



Available online at [www.sciencedirect.com](http://www.sciencedirect.com)



Journal of Statistical Planning and  
Inference III (III) III-III

journal of  
statistical planning  
and inference

[www.elsevier.com/locate/jspi](http://www.elsevier.com/locate/jspi)

1  
3  
Universal codes as a basis for nonparametric testing  
of serial independence for time series<sup>☆</sup>

5  
7  
Boris Ryabko<sup>a,\*</sup>, Jaakko Astola<sup>b</sup>

<sup>a</sup>*Siberian State University of Telecommunications and Computer Science, Russia*

<sup>b</sup>*Tampere University of Technology, Finland*

Received 26 August 2004; received in revised form 2 April 2005; accepted 24 July 2005

---

9  
11  
13  
15  
17  
**Abstract**

We address the problem of nonparametric testing of serial independence for time series and its generalization. More precisely, we consider a stationary and ergodic source  $p$ , which generates symbols  $x_1 \dots x_t$  from some finite set  $A$  and a null hypothesis  $H_0$  that  $p$  is a Markov source of order at most  $m$ , ( $m \geq 0$ ). The alternative hypothesis  $H_1$  is that the sequence is generated by a stationary and ergodic source, which differs from the source under  $H_0$ . In particular, if  $m = 0$  we have the null hypothesis  $H_0$  that the sequence is generated by a Bernoulli source (i.e. the hypothesis that  $x_1 \dots x_t$  are independent). In this paper some new tests that are based on so-called universal codes and universal predictors, are suggested.

© 2005 Elsevier B.V. All rights reserved.

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

19  
*Keywords:* Independence; Serial independence; Universal coding; Hypothesis testing; Information theory; Markov process; Random process; Prediction

---

<sup>☆</sup>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.).

\* Corresponding author.

E-mail address: [boris@ryabko.net](mailto:boris@ryabko.net) (B. Ryabko).

## 1. Introduction

Nonparametric testing of independence in time series is very important in statistical applications. There is an extensive literature dealing with nonparametric independence testing. We mention only the well-known methods that are based on the chi-square tests (see for review Kendall and Stuart, 1961) and the classical papers of Hoeffding (1948) and Blum et al. (1961); quite a full review can also be found in Ghoudi et al. (2001).

In this paper, we consider a source (or process), which generates elements from a finite set  $A$  and the following two hypotheses:  $H_0$  that the source is Markovian one of order not larger than  $m$ , ( $m \geq 0$ ), and the alternative hypothesis  $H_1$  that the sequence is generated by a stationary and ergodic source, which differs from the source under  $H_0$ . The test should be based on a sample  $x_1 \dots x_t$  generated by the source.

For example, the sequence  $x_1 \dots x_t$  might be a DNA-string and one can consider the question about the depth of the statistical dependence.

We suggest a family of tests that are based on so-called universal predictors (or universal data compression methods). The Type I errors of the tests are not larger than a given  $\alpha$  ( $\alpha \in (0, 1)$ ) for any source under  $H_0$ , whereas the Type II error for any source under  $H_1$  tends to 0, when the sample size  $t$  grows.

The tests are based on results and ideas of Information Theory and, especially, on those of universal coding. Informally, the idea of the tests can be described as follows. Suppose that the source generates letters from an alphabet  $A$  and one wants to test  $H_0$  (the source is Markovian of order  $m$ ,  $m \geq 0$ ). First we recall that there exist so-called universal codes which, loosely speaking can “compress” any sequence of length  $t$  generated by a stationary and ergodic source, to the length  $th_\infty$  bits, where  $h_\infty$  is the limiting Shannon entropy as  $t$  tends to infinity. Secondly, it is well known in Information Theory that  $h_\infty$  equals  $m$ th-order (conditional) Shannon entropy  $h_m$ , if  $H_0$  is true, and  $h_\infty$  is strictly less than  $h_m$  if  $H_1$  is true. So, the following test appears natural: compress the sample sequence  $x_1 \dots x_t$  by a universal code and compare the length of the obtained file with  $th_m^*$ , where  $h_m^*$  is an estimate of  $h_m$ . If the length of the compressed file is significantly less than  $th_m^*$ , then the hypothesis  $H_0$  should be rejected.

