{t} \gamma(x_i \mid x_1 \dots x_{i-1}). \tag{27}$$

For example, the Laplace measure L_0 of the word $x_1 \dots x_t = 0101$ is as follows: $L_0(0101) = \frac{1}{2}\frac{1}{3}\frac{1}{2}\frac{2}{5} = \frac{1}{30}$. By analogy with (24) and (25) we define

$$\rho_{\gamma,p}(x_1 \dots x_t) = t^{-1}(\log(p(x_1 \dots x_t)/\gamma(x_1 \dots x_t))) \tag{28}$$

and

$$\bar{\rho}_t(\gamma, p) = t^{-1} \sum_{x_1 \dots x_t \in A^t} p(x_1 \dots x_t) \log(p(x_1 \dots x_t)/\gamma(x_1 \dots x_t)). \tag{29}$$

For example, from those definitions and (26) we obtain the following estimation for Laplace predictor L_0 and any i.i.d. source p:

$$\bar{\rho}_t(L_0, p) < (\log t + c)/t, \tag{30}$$

where c is a constant.

The average error (29) has three interesting characteristics. Firstly, it can be easily seen from (24), (25) and (29) that $\bar{\rho}_t(\gamma, p)$ is the average error of the predictor γ when it is applied to the process p:

$$\bar{\rho}_t(\gamma, p) = t^{-1} \sum_{j=1}^{t} \rho^j(p\|\gamma).$$

Secondly, having taken into account the definition of the Shannon entropy (2), we can easily see that for $p \in M_0(A)$

$$\bar{\rho}_t(\gamma, p) = t^{-1} E_p(-\log \gamma(x_1 \dots x_t)) - h_0(p). \tag{31}$$

The third characteristic is connected with the theory of universal coding. One can construct a code with codelength $\gamma_{\text{code}}(a|x_1 \dots x_t) \approx -\log_2 \gamma(a|x_1 \dots x_n)$ for any letter $a \in A$ (since Shannon's original research, it has been well known, cf. e.g. [11], that, using block codes with large block length or more modern methods of arithmetic coding, the approximation may be as accurate as you like). If one knows the real distribution p, one can base coding on the true distribution p and not on the prediction γ. The difference in performance measured by average code length is given by

$$\sum_{a\in A} p(a|x_1 \dots x_t)(-\log_2 \gamma(a|x_1 \dots x_t)) - \sum_{a\in A} p(a|x_1 \dots x_t)(-\log_2 p(a|x_1 \dots x_t))$$
$$= \sum_{a\in A} p(a|x_1 \dots x_t) \log_2 \frac{p(a|x_1 \dots x_t)}{\gamma(a|x_1 \dots x_t)}.$$

Thus this excess, it is exactly the error (24) defined above. Analogously, if we encode the sequence $x_1 \dots x_t$ based on a predictor γ the redundancy per letter is defined by (28) and (29). So, from a mathematical point of view the universal prediction and universal coding are identical. But $-\log \gamma(x_1 \dots x_t)$ and $-\log p(x_1 \dots x_t)$ have a very natural interpretation. The first value is a codeword length (in bits), if the "code" γ is applied for compressing of the word $x_1 \dots x_t$ and the second one is the minimally possible codeword length. The difference is the redundancy of the code and, at the same time, the error of the predictor. It is worth noting that there are many other deep interrelations between universal coding, prediction and estimation, see [25,29].

As we saw in (30), the average error of the Laplace predictor is upper bounded by $(|A| - 1)(\log t + O(1))/(t + 1)$, when t grows. Krichevsky suggested the predictor $K_0(a|x_1 \dots x_t) = (\nu_{x_1 \dots x_t}(a) + 1/2)/(t + |A|/2)$ and showed that the error of this predictor is asymptotically less: $\bar{\rho}_t(K_0, p)$ is upper bounded by $(|A| - 1)(\log t + O(1))/(2t)$. Moreover, he showed that this predictor is asymptotically optimal in the sense that for any other predictor γ there exists a source $\hat{p}$ for which the error $\bar{\rho}_t(\gamma, \hat{p})$ is not less than $(|A| - 1)(\log t + O(1))/(2t)$, see [19].

