

Testing composite hypotheses about discrete ergodic processes

Daniil Ryabko

Received: 22 December 2010 / Accepted: 16 April 2011
© Sociedad de Estadística e Investigación Operativa 2011

Abstract Given a discrete-valued sample X_1, \dots, X_n , we wish to decide whether it was generated by a distribution belonging to a family H_0 , or it was generated by a distribution belonging to a family H_1 . In this work we assume that all distributions are stationary ergodic and do not make any further assumptions (in particular, no independence or mixing rate assumptions). We find some necessary and some sufficient conditions, formulated in terms of the topological properties of H_0 and H_1 , for the existence of a consistent test. For the case where H_1 is the complement of H_0 (to the set of all stationary ergodic processes), these necessary and sufficient conditions coincide, thereby providing a complete characterization of families of processes membership to which can be consistently tested, against their complement, based on sampling. This criterion includes as special cases several known and some new results on testing for membership to various parametric families, as well as testing identity, independence, and other hypotheses.

Keywords Hypothesis testing · Property testing · Stationary ergodic time series · Distributional distance

Mathematics Subject Classification (2000) 62G10 · 62G20 · 62M07

1 Introduction

Given a sample X_1, \dots, X_n (where X_i are from a finite alphabet A) which is known to be generated by a stationary ergodic process, we wish to decide whether it was

Communicated by Domingo Morales.

D. Ryabko (✉)
INRIA Lille-Nord Europe, 40 avenue Halley Parc Scientifique de la Haute Borne, 59650 Villeneuve
d'Ascq, France
e-mail: daniil@ryabko.net

generated by a distribution belonging to a family H_0 versus it was generated by a distribution belonging to a family H_1 . Unlike most of the works on the subject, we do not assume that X_i are i.i.d. but only make a much weaker assumption that the distribution generating the sample is stationary ergodic.

Examples Let us give some examples motivating the general problem in question. The most basic case of the hypothesis testing problem is testing a simple hypothesis $H_0 = \{\rho_0\}$ versus a simple hypothesis $H_1 = \{\rho_1\}$, where ρ_0 and ρ_1 are two stationary ergodic process distributions (which are assumed completely known theoretically). A more complex but more realistic problem is when only one of the hypotheses is simple, $H_0 = \{\rho_0\}$, but the alternative is general, for example, H_1 is the set of all stationary ergodic processes that are different from ρ_0 . One may also consider variants in which the alternative is the set of all stationary ergodic processes that differ from ρ_0 by at least ε in some distance. The described hypotheses are variants of the so-called goodness-of-fit, or identity testing problem. Another class of hypothesis testing problems is presented by the problem of *model verification*. Suppose that we have some relatively simple (possibly parametric) set of assumptions, and we wish to test whether the process generating the given sample satisfies these assumptions. As an example, H_0 can be the set of all k th-order Markov processes (fixed $k \in \mathbb{N}$), and H_1 is the set of all stationary ergodic processes that do not belong to H_0 ; one may also wish to consider more restrictive alternatives, for example, H_1 is the set of all k' th-order Markov processes where $k' > k$. Of course, instead of Markov processes, one can consider other models, e.g., hidden Markov processes. A similar problem is that of testing that the process has entropy less than some given ε versus that its entropy exceeds ε , or versus that its entropy is greater than $\varepsilon + \delta$ for some positive δ .

Yet another type of hypothesis testing problems concerns *property testing*. Suppose that we are given two samples, generated independently of each other by stationary ergodic distributions, and we wish to test the hypothesis that they are independent versus that they are not independent, or that they are generated by the same process versus that they are generated by different processes.

In all the considered cases, where the hypothesis testing problem turns out to be too difficult (i.e., there is no consistent test for the chosen notion of consistency) for the case of stationary ergodic processes, one may wish to restrict either H_0 , H_1 or both H_0 and H_1 to some smaller class of processes. Thus, one may wish to test the hypothesis of independence when, for example, both processes are known to have finite memory or to have certain mixing rates.