It is no surprise that the results and ideas of universal coding can be applied to some classical problems of mathematical statistics. In fact, methods of universal coding (and the closely connected universal prediction) extract information from observed data in order to compress (or predict) data efficiently in the case where the source statistics is unknown. Recently such a connection between universal coding and mathematical statistics was used by Csiszár and Shields (2000) for estimating the order of Markov sources and by Ryabko and Monarev (2005) for constructing efficient tests for randomness, i.e. for testing the hypothesis  $\hat{H}_0$  that a sequence is generated by a Bernoulli source and all letters have equal probabilities against  $\hat{H}_1$  that the sequence is generated by a stationary and ergodic source, which differs from the source under  $\hat{H}_0$ .

The outline of the paper is as follows. The next part contains definitions and necessary information from the theory of universal coding and universal prediction. Part three is devoted to testing the above described hypotheses. All proofs are given in the appendix.

1 **2. Definitions and preliminaries**

3 Consider an alphabet  $A = \{a_1, \dots, a_n\}$  with  $n \geq 2$  letters and denote by  $A^t$  the set of words  
 4  $x_1 \dots x_t$  of length  $t$  from  $A$ . Let  $p$  be a source which generates letters from  $A$ . Formally,  $p$  is  
 5 a probability distribution on the set of words of infinite length or, more simply,  $p = (p^t)_{t \geq 1}$   
 6 is a consistent set of probabilities over the sets  $A^t$ ;  $t \geq 1$ . By  $M_\infty(A)$  we denote the set of  
 7 all stationary and ergodic sources, which generate letters from  $A$ . Let  $M_m(A) \subset M_\infty(A)$   
 be the set of Markov sources of order  $m$ ,  $m \geq 0$ . More precisely,  $p \in M_m(A)$  if

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

9 for all  $t \geq m$  and  $a_{i_1}, a_{i_2}, \dots \in A$ . By definition,  $M_0(A)$  is the set of all Bernoulli (or i.i.d.)  
 sources over  $A$ .

11 *2.1. Universal prediction*

Now we briefly describe some results and methods of universal coding and prediction,  
 13 which will be used later. Let a source generate a message  $x_1 \dots x_{t-1} x_t \dots$  and let  $v^t(a)$   
 14 denote the count of letter  $a$  occurring in the word  $x_1 \dots x_{t-1} x_t$ . After the first  $t$  letters  
 15  $x_1, \dots, x_{t-1}, x_t$  have been processed the following letter  $x_{t+1}$  is to be predicted. By defini-  
 16 tion, a prediction is a set of nonnegative numbers  $\gamma(a_1 | x_1 \dots x_t), \dots, \gamma(a_n | x_1 \dots x_t)$  which  
 17 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  $p(x_{t+1} = a_i | x_1 \dots x_t)$ ;  $i = 1, \dots, n$ .

19 Laplace suggested the following predictor:

$$L(a | x_1 \dots x_t) = (v^t(a) + 1) / (t + |A|), \tag{1}$$

21 where  $|A|$  is the number of letters in the alphabet  $A$ , see Feller (1970). For example, if  
 22  $A = \{0, 1\}$ ,  $x_1 \dots x_5 = 01010$ , then the Laplace prediction is as follows:  $L(x_6 = 0 | 01010) =$   
 23  $(3 + 1) / (5 + 2) = \frac{4}{7}$ ,  $L(x_6 = 1 | 01010) = (2 + 1) / (5 + 2) = \frac{3}{7}$ .