From definitions (23) and (27) we can see that the Laplace predictor ascribes the following probabilities:

$$L_0(x_1 \dots x_t) = \prod_{i=1}^{t} \frac{\nu_{x_1 \dots x_{i-1}}(x_i) + 1}{i - 1 + |A|} = \frac{\prod_{a\in A} (\nu_{x_1 \dots x_t}(a))!}{((t + |A| - 1)!)/(|A| - 1)!}. \tag{32}$$

Analogously, for K_0 we obtain

$$K_0(x_1 \dots x_t) = \prod_{i=1}^{t} \frac{\nu_{x_1 \dots x_{i-1}}(x_i) + 1/2}{i - 1 + |A|/2} = \frac{\prod_{a \in A} \left(\prod_{j=1}^{\nu_{x_1 \dots x_t}(a)} (j - 1/2) \right)}{\prod_{i=0}^{t-1} (i + |A|/2)}. \tag{33}$$

The following simple example shows the difference between the predictors: if $A = \{0, 1\}$ and $x_1 \dots x_t = 0101$, then L_0 and K_0 ascribe the probabilities $\frac{1}{2}\frac{1}{3}\frac{1}{2}\frac{2}{5} = \frac{1}{30}$ and $\frac{1}{2}\frac{1}{4}\frac{1}{2}\frac{3}{8} = \frac{3}{128}$, correspondingly.

The product $(r + 1/2)((r + 1) + 1/2) \dots (s - 1/2)$ can be presented as a ratio $\frac{\Gamma(s+1/2)}{\Gamma(r+1/2)}$, where $\Gamma(\,)$ is the gamma function (see for definition, e.g., [17]). So, (33) can be presented as follows:

$$K_0(x_1 \dots x_t) = \frac{\Gamma(|A|/2)}{\Gamma(1/2)^{|A|}} \frac{\prod_{a \in A} \Gamma(\nu_{x_1 \dots x_t}(a) + 1/2)}{\Gamma((t + |A|/2))}. \tag{34}$$

As we mentioned above the average error of the Krichevsky predictor is asymptotically minimal. That is why we will focus our attention on this predictor and, for the sake of completeness, we prove an upper bound for its error.

Claim 1. *For any stationary and ergodic source generating letters from a finite alphabet A the average error of K_0 is upper bounded as follows:*

$$-t^{-1} \sum_{x_1 \dots x_t \in A^t} p(x_1 \dots x_t) \log(K_0(x_1 \dots x_t)) - h_0(p) \le ((|A| - 1) \log t + C)/(2t),$$

where C is a constant.

Proof is given in Appendix B.

Comment. In particular, if the source is i.i.d., the average error is less than $((|A| - 1) \log t + C)/(2t)$; see (31).

We indicated that extensions of both predictors to cover the general Markov case are possible. We take this up now. The trick is to view a Markov source $p \in M_m(A)$ as resulting from $|A|^m$ i.i.d. sources. We illustrate this idea by an example from [34]. So assume that $A = \{O, I\}$, $m = 2$ and assume that the source $p \in M_2(A)$ has generated the sequence

$$O O I O I I O O I I I O I O.$$

We represent this sequence by the following four subsequences:

$$* * I * * * * * I * * * * *,$$
$$* * * O * I * * * I * * * O,$$
$$* * * * I * * O * * * * I *,$$
$$* * * * * * O * * * I O * *.$$

These four subsequences contain letters which follow OO, OI, IO and II, respectively. By definition, $p \in M_m(A)$ if $p(a|x_1 \dots x_t) = p(a|x_{t-m+1} \dots x_t)$, for all $0 < m \le t$, all $a \in A$ and all $x_1 \dots x_t \in A^t$. Therefore, each of the four generated subsequences may be considered to be generated by a Bernoulli source. Further, it is possible to reconstruct the original sequence if we know the four $(=|A|^m)$ subsequences and the two $(=m)$ first letters of the original sequence.

Any predictor γ for i.i.d. sources can be applied for Markov sources. Indeed, in order to predict, it is enough to store in the memory $|A|^m$ sequences, one corresponding to each word in A^m. Thus, in the example, the letter x_3 which follows OO is predicted based on the Bernoulli method γ corresponding to the x_1x_2-subsequence ($=OO$), then x_4 is predicted based on the Bernoulli method corresponding to x_2x_3, i.e. to the OI-subsequence, and so forth. When this scheme is applied along with either L_0 or K_0 we denote the obtained predictors as L_m and K_m, correspondingly and define the probabilities for the first m letters as follows: $L_m(x_1) = L_m(x_2) = \dots L_m(x_m) = 1/|A|$, $K_m(x_1) = K_m(x_2) = \dots K_m(x_m) = 1/|A|$.

