

# Testing Statistical Hypotheses About Ergodic Processes

Daniil Ryabko\*, Boris Ryabko#

\*SequeL, INRIA-Lille Nord Europe, France, daniil@ryabko.net

#Institute of Computational Technologies of Siberian Branch of Russian Academy of Science, Siberian State University of Telecommunications and Informatics, Novosibirsk, Russia; boris@ryabko.net

## Abstract

We propose a method for statistical analysis of time series, that allows us to obtain solutions to some classical problems of mathematical statistics under the only assumption that the process generating the data is stationary ergodic. Namely, we consider three problems: goodness-of-fit (or identity) testing, process classification, and the change point problem. For each of the problems we construct a test that is asymptotically accurate for the case when the data is generated by stationary ergodic processes. The tests are based on empirical estimates of distributional distance.

Keywords: non-parametric hypothesis testing, stationary ergodic processes, goodness-of-fit test, process classification, change point problem

## 1 Introduction

<sup>1</sup> **Overview.** In this work we consider the problem of statistical analysis of time series, when nothing is known about the underlying process generating the data, except that it is stationary ergodic. There is a vast literature on time series analysis under various parametric assumptions, and also under such non-parametric assumptions as that the processes is finite-memory or has certain mixing rates. While under these settings most of the problems of statistical analysis are clearly solvable and efficient algorithms exist, in the general setting of stationary ergodic processes it is far less clear what can be done in principle, which problems of statistical analysis admit a solution and which do not. In this work we propose a method of statistical analysis of time series, that allows us to demonstrate that some classical statistical problems indeed admit a solution under the only assumption that the data is stationary ergodic, whereas before

---

<sup>1</sup>Some preliminary results appear in [16].

solutions only for more restricted cases were known. The solutions are always constructive, that is, we present asymptotically accurate algorithms for each of the considered problems. All the algorithms are based on empirical estimates of distributional distance, which is in the core of the suggested approach. We suggest that the proposed approach can be applied to other problems of statistical analysis of time series, with the view of establishing principled positive results, leaving the task of finding optimal algorithms for each particular problem as a topic for further research.

Here we concentrate on the following three problems: goodness-of-fit (or identity) testing, process classification, and the change point problem.

**Identity testing.** The first problem is the following problem of hypothesis testing. A stationary ergodic process distribution  $\rho$  is known theoretically. Given a data sample, it is required to test whether it was generated by  $\rho$ , versus it was generated by any other stationary ergodic distribution that is different from  $\rho$  (goodness-of-fit, or identity testing). The case of i.i.d. or finite-memory processes was widely studied (see e.g. [6]); in particular, when  $\rho$  has a finite memory [15] proposes a test against any stationary ergodic alternative: a test that can be based on an arbitrary universal code. It was noted in [17] that an asymptotically accurate test for the case of stationary ergodic processes over finite alphabet exists (but no test was proposed). Here we propose a concrete and simple asymptotically accurate goodness-of-fit test, which demonstrates the proposed approach: to use empirical distributional distance for hypotheses testing. By asymptotically accurate test we mean the following. First, the Type I error of the test (or its size) is fixed and is given as a parameter to the test. That is, given any  $\alpha > 0$  as an input, under  $H_0$  (that is, if the data sample was indeed generated by  $\rho$ ) the probability that the test says “ $H_1$ ” is not greater than  $\alpha$ . Second, under any hypothesis in  $H_1$  (that is, if the distribution generating the data is different from  $\rho$ ), the test will say “ $H_0$ ” not more than a finite number of times, with probability 1. In other words, the Type I error of the test is fixed and the Type II error can be made not more than a finite number of times, as the data sample increases, with probability 1 under any stationary ergodic alternative.