All the problems described above are special cases of the following general formulation: given two sets H_0 and H_1 which are contained in the set of all stationary ergodic process distributions and given a sample generated by a process that comes from either H_0 or H_1 , we would like to have a test that tells us which one is the case, H_0 or H_1 . The purpose of this paper is to characterize those pairs of H_0 , H_1 for which a consistent test exists. Ideally, the characterization should be complete, that is, in the form of necessary and sufficient conditions that can be verified for at least most of the problems outlined above. This goal is partially achieved: we find some necessary and some sufficient conditions that coincide in the case where H_1 is the complement of H_0 . We show that these conditions are indeed relatively easy to verify for some

of the considered hypotheses, such as identity testing, model verification, and testing independence.

Consistency A test is a function that takes, as input, a sample and a real-valued parameter (confidence level) and outputs a binary (possibly incorrect) answer: the sample was generated by a distribution from H_0 or from H_1 . An answer $i \in \{0, 1\}$ is correct if the sample is generated by a distribution that belongs to H_i . Here we are concerned with characterizing those pairs of H_0 and H_1 for which consistent tests exist.

Although there are many different (nonequivalent) ways to define consistency of a test, in this work we concentrate on one of them, which is perhaps the one most commonly used in mathematical statistics: the probability of Type I error is fixed (for any size of the sample), and the probability of Type II error is required to go to zero as the sample size increases. Type I error occurs if the test says “1” while the sample was generated by the distribution from H_0 . Type II error occurs if the test says “0” while H_1 is true. In many practical situations, these errors may have very different meanings: for example, this is the case where H_0 is interpreted as that a patient has a certain ailment, and H_1 that he does not. In such cases, one may wish to treat the errors asymmetrically. Also, H_0 can often be much simpler than the alternative H_1 ; for example, H_0 can be a simple parametric family, or it may consist of just one process distribution, while H_1 can be the complement of H_0 to the set of all stationary ergodic processes.

A test is called *consistent* if, for any prespecified level $\alpha \in (0, 1)$, any sample size n , and any distribution in H_0 , the probability of Type I error (the test says H_1) is not greater than α , while for every distribution in H_1 and every α , the Type II error is made only a finite number of times with probability 1 as the sample size goes to infinity.

Prior work There is a vast body of literature on hypothesis testing for i.i.d. (real- or discrete-valued) data (see, e.g., Lehmann 1986; Kendall and Stuart 1961). In the context of discrete-valued i.i.d. data, the necessary and sufficient conditions for the existence of a consistent test are rather simple. There is a consistent test if and only if the closure of H_0 does not intersect H_1 , where the topology is that of the parameter space (probabilities of each symbol), see, e.g., Csiszar (1967). Some extensions to Markov chains are also possible (Billingsley 1961; Anderson and Goodman 1957).

There is, however, much less literature on hypothesis testing beyond i.i.d. or parametric models, while the question of determining whether a consistent test exists, for various particular hypotheses as well as in general, is much less trivial. For a weaker notion of consistency, namely, requiring that the test should stabilize on the correct answer for a.e. realization of the process (under either H_0 or H_1), (Kieffer 1993) constructs a consistent test for so-called constrained finite-state model classes (including finite-state Markov and hidden Markov processes) against the general alternative of stationary ergodic processes. In Morvai and Weiss (2005) some results are presented on testing the hypothesis that the process has a finite memory and on some related problems. Consistent tests for some specific hypotheses, but under the general alternative of stationary ergodic processes, have been proposed in Ryabko

and Astola (2006), Ryabko et al. (2006), Ryabko and Ryabko (2010), which address problems of testing identity, independence, estimating the order of a Markov process, and also the change point problem. Noteworthy, a conceptually simple hypothesis of homogeneity (testing whether two sample are generated by the same or by different processes) does not admit a consistent test even in the weakest asymptotic sense, as was shown in Ryabko (2010) (whereas for i.i.d. data, it can of course be solved; see, e.g., Biau and Györfi 2005 and references therein).

The results Here we obtain some topological characterizations of the hypotheses for which consistent tests exist, for the case of stationary ergodic distributions. This characterization is rather similar to the one mentioned above for the case of i.i.d. data but is formulated with respect to the topology of distributional distance. The fact that necessary and sufficient conditions are obtained for the case of complementary hypotheses indicates that this topology is the right one to consider.