25 In Information Theory the error of prediction often is estimated by the Kullback–Leibler  
 (K–L) divergence between a distribution  $p$  and its estimate. Consider a source  $p$  and a  
 predictor  $\gamma$ . The error is characterized by the divergence

$$27 \rho_{\gamma,p}(x_1 \dots x_t) = \sum_{a \in A} p(a | x_1 \dots x_t) \log \frac{p(a | x_1 \dots x_t)}{\gamma(a | x_1 \dots x_t)}. \tag{2}$$

(Here and below  $\log \equiv \log_2$ .) It is well known that for any distributions  $p$  and  $\gamma$  the K–L  
 29 divergence is nonnegative and equals 0 if and only if  $p(a) = \gamma(a)$  for all  $a$ , see, for example,  
 30 Gallager (1968), that is why the K–L divergence is a natural estimate of the prediction error.  
 31 For a fixed  $t$ ,  $\rho_{\gamma,p}$  is a random variable, because  $x_1, x_2, \dots, x_t$  are random variables. We  
 define the average error at time  $t$  by

$$33 \rho^t(p \| \gamma) = E(\rho_{\gamma,p}(\cdot)) = \sum_{x_1 \dots x_t \in A^t} p(x_1 \dots x_t) \rho_{\gamma,p}(x_1 \dots x_t).$$

1 It is known that the error of the Laplace predictor goes to 0 for any Bernoulli source  $p$ . More  
precisely, it is proven that

$$3 \quad \rho^t(p \| L) < (|A| - 1)/(t + 1) \tag{3}$$

for any source  $p$ ; Ryabko (1990) (see also Ryabko and Topsoe, 2002).

5 Obviously, the convergence to 0 of a predictor's error for any source from some set  $M$   
is an important property. For example, we can see from (3) that it is true for the Laplace  
7 predictor and the set of Bernoulli sources  $M_0(A)$ . Unfortunately, it is known that a predictor,  
for which error (2) goes to 0 (with probability 1) for any stationary and ergodic source, does  
9 not exist. More precisely, for any predictor  $\gamma$  there exists a stationary and ergodic source  $\tilde{p}$ ,  
such that  $\limsup_{t \rightarrow \infty} \rho_{\gamma, \tilde{p}}(x_1 \dots x_t) \geq \text{const} > 0$  with probability 1; Ryabko (1988). (See  
11 also Algoet, 1999; Morvai et al., 1997; Nobel, 2003, where this result is generalized and  
the history of its discovery is described. In particular, they found out that such a result was  
13 described by Bailey, 1976 in his unpublished thesis.) That is why it is difficult to use (2) for  
comparison of different predictors. On the other hand, it is shown in Ryabko (1984, 1988)  
15 that there exists a predictor  $R$ , such that the following average  $t^{-1} \sum_{i=1}^t \rho_{R,p}(x_1 \dots x_i)$  goes  
to 0 (with probability 1) for any stationary and ergodic source  $p$ , where  $t$  goes to infinity.  
17 That is why we will focus our attention on such averages. First, we define for any predictor  
 $\gamma$  the following probability distribution:

$$19 \quad \gamma(x_1 \dots x_t) = \prod_{i=1}^t \gamma(x_i | x_1 \dots x_{i-1}).$$

For example, we obtain for the Laplace predictor  $L$  that  $L(0101) = \frac{1}{2} \frac{1}{3} \frac{1}{2} \frac{2}{5} = \frac{1}{30}$ , see (1).  
21 Then, by analogy with (2) we will estimate the error by K–L divergence and define

$$\bar{\rho}_{\gamma,p}(x_1 \dots x_t) = t^{-1} (\log(p(x_1 \dots x_t)/\gamma(x_1 \dots x_t))) \tag{4}$$

23 and

$$\bar{\rho}_t(\gamma, p) = t^{-1} \sum_{x_1 \dots x_t \in A^t} p(x_1 \dots x_t) \log(p(x_1 \dots x_t)/\gamma(x_1 \dots x_t)). \tag{5}$$

25 For example, from those definitions and (3) we obtain the following bound for the Laplace  
predictor  $L$  and any Bernoulli source  $p$ :  $\bar{\rho}_t(L, p) < ((|A| - 1) \log t + c)/t$ , where  $c$  is a  
27 constant.