A comment on this setting is in order. When the alternative  $H_1$  is less general, e.g. distributions that have finite-memory [10] or known mixing rates, one typically seeks a test that has optimal rates of decrease of probability of Type II error to 0. For our case, when the alternative is the set of all stationary ergodic processes, the rate of decrease of probability Type II error is necessarily non-uniform. In this sense, the property that we establish for our test is the strongest possible. Observe that it is strictly stronger than requiring that the test makes only a finite number of errors (either Type I or Type II), the setting considered, for example, in the cases when  $H_0$  is composite, or for the process classification problem that we address in this work.

**Process classification.** In the next problem that we consider, we again have to decide whether a data sample was generated by a process satisfying a hypothesis  $H_0$  or a hypothesis  $H_1$ . However, here  $H_0$  and  $H_1$  are not known theoretically, but are represented by two additional data samples. More pre-

cisely, the problem is that of process classification, which can be formulated as follows. We are given three samples  $X = (X_1, \dots, X_k)$ ,  $Y = (Y_1, \dots, Y_m)$  and  $Z = (Z_1, \dots, Z_n)$  generated by stationary ergodic processes with distributions  $\rho_X$ ,  $\rho_Y$  and  $\rho_Z$ . It is known that  $\rho_X \neq \rho_Y$  but either  $\rho_Z = \rho_X$  or  $\rho_Z = \rho_Y$ . It is required to test which one is the case. That is, we have to decide whether the sample  $Z$  was generated by the same process as the sample  $X$  or by the same process as the sample  $Y$ . This problem for the case of dependent time series was considered for example in [10], where a solution is presented under the finite-memory assumption. It is closely related to many important problems in statistics and application areas, such as pattern recognition, classification, etc. Apparently no asymptotically accurate procedure for process classification has been known so far for the general case of stationary ergodic processes. Here we propose a test that converges almost surely to the correct answer. In other words, the test makes only a finite number of errors with probability 1, with respect to any stationary ergodic processes generating the data. Unlike in the previous problem, here we do not explicitly distinguish between Type I and Type II error, since the hypotheses are by nature symmetric:  $H_0$  is “ $\rho_Z = \rho_X$ ” and  $H_1$  is “ $\rho_Z = \rho_Y$ ”.

**Change point estimation.** Finally, we consider the change point problem. It is another classical problem, with vast literature on both parametric (see e.g. [2]) and non-parametric (see e.g. [5]) methods for solving it. In this work we address the case where the data is dependent, its form and the structure of dependence is unknown, and marginal distributions before and after the change may be the same. We consider the following (off-line) setting of the problem: a (real-valued) sample  $Z_1, \dots, Z_n$  is given, where  $Z_1, \dots, Z_k$  are generated according to some distribution  $\rho_X$  and  $Z_{k+1}, \dots, Z_n$  are generated according to some distribution  $\rho_Y$  which is different from  $\rho_X$ . It is known that the distributions  $\rho_X$  and  $\rho_Y$  are stationary ergodic, but nothing else is known about them. Most literature on change point problem for dependent time series assumes that the marginal distributions before and after the change point are different, and often also make explicit restrictions on the dependence, such as requirements on mixing rates. Nonparametric methods used in these cases are typically based on Kolmogorov-Smirnov statistic, Cramer-von Mises statistic, or generalizations thereof [5, 3, 8]. The main difference of our results is that we do not assume that the single-dimensional marginals (or finite-dimensional marginals of any given fixed size) are different, and do not make any assumptions on the structure of dependence. The only assumption is that the (unknown) process distributions before and after the change point are stationary ergodic. Our result is a demonstration of that asymptotically accurate change point estimation is possible in this general setting.