A distributional distance between two process distributions is defined as a weighted sum of differences of probabilities of all possible tuples $X \in A^*$, where A is the alphabet, and the weights are positive and have a finite sum.

The tests we construct are based on empirical estimates of distributional distance. For a given level α , we take the smallest ε -neighborhood of the closure of H_0 that has probability not less than $1 - \alpha$ with respect to any distribution in it and outputs 0 if the sample falls into this neighborhood and 1 otherwise.

For the case of testing H_0 against its complement to the set \mathcal{E} of all stationary ergodic processes, we obtain the following necessary and sufficient condition (formalized in the next section).

Theorem There exists a consistent test for H_0 against $H_1 := \mathcal{E} \setminus H_0$ if and only if H_1 has probability 0 with respect to ergodic decomposition of every distribution from the closure of H_0 .

For the general case, we obtain some necessary and some sufficient conditions in the same terms. The main results are illustrated with derivations of several known and some new results for specific hypotheses. In particular, the established results generalize the known ones on testing membership to such parametric families as k -th-order Markov processes and k -state hidden Markov processes.

Although we provide an explicit definition of tests for any H_0, H_1 , the main value of the results is theoretical. They are intended to serve as a tool for checking whether a consistent test can be constructed in principle, rather than for providing actual testing procedures. At the same time, while we leave for further research the problem of converting the tests that we construct into efficient testing procedures for specific problems, we note that this may often indeed be possible, since empirical estimates of the distributional distance on which the test is based can be computed efficiently, as is demonstrated in Ryabko (2010) on the problem of clustering.

2 Preliminaries

Let A be a finite alphabet, and denote by A^* the set of words (or tuples) $\bigcup_{i=1}^{\infty} A^i$. For a word B , the symbol $|B|$ stands for the length of B . Denote by B_i the i th element of

A^* , enumerated in such a way that the elements of A^i appear before the elements of A^{i+1} for all $i \in \mathbb{N}$. *Distributions* or (*stochastic*) *processes* are probability measures on the space $(A^\infty, \mathcal{F}_{A^\infty})$, where \mathcal{F}_{A^∞} is the Borel sigma-algebra of A^∞ . Denote by $\#(X, B)$ the number of occurrences of a word B in a word $X \in A^*$, and by $\nu(X, B)$ its frequency:

$$\#(X, B) = \sum_{i=1}^{|X|-|B|+1} I_{\{(X_i, \dots, X_{i+|B|-1})=B\}},$$

and

$$\nu(X, B) = \begin{cases} \frac{1}{|X|-|B|+1} \#(X, B) & \text{if } |X| \geq |B|, \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where $X = (X_1, \dots, X_{|X|})$. For example, $\nu(0001, 00) = 2/3$.

We use the abbreviation $X_{1..k}$ for X_1, \dots, X_k . A process ρ is *stationary* if

$$\rho(X_{1..|B|} = B) = \rho(X_{t..t+|B|-1} = B)$$

for any $B \in A^*$ and $t \in \mathbb{N}$. Denote by \mathcal{S} the set of all stationary processes on $(A^\infty, \mathcal{F}_{A^\infty})$. A stationary process ρ is called (*stationary*) *ergodic* if the frequency of occurrence of each word B in a sequence X_1, X_2, \dots generated by ρ tends to its a priori (or limiting) probability a.s.: $\rho(\lim_{n \rightarrow \infty} \nu(X_{1..n}, B) = \rho(X_{1..|B|} = B)) = 1$. By virtue of the ergodic theorem (see, for example, Billingsley 1965), this definition can be shown to be equivalent to the standard definition of stationary ergodic processes (every shift-invariant set has measure 0 or 1; see, e.g., Csiszar and Shields 2004). Denote by \mathcal{E} the set of all stationary ergodic processes.