Having taken into account (32) and (34), we can present the Laplace and Krichevsky predictors for $M_m(A)$ as follows:

$$L_m(x_1 \dots x_t) = \begin{cases} \dfrac{1}{|A|^t}, & \text{if } t \le m; \\ \dfrac{1}{|A|^m} \displaystyle\prod_{v \in A^m} \frac{\prod_{a \in A} (\nu_x(va))!}{((\bar{\nu}_x(v) + |A|)!)/(|A| - 1)!}, & \text{if } t > m, \end{cases} \tag{35}$$

$$K_m(x_1 \dots x_t) = \begin{cases} \dfrac{1}{|A|^t}, & \text{if } t \le m; \\ \dfrac{1}{|A|^m} \left(\dfrac{\Gamma(|A|/2)}{\Gamma(1/2)^{|A|}} \right)^{|A|^m} \displaystyle\prod_{v \in A^m} \frac{\prod_{a \in A} (\Gamma(\nu_x(va) + 1/2))}{(\Gamma(\bar{\nu}_x(v) + |A|/2))}, & \text{if } t > m, \end{cases} \tag{36}$$

where $\bar{\nu}_x(v) = \sum_{a \in A} \nu_x(va)$, $x = x_1 \dots x_t$.

We have seen that any source from $M_m(A)$ can be presented as a "sum" of $|A|^m$ and i.i.d. sources. From this we can easily see that the error of a predictor for the source from $M_m(A)$ can be upper bounded by the error of i.i.d. source multiplied by $|A|^m$. In particular, we obtain from Claim 1 the following upper bound.

Claim 2. *For any stationary and ergodic source generated letters from a finite alphabet A the average error of the Krichevsky predictor K_m is upper bounded as follows:*

$$-t^{-1} \sum_{x_1 \dots x_t \in A^t} p(x_1 \dots x_t) \log(K_m(x_1 \dots x_t)) - h_m(p) \le |A|^m((|A| - 1) \log t + C)/(2t),$$

where C is a constant.

Now we can describe the universal predictor R and code R_{code} from [28,29]. By definition,

$$R(x_1 \dots x_t) = \sum_{i=0}^{\infty} \omega_{i+1} K_i(x_1 \dots x_t),$$
$$R(x_t \mid x_1 \dots x_{t-1}) = R(x_1 \dots x_t)/R(x_1 \dots x_{t-1})$$

and $|R_{\text{code}}(x_1 \dots x_t)| = -\log R(x_1 \dots x_t)$. It is worth noting that this construction can be applied to the Laplace predictor (if we use L_i instead of K_i) and any other family of predictors (or codes).

Claim 3. *Let the predictor R be applied to a source p. Then, for any stationary and ergodic source $p \in M_\infty(A)$ the error (29) of the predictor R goes to 0, when the sample size t goes to ∞.*

Proof can be derived from Claim 2 and the properties of the Shannon entropy. Indeed, we can see from the definition of R and Claim 2 that the average error is upper bounded as follows:

$$-t^{-1}\sum_{x_1\dots x_t\in A^t} p(x_1\dots x_t)\log(R(x_1\dots x_t)) - h_k(p)$$
$$\le (|A|^k(|A|-1)\log t + \log(1/\omega_i) + C)/(2t),$$

for any $k = 0, 1, 2, \dots$. Taking into account that for any $p \in M_\infty(A)$ $\lim_{k\to\infty} h_k(p) = h_\infty(p)$, we can see that

$$\left(\lim_{t\to\infty} t^{-1}\sum_{x_1\dots x_t\in A^t} p(x_1\dots x_t)\log(R(x_1\dots x_t)) - h_\infty(p)\right) = 0.$$

The main property of the universal codes (10) is also true for R_{code} and can be easily derived from Claim 3 using standard techniques of ergodic theory.

Appendix B. Proofs

Proof of the lemma. First we show that for any source $\theta^* \in M_0(A)$ and any words $x^1 = x^1_1\dots x^1_{t_1}, \dots, x^r = x^r_1\dots x^r_{t_r}$,

$$\theta^*(x^1\diamond\dots\diamond x^r) = \prod_{a\in A}(\theta^*(a))^{\nu_{x^1\diamond\dots\diamond x^r}(a)}$$
$$\le \prod_{a\in A}(\nu_{x^1\diamond\dots\diamond x^r}(a)/t)^{\nu_{x^1\diamond\dots\diamond x^r}(a)}, \tag{37}$$

where $t = \sum_{i=1}^r t_i$. Here the equality holds, because $\theta^* \in M_0(A)$. The inequality follows from (7). Indeed, if $p(a) = \nu_{x^1\diamond\dots\diamond x^r}(a)/t$ and $q(a) = \theta^*(a)$, then

$$\sum_{a\in A}\frac{\nu_{x^1\diamond\dots\diamond x^r}(a)}{t}\log\frac{(\nu_{x^1\diamond\dots\diamond x^r}(a)/t)}{\theta^*(a)} \ge 0.$$

From the latter inequality we obtain (37). Taking into account the definition (6) and (37), we can see that the statement of the lemma is true for this particular case.