**Methodology.** All the tests that we construct are based on empirical estimates of the so-called distributional distance. For two processes  $\rho_1, \rho_2$  a distributional distance is defined as  $\sum_{k=1}^{\infty} w_k |\rho_1(B_k) - \rho_2(B_k)|$ , where  $w_k$  are positive summable real weights, e.g.  $w_k = 2^{-k}$  and  $B_k$  range over a countable field that generates the sigma-algebra of the underlying probability space. For example, if we are talking about finite-alphabet processes with the binary alphabet

$A = \{0, 1\}$ ,  $B_k$  would range over the set  $A^* = \cup_{k \in \mathbb{N}} A^k$ ; that is, over all tuples  $0, 00, 01, 10, 000, 001, \dots$ ; therefore, the distributional distance in this case is the weighted sum of differences of probabilities of all possible tuples. In this work we consider real-valued processes,  $A = \mathbb{R}$ , so  $B_k$  can be taken to range over all intervals with rational endpoints, all pairs of such intervals, triples, etc. Although distributional distance is a natural concept that, for stochastic processes, has been studied for a while [9], its empirical estimates have not, to our knowledge, been used for statistical analysis of time series. We argue that this distance is rather natural for this kind of problems, first of all, since it can be consistently estimated (unlike, for example,  $\bar{d}$  distance, which cannot [13] be consistently estimated for the general case of stationary ergodic processes). Secondly, it is always bounded, unlike (empirical) KL divergence, which is often used for statistical inference for time series (e.g. [6, 15, 1, 7, 12] and others). Other approaches to statistical analysis of stationary dependent time series include the use of (universal) codes [11, 15, 14]. Here we first show that distributional distance between stationary ergodic processes can be consistently estimated based on sampling, and then apply it to construct a consistent test for the three problems of statistical analysis described above.

Although empirical estimates of the distributional distance involve taking an infinite sum, in practice it is obvious that only a finite number of summands has to be calculated. This is due to the fact that empirical estimates have to be compared to each other or to theoretically known probabilities, and since the (bounded) summands have (exponentially) decreasing weights, the result of the comparison is known after only finitely many evaluations. Therefore, the algorithms presented can be applied in practice. On the other hand, the main value of the results is in the demonstration of what is possible in principle; finding practically efficient procedures for each of the considered problems is an interesting problem for further research.

## 2 Preliminaries

We are considering (stationary ergodic) processes with the alphabet  $A = \mathbb{R}$ . The generalization to  $A = \mathbb{R}^d$  is straightforward; moreover, the results can be extended to the case when  $A$  is a complete separable metric space. We use the symbol  $A^*$  for  $\cup_{i=1}^{\infty} A^i$ . Elements of  $A^*$  are called words or sequences. For each  $k \in \mathbb{N}$ , let  $B^k$  be the set of all cylinders of the form  $A_1 \times \dots \times A_k$  where  $A_i \subset A$  are intervals with rational endpoints. Let  $\mathcal{B} = \cup_{k=1}^{\infty} B^k$ ; since this set is countable we can introduce an enumeration  $\mathcal{B} = \{B_i : i \in \mathbb{N}\}$ . The set  $\{B_i \times A^\infty : i \in \mathbb{N}\}$  generates the Borel  $\sigma$ -algebra on  $\mathbb{R}^\infty = A^\infty$ . For a set  $B \in \mathcal{B}$  let  $|B|$  be the index  $k$  of the set  $B^k$  that  $B$  comes from:  $|B| = k : B \in B^k$ .

For a sequence  $X \in A^n$  and a set  $B \in \mathcal{B}$  denote  $\nu(X, B)$  the frequency with which the sequence  $X$  falls in the set  $B$

$$\nu(X, B) := \begin{cases} \frac{1}{n-|B|+1} \sum_{i=1}^{n-|B|+1} I_{\{(X_i, \dots, X_{i+|B|-1}) \in B\}} & \text{if } n \geq |B|, \\ 0 & \text{otherwise} \end{cases}$$

where  $X = (X_1, \dots, X_n)$ . For example,

$$\nu((0.5, 1.5, 1.2, 1.4, 2.1), ([1.0, 2.0] \times [1.0, 2.0])) = 1/2.$$

We use the symbol  $\mathcal{S}$  for the set of all stationary ergodic processes on  $A^\infty$ . The ergodic theorem (see e.g. [4]) implies that for any process  $\rho \in \mathcal{S}$  generating a sequence  $X_1, X_2, \dots$  the frequency of observing a tuple that falls into each  $B \in \mathcal{B}$  tends to its limiting (or a priori) probability a.s.:

$$\nu((X_1, \dots, X_n), B) \rightarrow \rho((X_1, \dots, X_{|B|}) \in B)$$

as  $n \rightarrow \infty$ . We will often abbreviate  $\rho((X_1, \dots, X_{|B|}) \in B) =: \rho(B)$ .