The universal predictors will play a key role in the tests suggested below. By definition,  
29 a predictor  $\gamma$  is called *universal* (in average) for a class of sources  $M$ , if for any  $p \in M$   
the error  $\bar{\rho}_t(\gamma, p)$  goes to 0, when  $t$  goes to infinity. A predictor  $\gamma$  is called *universal with*  
31 *probability 1*, if the error  $\bar{\rho}_{\gamma,p}(x_1 \dots x_t)$  goes to 0 not only in average, but for almost all  
sequences  $x_1 x_2 \dots$ . In short, we will say that the predictor (or probability distribution)  $\gamma$  is  
33 *universal*, if  $\lim_{t \rightarrow \infty} \bar{\rho}_{\gamma,p}(x_1 \dots x_t) = 0$  is valid with probability 1 for any stationary and  
ergodic source  $p$  (i.e. for any  $p \in M_\infty(A)$ ). Now there are quite many known universal  
35 predictors. One of the first such predictors has been described in Ryabko (1984, 1988).

1 2.2. Universal coding

3 This short subparagraph is intended to give some explanation about why and how methods  
 4 of data compression can be used for testing of independence. The point is that the prediction  
 5 problem is deeply connected with the theory of universal coding. Moreover, practically used  
 6 data compression methods (or so-called archivers) can be directly applied for testing.

7 Let us give some definitions. Let, as before,  $A$  be a finite alphabet and, by definition,  
 8  $A^* = \bigcup_{n=1}^{\infty} A^n$  and  $A^{\infty}$  is the set of all infinite words  $x_1x_2 \dots$  over the alphabet  $A$ . A data  
 9 compression method (or code)  $\varphi$  is defined as a set of mappings  $\varphi_n$  such that  $\varphi_n : A^n \rightarrow$   
 10  $\{0, 1\}^*$ ,  $n = 1, 2, \dots$  and for each pair of different words  $x, y \in A^n$   $\varphi_n(x) \neq \varphi_n(y)$ .  
 11 Informally, it means that the code  $\varphi$  can be applied for compression of each message of any  
 12 length  $n$ ,  $n > 0$ , over alphabet  $A$  and the message can be decoded if its code is known. Further,  
 13 it is required that each sequence  $\varphi_n(x_1)\varphi_n(x_2) \dots \varphi_n(x_r)$ ,  $r \geq 1$ , of encoded words from  
 14 the set  $A^n$ ,  $n \geq 1$ , can be uniquely decoded into  $x_1x_2 \dots x_r$ . Such codes are called uniquely  
 15 decodable. For example, let  $A = \{a, b\}$ , the code  $\psi_1(a) = 0, \psi_1(b) = 00$ , obviously, is not  
 16 uniquely decodable. (Indeed, the word 000 can be decoded in both  $ab$  and  $ba$ .) It is well  
 17 known that if a code  $\varphi$  is uniquely decodable then the lengths of the codewords satisfy the  
 following inequality (the Kraft inequality):

$$\sum_{u \in A^n} 2^{-|\varphi_n(u)|} \leq 1,$$

19 see, for example, Gallager (1968). It will be convenient to reformulate this property as  
 follows:

21 **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*

$$23 \quad -\log \mu_{\varphi}(u) \leq |\varphi(u)| \tag{6}$$

24 *for any  $u$  from  $A^n$ . (Obviously, it is true for the measure  $\mu_{\varphi}(u) = 2^{-|\varphi(u)|} / \sum_{u \in A^n} 2^{-|\varphi(u)|}$ .)  
 25 It is well known that sequences  $x_1 \dots x_t$ , generated by a stationary and ergodic source  $p$ ,  
 26 can be “compressed” till the length  $-\log p(x_1 \dots x_t)$  bits. There exist so-called universal  
 27 codes, which, in a certain sense, are the best “compressors” for all stationary and ergodic  
 28 sources. The formal definition is as follows: a code  $\varphi$  is universal if for any stationary and  
 29 ergodic source  $p$*