For any $\theta \in M_m(A)$ and $x = x_1\dots x_s, s > m$, we present $\theta(x_1\dots x_s)$ as $\theta(x_1\dots x_s) = \theta(x_1\dots x_m)\prod_{u\in A^m}\prod_{a\in A}\theta(a|\,u)^{\nu_x(ua)}$, where $\theta(x_1\dots x_m)$ is the limit probability of the word $x_1\dots x_m$. Hence, $\theta(x_1\dots x_s) \le \prod_{u\in A^m}\prod_{a\in A}\theta(a|\,u)^{\nu_x(ua)}$. Taking into account the inequality (37), we obtain $\prod_{a\in A}\theta(a|\,u)^{\nu_x(ua)} \le \prod_{a\in A}(\nu_x(ua)/\bar\nu_x(u))^{\nu_x(ua)}$ for any word u. Hence,

$$\theta(x_1\dots x_s) \le \prod_{u\in A^m}\prod_{a\in A}\theta(a|u)^{\nu_x(ua)}$$
$$\le \prod_{u\in A^m}\prod_{a\in A}(\nu_x(ua)/\bar\nu_x(u))^{\nu_x(ua)}.$$

If we apply those inequalities to $\theta(x^1\diamond\dots\diamond x^r)$, we immediately obtain the following inequalities

$$\theta(x^1\diamond\dots\diamond x^r) \le \prod_{u\in A^m}\prod_{a\in A}\theta(a|\,u)^{\nu_{x^1\diamond\dots\diamond x^r}(ua)}$$
$$\le \prod_{u\in A^m}\prod_{a\in A}(\nu_{x^1\diamond\dots\diamond x^r}(ua)/\bar\nu_{x^1\diamond\dots\diamond x^r}(u))^{\nu_{x^1\diamond\dots\diamond x^r}(ua)}.$$

Now the statement of the lemma follows from the definition (6).

Proof of Theorem 1. In order to avoid cumbersome notations we first consider a case where the sample $\bar{x}$ is one sequence $x_1 \dots x_t$ and then note how the proof can be extended for the general case. Let C_α be a critical set of the test $T_\varphi^{id}(A, \alpha)$, 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 (9) is true. We define an auxiliary 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 (9) 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 Theorem 1 is proven.

Let us prove the second statement of the theorem. Suppose that the hypothesis $H_1^{id}(A)$ 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

$$\lim_{t \to \infty} - \log \pi(x_1 \dots x_t) - |\varphi(x_1 \dots x_t)| = \infty \tag{38}$$

with probability 1 (according to the measure τ). First we represent (38) as

$$\begin{aligned} &- \log \pi(x_1 \dots x_t) - |\varphi(x_1 \dots x_t)| \\ &\quad = t \left(\frac{1}{t} \log \frac{\tau(x_1 \dots x_t)}{\pi(x_1 \dots x_t)} + \frac{1}{t}(- \log \tau(x_1 \dots x_t) - |\varphi(x_1 \dots x_t)|) \right). \end{aligned}$$

From this equality and the property of a universal code (10) we obtain

$$- \log \pi(x_1 \dots x_t) - |\varphi(x_1 \dots x_t)| = t \left(\frac{1}{t} \log \frac{\tau(x_1 \dots x_t)}{\pi(x_1 \dots x_t)} + o(1) \right). \tag{39}$$

From (2)–(4) we can see that

$$\lim_{t \to \infty} - \log \tau(x_1 \dots x_t)/t \leq h_k(\tau) \tag{40}$$

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:

$$\begin{aligned} &- \log \pi(x_1 \dots x_t)/t = -t^{-1} \sum_{i=1}^{t} \log \pi(x_i | \, x_1 \dots x_{i-1}) \\ &\quad = -t^{-1} \left(\sum_{i=1}^{k} \log \pi(x_i | \, x_1 \dots x_{i-1}) + \sum_{i=k+1}^{t} \log \pi(x_i | \, x_{i-k} \dots x_{i-1}) \right) \end{aligned}$$

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 $h_k(\tau)$, see [2,11]. 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, (40) and (39), we can see that

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

for any $k \geq s$. From this inequality and (7) we can obtain that $-\log \pi(x_1 \dots x_t) - |\varphi(x_1 \dots x_t)| \geq ct + o(t)$, where c is a positive constant, $t \to \infty$. Hence, (38) is true.

Let us consider a case where $\bar{x}$ is a sequence $x^1 = x_1^1 \dots x_{t_1}^1, \dots, x^l = x_1^l \dots x_{t_l}^l$ (i.e. $\bar{x} = x^1 \diamond \dots \diamond x^l$). The proof of the first statement of the theorem is analogical and can be simply repeated for this case. In order to prove the second statement we note that the length of at least one sequence x^i goes to infinity and, hence, the equality (38) is true for that sequence, whereas for all other sequences the differences $\log \pi(x^j) - |\varphi(x^j)|$ are either bounded or go to infinity. The theorem is proven.