**Definition 1** (distributional distance). *The distributional distance is defined for a pair of processes  $\rho_1, \rho_2$  as follows [9]:*

$$d(\rho_1, \rho_2) = \sum_{i=1}^{\infty} w_i |\rho_1(B_i) - \rho_2(B_i)|, \quad (1)$$

where  $w_i$  are summable positive real weights (e.g.  $w_k = 2^{-k}$ ).

It is easy to see that  $d$  is a metric. The reader is referred to [9] for more information about  $d$  and its properties.

**Definition 2** (empirical distributional distance). *For  $X, Y \in A^*$ , define empirical distributional distance  $\hat{d}(X, Y)$  as*

$$\hat{d}(X, Y) := \sum_{i=1}^{\infty} w_i |\nu(X, B_i) - \nu(Y, B_i)|. \quad (2)$$

Similarly, we can define the empirical distance when only one of the process measures is unknown:

$$\hat{d}(X, \rho) := \sum_{i=1}^{\infty} w_i |\nu(X, B_i) - \rho(B_i)|, \quad (3)$$

where  $\rho \in \mathcal{S}$  and  $X \in A^*$ .

The following lemma will play a key role in establishing the main results.

**Lemma 1.** *Let two samples  $X = (X_1, \dots, X_k)$  and  $Y = (Y_1, \dots, Y_m)$  be generated by stationary ergodic processes  $\rho_X$  and  $\rho_Y$  respectively. Then*

$$(i) \lim_{k, m \rightarrow \infty} \hat{d}(X, Y) = d(\rho_X, \rho_Y) \text{ a.s.}$$

$$(ii) \lim_{k \rightarrow \infty} \hat{d}(X, \rho_Y) = d(\rho_X, \rho_Y) \text{ a.s.}$$

*Proof.* For any  $\varepsilon > 0$  we can find such an index  $J$  that  $\sum_{i=J}^{\infty} w_i < \varepsilon/2$ . Moreover, for each  $j$  we have  $\nu((X_1, \dots, X_k), B_j) \rightarrow \rho_X(B_j)$  a.s., so that

$$|\nu((X_1, \dots, X_k), B_j) - \rho(B_j)| < \varepsilon/(4Jw_j)$$

from some step  $k$  on; define  $K_j := k$ . Let  $K := \max_{j < J} K_j$  ( $K$  depends on the realization  $X_1, X_2, \dots$ ). Define analogously  $M$  for the sequence  $(Y_1, \dots, Y_m, \dots)$ . Thus for  $k > K$  and  $m > M$  we have

$$\begin{aligned} |\hat{d}(X, Y) - d(\rho_X, \rho_Y)| &= \\ & \left| \sum_{i=1}^{\infty} w_i (|\nu(X, B_i) - \nu(Y, B_i)| - |\rho_X(B_i) - \rho_Y(B_i)|) \right| \\ & \leq \sum_{i=1}^{\infty} w_i (|\nu(X, B_i) - \rho_X(B_i)| + |\nu(Y, B_i) - \rho_Y(B_i)|) \\ & \leq \sum_{i=1}^J w_i (|\nu(X, B_i) - \rho_X(B_i)| + |\nu(Y, B_i) - \rho_Y(B_i)|) + \varepsilon/2 \\ & \leq \sum_{i=1}^J w_i (\varepsilon/(4Jw_i) + \varepsilon/(4Jw_i)) + \varepsilon/2 = \varepsilon, \end{aligned}$$

which proves the first statement. The second statement can be proven analogously.  $\square$

### 3 Main results

#### 3.1 Goodness-of-fit Test

For a given stationary ergodic process measure  $\rho$  and a sample  $X = (X_1, \dots, X_n)$  we wish to test the hypothesis  $H_0$  that the sample was generated by  $\rho$  versus  $H_1$  that it was generated by a stationary ergodic distribution that is different from  $\rho$ . Thus,  $H_0 = \{\rho\}$  and  $H_1 = \mathcal{S} \setminus H_0$ .