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

31 *with probability 1. So, informally speaking, the universal codes estimate the probability  
 characteristics of the source  $p$  and use them for efficient “compression”.*

33 3. The tests

35 In this paragraph we describe the suggested tests. First, we give some definitions. Let  
 $v$  be a word  $v = v_1 \dots v_k, k \leq t, v_i \in A$ . Denote the rate of a word  $v$  occurring in the

1 sequence  $x_1 x_2 \dots x_k, x_2 x_3 \dots x_{k+1}, x_3 x_4 \dots x_{k+2}, \dots, x_{t-k+1} \dots x_t$  by  $v^t(v)$ . For example,  
 2 if  $x_1 \dots x_t = 000100$  and  $v = 00$ , then  $v^6(00) = 3$ . Now we define for any  $k \geq 0$  the so-called  
 3 empirical Shannon entropy of order  $k$  as follows:

$$h_k^*(x_1 \dots x_t) = -\frac{1}{(t-k)} \sum_{v \in A^k} \bar{v}^t(v) \sum_{a \in A} (v^t(va)/\bar{v}^t(v)) \log(v^t(va)/\bar{v}^t(v)), \quad (7)$$

5 where  $k < t$  and  $\bar{v}^t(v) = \sum_{a \in A} v^t(va)$ . In particular, if  $k = 0$ , we obtain

$$h_0^*(x_1 \dots x_t) = -\frac{1}{t} \sum_{a \in A} v^t(a) \log(v^t(a)/t).$$

7 The suggested test is as follows.

Let  $\gamma$  be any probability distribution over  $A^t$ . The hypothesis  $H_0$  is accepted if

$$9 \quad (t-m)h_m^*(x_1 \dots x_t) - \log(1/\gamma(x_1 \dots x_t)) \leq \log(1/\alpha), \quad (8)$$

where  $0 < \alpha < 1$ . Otherwise,  $H_0$  is rejected. We denote this test by  $\Upsilon_{\alpha, \gamma, m}^t$ .

11 **Theorem.** (i) For any predictor (or measure)  $\gamma$  the Type I error of the test  $\Upsilon_{\alpha, \gamma, m}^t$  is less  
 than or equal to  $\alpha$ ,  $\alpha \in (0, 1)$ .

13 (ii) If  $\gamma$  is a universal predictor (measure) (i.e., by definition, for any  $p \in M_\infty(A)$ )

$$\lim_{t \rightarrow \infty} t^{-1}(-\log p(x_1 \dots x_t) - \log(1/\gamma(x_1 \dots x_t))) = 0 \quad (9)$$

15 with probability 1), then the Type II error goes to 0, when  $t$  goes to infinity.

The proof is given in Appendix.

17 **Comment.** Let  $\varphi$  be a uniquely decodable code (or a data compression method). Define  
 the test  $\hat{\Upsilon}_{\alpha, \varphi, m}^t$  as follows: The hypothesis  $H_0$  is accepted if

$$19 \quad (t-m)h_m^*(x_1 \dots x_t) - |\varphi(x_1 \dots x_t)| \leq \log(1/\alpha), \quad (10)$$

where  $\alpha \in (0, 1)$ . Otherwise,  $H_0$  is rejected.

21 We immediately obtain from the Theorem 1 and the Claim 1 the following statement.

23 **Claim 2.** (i) For any uniquely decodable code  $\varphi$  the Type I error of the test  $\hat{\Upsilon}_{\alpha, \varphi, m}^t$  is less  
 than or equal to  $\alpha$ ,  $\alpha \in (0, 1)$ .

(ii) If  $\varphi$  is a universal code, then the Type II error goes to 0, where  $t$  goes to infinity.