A *distributional distance* is defined for a pair of processes ρ_1, ρ_2 as follows (Gray 1988):

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

where w_i are summable positive real weights (e.g., $w_k = 2^{-k}$: we fix this choice for the sake of concreteness). It is easy to see that d is a metric. Equipped with this metric, the space of all stochastic processes is separable and complete; moreover, it is a compact. The set of stationary processes \mathcal{S} is its convex closed subset (hence a compact too). The set of all finite-memory stationary distributions is dense in \mathcal{S} . (Taking only those that have rational transition probabilities, we obtain a countable dense subset of \mathcal{S} .) The set \mathcal{E} is not convex (a mixture of stationary ergodic distributions is always stationary but never ergodic) and is not closed (its closure is \mathcal{S}). We refer to Gray (1988) for more details and proofs of these facts.

When talking about closed and open subsets of \mathcal{S} , we assume the topology of d . The compactness of the set \mathcal{S} is one of the main ingredients in the proofs of the main results. Another is that the distance d can be consistently estimated, as the next lemma shows.

Define the *empirical estimates of the distributional distance* d :

$$\hat{d}(X_{1..n}, \rho) = \sum_{i=1}^{\infty} w_i |v(X_{1..n}, B_i) - \rho(B_i)|,$$

where $n \in \mathbb{N}$, $\rho \in \mathcal{S}$, and $X_{1..n} \in A^n$. That is, $\hat{d}(X_{1..n}, \rho)$ measures the discrepancy between empirically estimated and theoretical probabilities.

Lemma 1 (*\hat{d} is consistent* (Ryabko and Ryabko 2010)) *Let $\rho, \xi \in \mathcal{E}$, and let a sample $X_{1..k}$ be generated by ρ . Then $\lim_{k \rightarrow \infty} \hat{d}(X_{1..k}, \xi) = d(\rho, \xi)$ ρ -a.s.*

Considering the Borel (with respect to the metric d) sigma-algebra $\mathcal{F}_{\mathcal{S}}$ on the set \mathcal{S} , we obtain a standard measurable space $(\mathcal{S}, \mathcal{F}_{\mathcal{S}})$. When talking about measurable subsets of \mathcal{S} , we refer to this space. In particular, \mathcal{E} is measurable. An important tool that will be used in the analysis is *ergodic decomposition* of stationary processes (see, e.g., Gray 1988; Billingsley 1965): any stationary process can be expressed as a mixture of stationary ergodic processes. More formally, for any $\rho \in \mathcal{S}$, there is a measure W_{ρ} on $(\mathcal{S}, \mathcal{F}_{\mathcal{S}})$ such that $W_{\rho}(\mathcal{E}) = 1$ and $\rho(B) = \int dW_{\rho}(\mu)\mu(B)$ for any $B \in \mathcal{F}_{A^{\infty}}$.

A *test* is a function $\phi : A^* \rightarrow \{0, 1\}$ that takes a sample and outputs a binary answer, where the answer i is interpreted as “the sample was generated by a distribution that belongs to H_i .” The answer i is correct if the sample was indeed generated by a distribution from H_i ; otherwise we say that the test made an *error*. A test ϕ makes the *Type I* error if it says 1 while H_0 is true, and it makes *Type II* error if it says 0 while H_1 is true.

Definition 1 (*consistency*) A family of tests ψ^{α} , $\alpha \in (0, 1)$, is called *consistent* if:

- (i) The probability of Type I error is always bounded by α : $\rho(\psi^{\alpha}(X_{1..n}) = 1) \leq \alpha$ for all $n \in \mathbb{N}$, all $\rho \in H_0$, and all $\alpha \in (0, 1)$, and
- (ii) Type II error is made not more than a finite number of times with probability 1: $\rho(\lim_{n \rightarrow \infty} \psi^{\alpha}(X_{1..n}) = 1) = 1$ for every $\rho \in H_1$ and every $\alpha \in (0, 1)$.

Note that the treatment of the error probabilities is asymmetric: the probability of Type I error is upper-bounded by α for all n (nonsymptotic), while the Type II error is guaranteed to be made only a finite number of times (only an asymptotic guarantee). Abusing the notation, we will sometimes call families of tests ψ^{α} , $\alpha \in (0, 1)$, simply *tests*.

3 Main results

To define the tests whose consistency will be established in this section, we first need to define an empirical distance between a sample and a family of processes.