Define the set  $D_\delta^n$  as the set of all samples of length  $n$  that are at least  $\delta$ -far from  $\rho$  in empirical distributional distance:

$$D_\delta^n := \{X \in A^n : \hat{d}(X, \rho) \geq \delta\}.$$

For each  $n$  and each given confidence level  $\alpha$  define the critical region  $C_\alpha^n$  of the test as  $C_\alpha^n := D_\gamma^n$  where

$$\gamma := \inf\{\delta : \rho(D_\delta^n) \leq \alpha\}.$$

The test rejects  $H_0$  at confidence level  $\alpha$  if  $(X_1, \dots, X_n) \in C_\alpha^n$  and accepts it otherwise. In words, for each sequence we measure the distance between the empirical probabilities (frequencies) and the measure  $\rho$  (that is, the theoretical  $\rho$ -probabilities); we then take a largest ball (with respect to this distance) around  $\rho$  that has  $\rho$ -probability not greater than  $1 - \alpha$ . The test rejects all sequences outside this ball.

**Definition 3** (Goodness-of-fit test). For each  $n \in \mathbb{N}$  and  $\alpha \in (0, 1)$  the goodness-of-fit test  $G_n^\alpha : A^n \rightarrow \{0, 1\}$  is defined as

$$G_n^\alpha(X_1, \dots, X_n) := \begin{cases} 1 & \text{if } (X_1, \dots, X_n) \in C_\alpha^n, \\ 0 & \text{otherwise.} \end{cases}$$

**Theorem 1.** The test  $G_n^\alpha$  has the following properties.

- (i) For every  $\alpha \in (0, 1)$  and every  $n \in \mathbb{N}$  the Type I error of the test is not greater than  $\alpha$ :  $\rho(G_n^\alpha = 1) \leq \alpha$ .
- (ii) For every  $\alpha \in (0, 1)$  the Type II error goes to 0 almost surely: for every  $\rho' \neq \rho$  we have  $\lim_{n \rightarrow \infty} G_n^\alpha = 1$  with  $\rho'$  probability 1.

Note that using an appropriate randomization in the definition of  $C_\alpha^n$  we can make the Type I error exactly  $\alpha$ .

*Proof.* The first statement holds by construction. To prove the second statement, let the sample  $X$  be generated by  $\rho' \in \mathcal{S}$ ,  $\rho' \neq \rho$ , and define  $\delta = d(\rho, \rho')/2$ . By Lemma 1 we have  $\rho(D_\delta^n) \rightarrow 0$ , so that  $\rho(D_\delta^n) < \alpha$  from some  $n$  on; denote it  $n_1$ . Thus, for  $n > n_1$  we have  $D_\delta^n \subset C_\alpha^n$ . At the same time, by Lemma 1 we have  $\hat{d}(X, \rho) > \delta$  from some  $n$  on, which we denote  $n_2(X)$ , with  $\rho'$ -probability 1. So, for  $n > \max\{n_1, n_2(X)\}$  we have  $X \in D_\delta^n \subset C_\alpha^n$ , which proves the statement (ii).  $\square$

### 3.2 Process classification

Let there be given three samples  $X = (X_1, \dots, X_k)$ ,  $Y = (Y_1, \dots, Y_m)$  and  $Z = (Z_1, \dots, Z_n)$ . Each sample is generated by a stationary ergodic process  $\rho_X$ ,  $\rho_Y$  and  $\rho_Z$  respectively. Moreover, it is known that either  $\rho_Z = \rho_X$  or  $\rho_Z = \rho_Y$ , but  $\rho_X \neq \rho_Y$ . We wish to construct a test that, based on the finite samples  $X, Y$  and  $Z$  will tell whether  $\rho_Z = \rho_X$  or  $\rho_Z = \rho_Y$ .