Proof of Theorem 2. We only consider a case where the sample $\bar{x}$ is one sequence $x_1 \dots x_t$, because the general case is analogical, but requires cumbersome notations. Let us denote the critical set of the test $T_\varphi^{SI}(A, \alpha)$ as C_α, i.e., by definition, $C_\alpha = \{x_1 \dots x_t : (t-m)h_m^*(x_1 \dots x_t) - |\varphi(x_1 \dots x_t)| > \log(1/\alpha)\}$. From (9) we can see that there exists such a measure μ_φ that $-\log \mu_\varphi(x_1 \dots x_t) \leq |\varphi(x_1 \dots x_t)|$. We also define

$$\hat{C}_\alpha = \{x_1 \dots x_t : (t-m)h_m^*(x_1 \dots x_t) - (-\log \mu_\varphi(x_1 \dots x_t)) > \log(1/\alpha)\}. \tag{41}$$

Obviously, $\hat{C}_\alpha \supset C_\alpha$. Let θ be any source from $M_m(A)$. The following chain of equalities and inequalities is true:

$$1 \geq \mu_\varphi(\hat{C}_\alpha) = \sum_{x_1 \dots x_t \in \hat{C}_\alpha} \mu_\varphi(x_1 \dots x_t) \geq \alpha^{-1} \sum_{x_1 \dots x_t \in \hat{C}_\alpha} 2^{(t-m)h_m^*(x_1 \dots x_t)} \geq \alpha^{-1} \sum_{x_1 \dots x_t \in \hat{C}_\alpha} \theta(x_1 \dots x_t) = \theta(\hat{C}_\alpha).$$

(Here both equalities and the first inequality are obvious, the second and the third inequalities follow from (41) and the lemma, correspondingly.) So, we obtain that $\theta(\hat{C}_\alpha) \leq \alpha$ for any source $\theta \in M_m(A)$. 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 Type I error is not greater than α. The first statement of the theorem is proven.

The proof of the second statement will be based on some results of Information Theory. We obtain from (10) and (4) that for any stationary and ergodic p

$$\lim_{t \to \infty} t^{-1} |\varphi(x_1 \dots x_t)| = h_\infty(p) \tag{42}$$

with probability 1. It can be seen from (5) that h_m^* is an estimate for the m-order Shannon entropy (2). Applying the ergodic theorem we obtain $\lim_{t \to \infty} h_m^*(x_1 \dots x_t) = h_m(p)$ with probability 1; see [2,11]. It is known in Information Theory that $h_m(\varrho) - h_\infty(\varrho) > 0$, if ϱ belongs to $M_\infty(A) \setminus M_m(A)$, see [2,11]. It is supposed that $H_1^{SI}(A)$ is true, i.e. the considered process belongs to $M_\infty(A) \setminus M_m(A)$. So, from (42) and the last equality we obtain that $\lim_{t \to \infty}((t-m)h_m^*(x_1 \dots x_t) - |\varphi(x_1 \dots x_t)|) = \infty$. This proves the second statement of the theorem.

Proof of Theorem 3. As before, we only consider a case where the sample $\bar{x}$ is one sequence $x_1 \dots x_t$, because the general case is analogical. Let C_α be a critical set of the test, i.e., by definition, $C_\alpha = \{(x_1, \dots, x_t) : \sum_{i=1}^{d}(t-m)h_m^*(x_1^{(i)} \dots x_t^{(i)}) - |\varphi(x_1 \dots x_t)| > \log(1/\alpha)\}$. There exists a measure μ_φ, for which (9) is valid. Hence,

$$C_\alpha \subset C_\alpha^* \equiv \left\{(x_1, \dots, x_t) : \sum_{i=1}^{d}(t-m)h_m^*(x_1^{(i)} \dots x_t^{(i)}) - \log(1/\mu_\varphi(x_1, \dots, x_t)) > \log(1/\alpha)\right\}. \tag{43}$$

Let θ be any measure from $M_m(A)$. Then

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

Having taken into account the lemma, we obtain

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

It is supposed that H_0^{ind} is true and, hence, (13) is valid. So, from the latter inequalities we can see that $1 \geq \mu_\varphi(C_\alpha^*) \geq \sum_{x_1,\dots,x_t \in C_\alpha^*} \theta(x_1, \dots, x_t)$. Taking into account that $\sum_{x_1,\dots,x_t \in C_\alpha^*} \theta(x_1, \dots, x_t) = \theta(C_\alpha^*)$ and (43), we obtain that $\theta(C_\alpha) \leq \alpha$. 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 $th_m(\mu)$. (Informally, it uses dependence for the better compression.) That is why the difference $(\sum_{i=1}^{d} h_m(\mu^i) - th_m(\mu))$ goes to infinity, when t increases and, hence, H_0^{ind} will be rejected.