#### 25 4. Conclusion

The tests described above can be based on known universal codes (or so-called archivers)  
 27 which are widely used for text compression everywhere. It is important to note that, on the  
 one hand, the universal codes and archivers are based on results of Information Theory, the  
 29 theory of algorithms and some other branches of mathematics; see, for example, [Rissanen](#)

1 (1984), Kieffer (1998), Kieffer and Yang (2000), Effros et al. (2002). On the other hand, the  
 3 archivers have shown high efficiency in practice as compressors of texts, DNA sequences  
 5 and many other types of real data. In fact, the archivers can find many kinds of latent  
 regularities, which is why they look like a promising tool for independence testing and its  
 generalizations.

The natural question is the possibility of generalizing the suggested tests for the case of  
 an infinite source alphabet  $A$  (say,  $A$  is a metric space.) Apparently, such a generalization  
 can be done for the case of independence testing, if we will use a known technique of  
 partitioning; see Darbellay and Vajda (1998, 1999). But we do not know how to generalize  
 the suggested tests for the case where  $H_0$  is that the source is Markovian. The point is that  
 the partitioning can increase the source order. For example, even if the alphabet  $A$  contains  
 three letters and we combine two of them in one subset (i.e. a new letter) the order of the  
 obtained source can increase till infinity. Hence, the generalization to Markov sources with  
 infinite alphabet can be considered as an open problem.

15 **Appendix**

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

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

19 Indeed, the equality is true, because  $\tau^*$  is a Bernoulli measure. The inequality follows from  
 the well-known inequality  $\sum_{a \in A} p(a) \log(p(a)/q(a)) \geq 0$ , for K–L divergence, which is  
 21 true for any distributions  $p$  and  $q$  (see, for example, Gallager (1968)). So, if  $p(a) = v^t(a)/t$   
 and  $q(a) = \tau^*(a)$ , then

$$\sum_{a \in A} \frac{v^t(a)}{t} \log \frac{(v^t(a)/t)}{\tau^*(a)} \geq 0.$$

From the last inequality we obtain (11).

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

$$\tau(x_1 \dots x_t) \leq \prod_{u \in A^m} \prod_{a \in A} (v^t(ua)/\bar{v}^t(u))^{v^t(ua)}. \tag{12}$$

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

$$\tau(x_1 \dots x_t) = \tau_\infty(x_1 \dots x_m) \prod_{u \in A^m} \prod_{a \in A} \tau(a/u)^{v^t(ua)},$$

29 where  $\tau_\infty(x_1 \dots x_m)$  is the limit probability of the word  $x_1 \dots x_m$ . From the last equality  
 we can see that

$$\tau(x_1 \dots x_t) \leq \prod_{u \in A^m} \prod_{a \in A} \tau(a/u)^{v^t(ua)}.$$

31

1 Taking into account inequality (11), we obtain

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

3 for any word  $u$ . So, from the last two inequalities we obtain (12).

It will be convenient to define an auxiliary measure on  $A^t$  as follows:

$$5 \quad \pi_m(x_1 \dots x_t) = \Delta 2^{-(t-m)h_m^*(x_1 \dots x_t)}, \tag{13}$$

where  $x_1 \dots x_t \in A^t$  and  $\Delta = (\sum_{x_1 \dots x_t \in A^t} 2^{-(t-m)h_m^*(x_1 \dots x_t)})^{-1}$ . If we take into account  
7 that  $2^{-(t-m)h_m^*(x_1 \dots x_t)} = \prod_{u \in A^m} \prod_{a \in A} (v^t(ua)/\bar{v}^t(u))^{v^t(ua)}$ , we can see from (12) and (13)  
that, for any measure  $\tau \in M_m(A)$  and any  $x_1 \dots x_t \in A^t$ ,

$$9 \quad \tau(x_1 \dots x_t) \leq \pi_m(x_1 \dots x_t)/\Delta. \tag{14}$$

Let us denote the critical set of the test  $\mathcal{Y}_{\alpha, \gamma, m}^t$  as  $C_\alpha$  i.e., by definition,

$$11 \quad C_\alpha = \{x_1 \dots x_t : (t-m)h_m^*(x_1 \dots x_t) - \log(1/\gamma(x_1 \dots x_t)) > \log(1/\alpha)\}. \tag{15}$$