The test chooses the sample  $X$  or  $Y$  according to whichever is closer to  $Z$  in  $\hat{d}$ . That is, we define the test  $G(X, Y, Z)$  as follows. If  $\hat{d}(X, Z) \leq \hat{d}(Y, Z)$  then the test says that the sample  $Z$  is generated by the same process as the sample  $X$ , otherwise it says that the sample  $Z$  is generated by the same process as the sample  $Y$ .

**Definition 4** (Process classifier). Define the classifier  $L : A^* \times A^* \times A^* \rightarrow \{1, 2\}$  as follows

$$L(X, Y, Z) := \begin{cases} 1 & \text{if } \hat{d}(X, Z) \leq \hat{d}(Y, Z) \\ 2 & \text{otherwise,} \end{cases}$$

for  $X, Y, Z \in A^*$ .

**Theorem 2.** The test  $L(X, Y, Z)$  makes only a finite number of errors when  $|X|, |Y|$  and  $|Z|$  go to infinity, with probability 1: if  $\rho_X = \rho_Z$  then  $L(X, Y, Z) = 1$  from some  $|X|, |Y|, |Z|$  on with probability 1; otherwise  $L(X, Y, Z) = 2$  from some  $|X|, |Y|, |Z|$  on with probability 1.

*Proof.* From the fact that  $d$  is a metric and from Lemma 1 we conclude that  $\hat{d}(X, Z) \rightarrow 0$  (with probability 1) if and only if  $\rho_X = \rho_Z$ . So, if  $\rho_X = \rho_Z$  then by assumption  $\rho_Y \neq \rho_Z$  and  $\hat{d}(X, Z) \rightarrow 0$  a.s. while

$$\hat{d}(Y, Z) \rightarrow d(\rho_Y, \rho_Z) \neq 0.$$

Thus in this case  $\hat{d}(Y, Z) > \hat{d}(X, Z)$  from some  $|X|, |Y|, |Z|$  on with probability 1, from which moment we have  $L(X, Y, Z) = 1$ . The opposite case is analogous.  $\square$

### 3.3 Change point problem

The sample  $Z = (Z_1, \dots, Z_n)$  consists of two concatenated parts  $X = (X_1, \dots, X_k)$  and  $Y = (Y_1, \dots, Y_m)$ , where  $m = n - k$ , so that  $Z_i = X_i$  for  $1 \leq i \leq k$  and  $Z_{k+j} = Y_j$  for  $1 \leq j \leq m$ . The samples  $X$  and  $Y$  are generated independently by two different stationary ergodic processes with alphabet  $A = \mathbb{R}$ . The distributions of the processes are unknown. The value  $k$  is called the *change point*. It is assumed that  $k$  is linear in  $n$ ; more precisely,  $\alpha n < k < \beta n$  for some  $0 < \alpha \leq \beta < 1$  from some  $n$  on.

It is required to estimate the change point  $k$  based on the sample  $Z$ .

For each  $t$ ,  $1 \leq t \leq n$ , denote  $U^t$  the sample  $(Z_1, \dots, Z_t)$  consisting of the first  $t$  elements of the sample  $Z$ , and denote  $V^t$  the remainder  $(Z_{t+1}, \dots, Z_n)$ .

**Definition 5** (Change point estimator). *Define the change point estimate  $\hat{k} : A^* \rightarrow \mathbb{N}$  as follows:*

$$\hat{k}(X_1, \dots, X_n) := \operatorname{argmax}_{t \in [\sqrt{n}, n - \sqrt{n}]} \hat{d}(U^t, V^t).$$

It should be noted that the term  $\sqrt{n}$  in this definition can be replaced by any  $o(n)$  function that goes to infinity with  $n$ ; this, in particular, does not affect the theorem below. Alternative approaches used in the literature on the change point problem are to introduce weights near the ends of the sample, or to assume known linear bounds on the change point (see e.g. [5]).