Proof of Theorem 4. In short, we consider a case of two samples and i.i.d. sources (i.e. $m = 0$), because a generalization is obvious. So, there are two samples $x^1 = x_1^1 \dots x_{t_1}^1$ and $x^2 = x_1^2 \dots x_{t_2}^2$ generated by sources from $M_0(A)$. As before, let C_α be a critical set of the test, i.e., by definition, $C_\alpha = \{(x^1, x^2) : (t_1+t_2)h_0(x^1 \diamond x^2) - (|\varphi(x^1)| + |\varphi(x^2)|) > \log(1/\alpha)\}$. There exists a measure μ_φ for which (9) is valid. So, $C_\alpha \supset C_\alpha^* \equiv \{(x^1, x^2) : (t_1+t_2)h_0^*(x^1 \diamond x^2) - (\log(1/\mu_\varphi(x^1)) + \log(1/\mu_\varphi(x^2))) > \log(1/\alpha)\}$. Let us suppose that H_0^{hom} is true. It means that (x^1, x^2) are created by some source $\theta \in M_0(A)$. Having taken into account the definition of the set C_α^* and the lemma, we obtain the following chain of inequalities:

$$1 \geq \mu_\varphi(C_\alpha^*) = \sum_{(x^1,x^2) \in C_\alpha^*} \mu_\varphi(x^1 \diamond x^2) \geq \alpha^{-1} \sum_{(x^1,x^2) \in C_\alpha^*} 2^{-(t_1+t_2)h_0^*(x^1 \diamond x^2)} \geq \sum_{(x^1,x^2) \in C_\alpha^*} \theta(x^1 \diamond x^2) = \theta(C_\alpha^*).$$

Hence, $\theta(C_\alpha^*) \leq \alpha$ and, taking into account that the critical set $C_\alpha \subset C_\alpha^*$, we finish the proof of the first statement of the theorem.

Let us suppose that H_1^{hom} is true, i.e. the samples x^1, x^2 are generated by different sources θ_1, θ_2, correspondingly. For any $\gamma \in (0, 1)$ we define $\theta_\gamma = \gamma\theta_1 + (1-\gamma)\theta_2$ and let

$$\delta = \inf_{\gamma\in[c,1-c]}(h_0(\theta_\gamma) - (h_0(\theta_1) + h_0(\theta_2))), \tag{44}$$

where c is defined in (17). Due to the Jensen inequality for the Shannon entropy, we can easily see that $\delta > 0$. Having taken into account the definition of a universal code and ergodicity of θ_1, θ_2 we obtain that

$$(t_1+t_2)h_0^*(x^1 \diamond x^2) - (|\varphi(x^1)| + |\varphi(x^2)|) = (t_1+t_2)\left(h_0\left(\frac{t_1}{t_1+t_2}\theta_1 + \frac{t_2}{t_1+t_2}\theta_2\right)\right.$$
$$\left. - \left(\frac{t_1}{t_1+t_2}h_0(\theta_1) + \frac{t_2}{t_1+t_2}h_0(\theta_2)\right)\right) + o(t_1+t_2),$$

(with probability 1), if $(t_1+t_2) \to \infty$. Taking into account the definition (44) and (17) we obtain from the last equality that

$$(t_1+t_2)h_0^*(x^1 \diamond x^2) - (|\varphi(x^1)| + |\varphi(x^2)|) > \delta(t_1+t_2) + o(t_1+t_2).$$

Hence, the difference

$$(t_1+t_2)h_0^*(x^1 \diamond x^2) - (|\varphi(x^1)| + |\varphi(x^2)|)$$

goes to infinity and the second statement of the theorem is proven.

Proof of Theorem 5. The following chain proves the first statement of the theorem:

$$\Pr\left\{H_0^{\aleph}(A) \textit{ is rejected}/H_0 \textit{ is true}\right\}$$
$$= \Pr\left\{\bigcup_{i=1}^{\infty}\{H_0^{\aleph}(\Lambda_i) \textit{ is rejected}/H_0 \textit{ is true}\}\right\}$$
$$\le \sum_{i=1}^{\infty}\Pr\{H_0^{\aleph}(\Lambda_i)/H_0 \textit{ is true}\} \le \sum_{i=1}^{\infty}(\alpha\omega_i) = \alpha.$$

(Here both inequalities follow from the description of the test, whereas the last equality follows from (18).)