For a sample $X_{1..n} \in A^n$ and a hypothesis $H \subset \mathcal{E}$, define

$$\hat{d}(X_{1..n}, H) = \inf_{\rho \in H} \hat{d}(X_{1..n}, \rho).$$

For $H \subset \mathcal{S}$, denote by $\text{cl } H$ the closure of H (with respect to the topology of d).

Construct the test ψ_{H_0, H_1}^α , $\alpha \in (0, 1)$, as follows. For all $n \in \mathbb{N}$, $\delta > 0$, and $H \subset \mathcal{E}$, define the neighborhood $b_\delta^n(H)$ of n -tuples around H as

$$b_\delta^n(H) := \{X \in A^n : \hat{d}(X, H) \leq \delta\}.$$

Moreover, let

$$\gamma_n(H, \theta) := \inf \left\{ \delta : \inf_{\rho \in H} \rho(b_\delta^n(H)) \geq \theta \right\}$$

be the smallest radius of a neighborhood around H that has the probability not less than θ with respect to any process in H (clearly, it exists and is positive), and let $C^n(H, \theta) := b_{\gamma_n(H, \theta)}^n(H)$ be the neighborhood of this radius. Define

$$\psi_{H_0, H_1}^\alpha(X_{1..n}) := \begin{cases} 0 & \text{if } X_{1..n} \in C^n(\text{cl } H_0 \cap \mathcal{E}, 1 - \alpha), \\ 1 & \text{otherwise.} \end{cases}$$

Since the set \mathcal{S} is a complete separable metric space, it is easy to see that the function $\psi_{H_0, H_1}^\alpha(X_{1..n})$ is measurable, provided that $\text{cl } H_0$ is measurable. We will often omit the subscript H_0, H_1 from ψ_{H_0, H_1}^α when it can cause no confusion.

Theorem 1 *Let $H_0, H_1 \subset \mathcal{E}$ with H_0 measurable. If $W_\rho(H_0) = 1$ for every $\rho \in \text{cl } H_0$, then the test ψ_{H_0, H_1}^α is consistent. Conversely, if there is a consistent test for H_0 against H_1 then $W_\rho(H_1) = 0$ for any $\rho \in \text{cl } H_0$.*

For the case where H_1 is the complement of H_0 , the necessary and sufficient conditions of Theorem 1 coincide and give the following criterion.

Corollary 1 *Let H_0 be a measurable subset of E , and let $H_1 = \mathcal{E} \setminus H_0$. The following statements are equivalent:*

- (i) *There exists a consistent test for H_0 against H_1 .*
- (ii) *The test ψ_{H_0, H_1}^α is consistent.*
- (iii) *The set H_1 has probability 0 with respect to ergodic decomposition of every ρ in the closure of H_0 : $W_\rho(H_1) = 0$ for each $\rho \in \text{cl } H_0$.*

4 Examples

Theorem 1 can be used to check whether a consistent test exists for such problems as identity, independence, estimating the order of a (hidden) Markov model, bounding the entropy, bounding the distance, uniformity, monotonicity, etc. Some of these examples are considered in this section.

Example 1: Simple hypotheses, identity. First of all, it is obvious that sets that consist of just one or finitely many stationary ergodic processes are closed and closed under ergodic decompositions; therefore, for any pair of disjoint sets of this type, there exists a consistent test. (In particular, there is a consistent test for $H_0 = \{\rho_0\}$)

against $H_1 = \{\rho_1\}$ for any $\rho_0, \rho_1 \in \mathcal{E}$.) A more interesting case is identity testing, or goodness-of-fit; the problem here consists in testing whether a distribution generating the sample obeys a certain given law versus it does not. Let $\rho \in \mathcal{E}$, $H_0 = \{\rho\}$, and $H_1 = \mathcal{E} \setminus H_0$. For this problem, Theorem 1 implies that there is a consistent test for H_0 against H_1 . (Indeed, the conditions of the theorem are easily verified.) Identity testing is a classical problem of mathematical statistics, with solutions (e.g., based on Pearson's χ^2 statistic) for i.i.d. data (e.g., Lehmann 1986) and Markov chains (Billingsley 1961). For stationary ergodic processes, Ryabko et al. (2006) gives a consistent test when H_0 has a finite and bounded memory, and Ryabko and Ryabko (2010) for the general case of stationary ergodic real-valued processes.