**Theorem 3.** *For the estimate  $\hat{k}$  of the change point  $k$  we have*

$$|\hat{k} - k| = o(n) \text{ a.s.}$$

where  $n$  is the size of the sample, and when  $k, n - k \rightarrow \infty$  in such a way that  $\alpha < \frac{k}{n} < \beta$  for some  $\alpha, \beta \in (0, 1)$  from some  $n$  on.

*Proof.* To prove the statement, we will show that for every  $\gamma$ ,  $0 < \gamma < 1$  with probability 1 the inequality  $\hat{d}(U^t, V^t) < \hat{d}(X, Y)$  holds for each  $t$  such that  $\sqrt{n} \leq t < \gamma k$  possibly except for a finite number of times. Thus we will show that linear  $\gamma$ -underestimates occur only a finite number of times, and for overestimate it is analogous. Fix some  $\gamma$ ,  $0 < \gamma < 1$  and  $\varepsilon > 0$ . Let  $J$  be big enough to have  $\sum_{i=J}^{\infty} w_i < \varepsilon/2$  and also big enough to have an index  $j < J$  for which  $\rho_X(B_j) \neq \rho_Y(B_j)$ . Take  $M_\varepsilon \in \mathbb{N}$  large enough to have  $|\nu(Y, B_i) - \rho_Y(B_i)| \leq \varepsilon/2J$  for all  $m > M_\varepsilon$  and for each  $i$ ,  $1 \leq i \leq J$ , and also to have  $|B_j|/m < \varepsilon/J$ . This

is possible since empirical frequencies converge to the limiting probabilities a.s. (that is,  $M_\varepsilon$  depends on the realizations  $Y_1, Y_2, \dots$ ) (cf. the proof of Lemma 1). Observe that the distribution of the sample  $X_s, X_{s+1}, \dots, X_k$ , where  $s$  is chosen independently of the sample, is governed by the same stationary ergodic process as  $X_1, \dots, X_k$ . Therefore, we can find such a  $K_\varepsilon$  (that depends on  $X$ ) that for all  $k > K_\varepsilon$  and for all  $i, 1 \leq i \leq J$  we will have  $|\nu(U^t, B_i) - \rho_X(B_i)| \leq \varepsilon/2J$  for each  $t \geq \sqrt{n}$ , and  $|\nu((X_s, X_{s+1}, \dots, X_k), B_i) - \rho_X(B_i)| \leq \varepsilon/2J$  for each  $s \leq \gamma k$ . So, for each  $s \in [\sqrt{n}, \gamma k]$  we have

$$\begin{aligned} & \left| \nu(V^s, B_j) - \frac{(1-\gamma)k\rho_X(B_j) + m\rho_Y(B_j)}{(1-\gamma)k + m} \right| \\ & \leq \left| \frac{(1-\gamma)k\nu((X_s, \dots, X_k), B_j) + m\nu(Y, B_j)}{(1-\gamma)k + m} - \frac{(1-\gamma)k\rho_X(B_j) + m\rho_Y(B_j)}{(1-\gamma)k + m} \right| + \frac{|B_j|}{m + \gamma k} \leq 3\varepsilon/J, \end{aligned}$$

for  $k > K_\varepsilon$  and  $m > M_\varepsilon$  (from the definitions of  $K_\varepsilon$  and  $M_\varepsilon$ ). Hence

$$\begin{aligned} & |\nu(X, B_j) - \nu(Y, B_j)| - |\nu(U^s, B_j) - \nu(V^s, B_j)| \\ & \geq |\nu(X, B_j) - \nu(Y, B_j)| \\ & \quad - \left| \nu(U^s, B_j) - \frac{(1-\gamma)k\rho_X(B_j) + m\rho_Y(B_j)}{(1-\gamma)k + m} \right| - 3\varepsilon/J \\ & \geq |\rho_X(B_j) - \rho_Y(B_j)| \\ & \quad - \left| \rho_X(B_j) - \frac{(1-\gamma)k\rho_X(B_j) + m\rho_Y(B_j)}{(1-\gamma)k + m} \right| - 4\varepsilon/J \\ & \qquad \qquad \qquad = \delta_j - 4\varepsilon/J, \end{aligned}$$