The second statement also follows from the description of the test. Indeed, let a sample be created by a source ϱ, for which $H_1(A)^{\aleph}$ is true. It is supposed that the sequence of partitions $\hat{\Lambda}$ discriminates between $H_0^{\aleph}(A)$, $H_1^{\aleph}(A)$. By definition, it means that there exists j for which $H_1^{\aleph}(\Lambda_j)$ is true for the process ϱ_{Λ_j}. It immediately follows from Theorems 1–4 that the Type II error of the test $T_\varphi^{\aleph}(\Lambda_j, \alpha\omega_j)$ goes to 0, when the sample size tends to infinity.

Proof of Claim 1. From (34) we obtain:

$$-\log K_0(x_1\dots x_t) = -\log\left(\frac{\Gamma(|A|/2)}{\Gamma(1/2)^{|A|}}\frac{\prod_{a\in A}\Gamma(\nu_{x_1\dots x_t}(a)+1/2)}{\Gamma(t+|A|/2)}\right)$$
$$= c_1 + c_2|A| + \log\Gamma(t+|A|/2) - \sum_{a\in A}\Gamma(\nu_{x_1\dots x_t}(a)+1/2),$$

where c_1, c_2 are constants. Now we use the well known Stirling formula

$$\ln\Gamma(s) = \ln\sqrt{2\pi} + (s-1/2)\ln s - s + \theta/12,$$

where $\theta \in (0, 1)$, see, e.g., [17]. Using this formula we rewrite the previous equality as

$$-\log K_0(x_1 \dots x_t) = -\sum_{a\in A} \nu_{x_1\dots x_t}(a) \log(\nu_{x_1\dots x_t}(a)/t) + (|A| - 1)\log t/2 + \bar{c}_1 + \bar{c}_2|A|,$$

where $\bar{c}_1, \bar{c}_2$ are constants. Having taken into account the definition of the empirical entropy (5), we obtain

$$-\log K_0(x_1 \dots x_t) \le t h_0^*(x_1 \dots x_t) + (|A| - 1)\log t/2 + c|A|.$$

Hence,

$$\sum_{x_1\dots x_t\in A^t} p(x_1 \dots x_t)(-\log(K_0(x_1 \dots x_t))) \le t \sum_{x_1\dots x_t\in A^t} p(x_1 \dots x_t) h_0^*(x_1 \dots x_t) + (|A| - 1)\log t/2 + c|A|.$$

Having taken into account the definition (5), we apply the well known Jensen inequality for the concave function $-x \log x$ and obtain the following inequality:

$$\sum_{x_1\dots x_t\in A^t} p(x_1 \dots x_t)(-\log(K_0(x_1 \dots x_t))) \le -t \sum_{x_1\dots x_t\in A^t} p(x_1 \dots x_t)((\nu_{x_1\dots x_t}(a)/t)) \log \sum_{x_1\dots x_t\in A^t} p(x_1 \dots x_t)(\nu_{x_1\dots x_t}(a)/t) + (|A| - 1)\log t/2 + c|A|.$$

The source p is stationary and ergodic, so the average frequency $\sum_{x_1\dots x_t\in A^t} p(x_1 \dots x_t)\nu_{x_1\dots x_t}(a)$ is equal to $p(a)$ for any $a \in A$ and we obtain from the two last formulas the following inequality:

$$\sum_{x_1\dots x_t\in A^t} p(x_1 \dots x_t)(-\log(K_0(x_1 \dots x_t))) \le t h_0(p) + (|A| - 1)\log t/2 + c|A|$$

(where $h_0(p) = -\sum_{a\in A} p(a)\log p(a)$ is the Shannon entropy). Claim 1 is proven.

References

[1] G.J. Babu, A. Boyarsky, Y.P. Chaubey, P. Gora, New statistical method for filtering and entropy estimation of a chaotic map from noisy data, International Journal of Bifurcation and Chaos 14 (11) (2004) 3989–3994.

[2] P. Billingsley, Ergodic Theory and Information, John Wiley & Sons, 1965.

[3] R. Cilibrasi, P.M.B. Vitanyi, Clustering by compression, IEEE Transactions on Information Theory 51 (4) (2005) 1523–1545.

[4] R. Cilibrasi, R. de Wolf, P.M.B. Vitanyi, Algorithmic clustering of music, Computer Music Journal 28 (4) (2004) 49–67.

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

[6] G.A. Darbellay, I. Vajda, Entropy expressions for multivariate continuous distributions, Research Report no. 1920, UTIA, Academy of Science, Prague, 1998. library@utia.cas.cz.

[7] G.A. Darbellay, I. Vajda, Estimation of the mutual information with data-dependent partitions, IEEE Transactions on Information Theory 48 (5) (1999) 1061–1081.