From (14) and this definition we can see that for any measure  $\tau \in M_m(A)$

$$13 \quad \tau(C_\alpha) \leq \pi_m(C_\alpha)/\Delta. \tag{16}$$

From definitions (15) and (13) we obtain

$$15 \quad \begin{aligned} C_\alpha &= \{x_1 \dots x_t : 2^{(t-m)h_m^*(x_1 \dots x_t)} > (\alpha\gamma(x_1 \dots x_t))^{-1}\} \\ &= \{x_1 \dots x_t : (\pi_m(x_1 \dots x_t)/\Delta)^{-1} > (\alpha\gamma(x_1 \dots x_t))^{-1}\}. \end{aligned}$$

Finally,

$$17 \quad C_\alpha = \{x_1 \dots x_t : \gamma(x_1 \dots x_t) > \pi_m(x_1 \dots x_t)/(\alpha\Delta)\}. \tag{17}$$

The following inequalities and equalities are valid:

$$19 \quad \begin{aligned} 1 &\geq \sum_{x_1 \dots x_t \in C_\alpha} \gamma(x_1 \dots x_t) \geq \sum_{x_1 \dots x_t \in C_\alpha} \pi_m(x_1 \dots x_t)/(\alpha\Delta) \\ &= \pi_m(C_\alpha)/(\alpha\Delta) \geq \tau(C_\alpha)\Delta/(\alpha\Delta) = \tau(C_\alpha)/\alpha. \end{aligned}$$

(Here both equalities and the first inequality are obvious, the second inequality and the  
21 third one follow from (17) and (16), correspondingly.) So, we obtain that  $\tau(C_\alpha) \leq \alpha$  for any  
23 measure  $\tau \in M_m(A)$ . Taking into account that  $C_\alpha$  is the critical set of the test, we can see  
that the probability of the Type I error is not greater than  $\alpha$ . The first claim of the theorem  
is proven.

25 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:

$$27 \quad h_t(p) = - \sum_{x_1 \dots x_t \in A^t} p(x_1 \dots x_t) \sum_{a \in A} p(a/x_1 \dots x_t) \log p(a/x_1 \dots x_t), \tag{18}$$

where  $p \in M_\infty(A)$ . It is known that for any  $p \in M_\infty(A)$  firstly,  $\log |A| \geq$   
29  $h_0(p) \geq h_1(p) \geq \dots$ , secondly, the following limit Shannon entropy  $h_\infty(p) = \lim_{t \rightarrow \infty} h_t(p)$



1 exists, thirdly,  $\lim_{t \rightarrow \infty} -t^{-1} \log p(x_1 \dots x_t) = h_\infty(p)$  with probability 1 and, finally,  $h_m(p)$   
 2 is strictly greater than  $h_\infty(p)$ , if the memory of  $p$  is larger  $m$ , (i.e.  $p \in M_\infty(A) \setminus M_m(A)$ ),  
 3 see, for example, Billingsley (1965), Gallager (1968).

4 Taking into account the definition of the universal predictor (see (9)), we obtain from the  
 5 above described properties of the entropy that

$$\lim_{t \rightarrow \infty} -t^{-1} \log \gamma(x_1 \dots x_t) = h_\infty(p) \quad (19)$$

7 with probability 1. It can be seen that  $h_m^*$  (7) is a consistent estimate for the  $m$ - order  
 8 Shannon entropy (18), i.e.  $\lim_{t \rightarrow \infty} h_m^*(x_1 \dots x_t) = h_m(p)$  with probability 1; see Billingsley  
 9 (1965), Gallager (1968). Having taken into account that  $h_m(p) > h_\infty(p)$  and (19) we obtain  
 10 from the last equality that  $\lim_{t \rightarrow \infty} ((t - m) h_m^*(x_1 \dots x_t) - \log(1/\gamma(x_1 \dots x_t))) = \infty$ . This  
 11 proves the second statement of the theorem.  $\square$