for some  $\delta_j$  that depends only on  $k/m$  and  $\gamma$ . Summing over all  $B_i, i \in \mathbb{N}$ , we get

$$\hat{d}(X, Y) - \hat{d}(U^s, V^s) \geq w_j \delta_j - 5\varepsilon,$$

for all  $n$  such that  $k > K_\varepsilon$  and  $m > M_\varepsilon$ , which is positive for small enough  $\varepsilon$ .  $\square$

## References

- [1] R. Ahlswede, I. Csiszar, Hypothesis testing with communication constraints. *IEEE Transactions on Information Theory*, vol 32 (4), pp. 533–542, 1986.
- [2] M. Basseville, I. Nikiforov. *Detection of Abrupt Changes: Theory and Applications*. Prentice Hall, 1993.
- [3] E. Carlstein, S. Lele, Nonparametric change-point estimation for data from an ergodic sequence. *Teor. Veroyatnost. i Primenen.* 38 (1993), no. 4, pp.

- 910–917; translation in *Theory Probab. Appl.* vol. 38 (4), pp. 726–733, 1993.
- [4] P. Billingsley, *Ergodic theory and information*. Wiley, New York, 1965.
- [5] B. Brodsky, B. Darkhovsky. *Nonparametric Methods in Change-Point Problems*. Kluwer Academic Publishers, 1993.
- [6] I. Csiszár, P. Shields, *Notes on Information Theory and Statistics: A tutorial*, *Foundations and Trends in Communications and Information Theory* 1 (2004), p. 1–111.
- [7] I. Csiszar, *Information Theoretic Methods in Probability and Statistics*. *Information Theory Soc. Rev. articles*, <http://www.itsoc.org/review/frrev.html>, 1997.
- [8] L. Giraitisa, R. Leipusb, D. Surgailis. The change-point problem for dependent observations. *Journal of Statistical Planning and Inference* Vol.53 (3), pp. 297-310, 1996.
- [9] R. Gray. *Probability, Random Processes, and Ergodic Properties*. Springer Verlag, 1988.
- [10] M. Gutman. Asymptotically Optimal Classification for Multiple Tests with Empirically Observed Statistics. *IEEE Trans. Information Theory*, vol. 35 no. 2, pp. 402–408, 1989.
- [11] J. C. Kieffer, Strongly consistent code-based identification and order estimation for constrained finite-state model classes. *IEEE Trans. Inform. Theory* vol.39 (3), pp. 893–902, 1993.
- [12] L. Gyorfi, G. Morvai, I. Vajda, Information-theoretic methods in testing the goodness of fit. In *proceedings of IEEE International Symposium on Information Theory*, 2000.
- [13] Ornstein, D. S. and Weiss, B.(1990). *How Sampling Reveals a Process*. *Annals of Probability* 18(3), pp. 905–930.
- [14] B. Ryabko, J. Astola. Universal codes as a basis for nonparametric testing of serial independence for time series. *Journal of Statistical Planning and Inference*, Vol. 136 (12), pp. 4119-4128, 2006.
- [15] B. Ryabko, J. Astola, A. Gammerman. Application of Kolmogorov complexity and universal codes to identity testing and nonparametric testing of serial independence for time series. *Theoretical Computer Science*, 359, 2006, pp. 440–448.
- [16] D. Ryabko, B. Ryabko. On hypotheses testing for ergodic processes. In *Proceedings of Information Theory Workshop (2008)*, Porto, Portugal, pp. 281–283.

- [17] P. Shields, The Interactions Between Ergodic Theory and Information Theory. IEEE Trans. on Information Theory, vol. 44, no. 6 (1998), pp. 2079–2093.