[8] M. Effros, K. Visweswariah, S.R. Kulkarni, S. Verdu, Universal lossless source coding with the Burrows Wheeler transform, IEEE Transactions on Information Theory 48 (5) (2002) 1061–1081.

[9] W. Feller, An Introduction to Probability Theory and Its Applications, vol. 1, John Wiley & Sons, New York, 1970.

[10] B.M. Fitingof, Optimal encoding for unknown and changing statistics of messages, Problems of Information Transmission 2 (2) (1966) 3–11.

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

[12] K. Ghoudi, R.J. Kulperger, B. Remillard, A nonparametric test of serial independence for time series and residuals, Journal of Multivariate Analysis 79 (2) (2001) 191–218.
[13] P. Jacquet, W. Szpankowski, L. Apostol, Universal predictor based on pattern matching, IEEE Transactions on Information Theory 48 (2002) 1462–1472.
[14] M.G. Kendall, A. Stuart, The advanced theory of statistics; vol. 2: Inference and relationship, London, 1961.
[15] J. Kieffer, Prediction and Information Theory, 1998 (preprint). Available at ftp://oz.ee.umn.edu/users/kieffer/papers/prediction.pdf/.
[16] J.C. Kieffer, E.-H. Yang, Grammar-based codes: a new class of universal lossless source codes, IEEE Transactions on Information Theory 46 (3) (2000) 737–754.
[17] D.E. Knuth, The Art of Computer Programming, vol. 2, Addison Wesley, 1981.
[18] A.N. Kolmogorov, Three approaches to the quantitative definition of information, Problems of Information Transmission 1 (1965) 3–11.
[19] R. Krichevsky, Universal Compression and Retrieval, Kluver Academic Publishers, 1993.
[20] S. Kullback, Information Theory and Statistics, Wiley, New York, 1959.
[21] U. Maurer, Information-theoretic cryptography, in: Advances in Cryptology — CRYPTO'99, in: Lecture Notes in Computer Science, vol. 1666, Springer-Verlag, 1999, pp. 47–64.
[22] O. Moeschlin, E. Grycko, C. Pohl, F. Steinert, Experimental Stochastics, Springer-Verlag, Berlin, 1998.
[23] G. Morvai, S.J. Yakowitz, P.H. Algoet, Weakly convergent nonparametric forecasting of stationary time series, IEEE Transactions on Information Theory 43 (1997) 483–498.
[24] A.B. Nobel, On optimal sequential prediction, IEEE Transactions on Information Theory 49 (1) (2003) 83–98.
[25] J. Rissanen, Universal coding, information, prediction, and estimation, IEEE Transaction on Information Theory 30 (4) (1984) 629–636.
[26] J. Rissanen, Hypothesis selection and testing by the MDL principle, The Computer Journal 42 (4) (1999) 260–269.
[27] A. Rukhin, J. Soto, J. Nechvatal, M. Smid, D. Banks, A statistical test suite for random and pseudorandom number generators for cryptographic applications, NIST Special Publication 800-22, 2001. http://csrc.nist.gov/rng/SP800-22b.pdf.
[28] B.Ya. Ryabko, Twice-universal coding, Problems of Information Transmission 20 (3) (1984) 173–177.
[29] B.Ya. Ryabko, Prediction of random sequences and universal coding, Problems of Information Transmission 24 (2) (1988) 87–96.
[30] B.Ya. Ryabko, A fast adaptive coding algorithm, Problems of Information Transmission 26 (4) (1990) 305–317.
[31] B. Ryabko, J. Astola, Universal codes as a basis for nonparametric testing of serial independence for time series, Journal of Statistical Planning and Inference (in press).
[32] B. Ryabko, V. Monarev, Using information theory approach to randomness testing, Journal of Statistical Planning and Inference 133 (1) (2005) 95–110.
[33] B. Ryabko, Zh. Reznikova, Using Shannon entropy and Kolmogorov complexity to study the communicative system and cognitive capacities in ants, Complexity 2 (2) (1996) 37–42.
[34] B. Ryabko, F. Topsoe, On asymptotically optimal methods of prediction and adaptive coding for Markov sources, Journal of Complexity 18 (1) (2002) 224–241.
[35] S.A. Savari, A probabilistic approach to some asymptotics in noiseless communication, IEEE Transactions on Information Theory 46 (4) (2000) 1246–1262.
[36] C.E. Shannon, A mathematical theory of communication, Bell System Technical Journal 27 (1948) 379–423, 623–656.
[37] C.E. Shannon, Communication theory of secrecy systems, Bell System Technical Journal 28 (1948) 656–715.
[38] P.C. Shields, The interactions between ergodic theory and information theory, IEEE Transactions on Information Theory 44 (6) (1998) 2079–2093.