## References

- 13 Algoet, P., 1999. Universal schemes for learning the best nonlinear predictor given the infinite past and side  
 information. *IEEE Trans. Inform. Theory* 45, 1165–1185.
- 15 Bailey, D.H., 1976. Sequential schemes for classifying and predicting ergodic processes. Ph.D. Dissertation,  
 Stanford University.
- 17 Billingsley, P., 1965. *Ergodic Theory and Information*. Wiley, New York.
- 18 Blum, J.R., Kiefer, J., Rosenblatt, M., 1961. Distribution free tests of independence based on the sample distribution  
 19 function. *Ann. Math. Statist.* 32, 485–498.
- 20 Csiszár, I., Shields, P., 2000. The consistency of the BIC Markov order estimation. *Ann. Statist.* 6, 1601–1619.
- 21 Darbellay, G.A., Vajda, I., 1998. Entropy expressions for multivariate continuous distributions. Research Report  
 No. 1920, UTIA, Academy of Science, Prague (library@utia.cas.cz).
- 23 Darbellay, G.A., Vajda, I., 1999. Estimation of the mutual information with data-dependent partitions. *IEEE Trans.*  
*Inform. Theory* 48 (5), 1061–1081.
- 25 Effros, M., Visweswariah, K., Kulkarni, S.R., Verdu, S., 2002. Universal lossless source coding with the Burrows  
 Wheeler transform. *IEEE Trans. Inform. Theory* 48 (5), 1061–1081.
- 27 Feller, W., 1970. *An Introduction to Probability Theory and its Applications*, vol. 1. Wiley, New York.
- 28 Gallager, R.G., 1968. *Information Theory and Reliable Communication*. Wiley, New York.
- 29 Ghoudi, K., Kulperger, R.J., Remillard, B., 2001. A nonparametric test of serial independence for time series and  
 residuals. *J. Multivariate Anal.* 79 (2), 191–218.
- 31 Hoeffding, W., 1948. A nonparametric test of independence. *Ann. Math. Statist.* 19, 546–557.
- 32 Kendall, M.G., Stuart, A., 1961. *The Advanced Theory of Statistics*, vol. 2. Inference and Relationship. Charles  
 33 Griffin, London.
- 34 Kieffer, J., 1998. Prediction and Information Theory, Preprint (available at  
 35 <ftp://oz.ee.umn.edu/users/kieffer/papers/prediction.pdf>).
- 36 Kieffer, J.C., Yang, E.H., 2000. Grammar-based codes: a new class of universal lossless source codes. *IEEE Trans.*  
 37 *Inform. Theory* 46 (3), 737–754.
- 38 Morvai, G., Yakowitz, S.J., Algoet, P.H., 1997. Weakly convergent nonparametric forecasting of stationary time  
 39 series. *IEEE Trans. Inform. Theory* 43, 483–498.
- 40 Nobel, A.B., 2003. On optimal sequential prediction. *IEEE Trans. Inform. Theory* 49 (1), 83–98.
- 41 Rissanen, J., 1984. Universal coding, information, prediction, and estimation. *IEEE Trans. Inform. Theory* 30 (4),  
 629–636.
- 43 Ryabko, B., Monarev, V., 2005. Using information theory approach to randomness testing. *J. Statist. Plann.*  
*Inference* 133 (1), 95–110.
- 45 Ryabko, B., Topsoe, F., 2002. On asymptotically optimal methods of prediction and adaptive coding for Markov  
 sources. *J. Complexity* 18 (1), 224–241.

- 1 Ryabko, B. Ya., 1984. Twice-universal coding. *Problems Inform. Transmission* 20 (3),
- 3 Ryabko, B. Ya., 1988. Prediction of random sequences and universal coding. *Problems Inform. Transmission* 24  
(2), 87–96.
- Ryabko, B. Ya., 1990. A fast adaptive coding algorithm. *Problems Inform. Transmission* 26 (4), 305–317.

UNCORRECTED PROOF