Example 2: Markov and hidden Markov processes: bounding the order. For any k , there exists a consistent test of the hypothesis $\mathcal{M}^k =$ “the process is Markov of order not greater than k ” against $\mathcal{E} \setminus \mathcal{M}^k$. For any k , there exists a consistent test of $\mathcal{HM}^k =$ “the process is given by a hidden Markov process with not more than k states” against $H_1 = \mathcal{E} \setminus \mathcal{HM}^k$. Indeed, in both cases (k th-order Markov, hidden Markov with no more than k states), the hypothesis H_0 is a continuously parameterized family with a compact set of parameters; that is, H_0 is a closed subset of \mathcal{S} . Moreover, in both cases it is easy to see that H_0 is closed under taking ergodic decompositions. Thus, by Theorem 1, there exists a consistent test.

The problem of estimating the order of a (hidden) Markov process, based on a sample from it, was addressed in a number of works. In the contest of hypothesis testing, consistent tests for \mathcal{M}^k against \mathcal{M}^t with $t > k$ were given in Anderson and Goodman (1957), see also Billingsley (1961). The existence of asymptotically consistent tests for \mathcal{M}^k against $\mathcal{E} \setminus \mathcal{M}^k$, and of \mathcal{HM}^k against $\mathcal{E} \setminus \mathcal{HM}^k$, was established in Kieffer (1993), see also Csiszar and Shields (2000). Consistent tests for \mathcal{M}^k against $\mathcal{E} \setminus \mathcal{M}^k$ were obtained in Ryabko and Astola (2006), while for the case of testing for \mathcal{HM}^k against $\mathcal{E} \setminus \mathcal{HM}^k$ the positive result above is apparently new.

Example 3: Smooth parametric families. From the discussion in the previous example we can see that the following generalization is valid. Let $H_0 \subset \mathcal{S}$ be a set of processes that is continuously parameterized by a compact set of parameters. If H_0 is closed under taking ergodic decompositions, then there exists a consistent test for H_0 against $\mathcal{E} \setminus H_0$. In particular, this strengthens the mentioned result of Kieffer (1993), since a stronger notion of consistency is used, as well as a more general class of parametric families is considered.

Example 4: Independence. Suppose that $A = A_1 \times A_2$, so that a sample $X_{1..n}$ consists of two processes $X_{1..n}^1$ and $X_{1..n}^2$, which we call features. The hypothesis of independence is that the first feature is independent from the second: $\rho(X_{1..t}^1 \in T_1, X_{1..t}^2 \in T_2) = \rho(X_{1..t}^1 \in T_1)\rho(X_{1..t}^2 \in T_2)$ for any $(T_1, T_2) \in A^n$ and any $n \in \mathbb{N}$. Let \mathcal{I} be the set of all stationary ergodic processes satisfying this property. It is easy to see that Theorem 1 implies that there exists a consistent test for $\mathcal{I} \cap \mathcal{M}^k$ against $\mathcal{E} \setminus \mathcal{I}$ for any given $k \in \mathbb{N}$. Analogously, if we confine H_0 to hidden Markov processes of a given order, then consistent testing is possible. That is, there exists a consistent test for $\mathcal{I} \cap \mathcal{HM}^k$ against $\mathcal{E} \setminus \mathcal{I}$ for any given $k \in \mathbb{N}$. The question of whether \mathcal{I} can be tested

against $\mathcal{E} \setminus \mathcal{I}$ is more difficult. It is clear that the closure of \mathcal{I} only contains processes with independent features. It is not clear whether any of the limiting points of \mathcal{I} has ergodic components whose features are not independent. If there are none, this would prove that there exists a consistent test for independence, for the class of stationary ergodic process.

The existence of a consistent test for $\mathcal{I} \cap \mathcal{M}^k$ against $\mathcal{E} \setminus \mathcal{I}$ is due to Ryabko and Astola (2006); the extension to hidden Markov processes is apparently new. On the case of i.i.d. variables, see, for example, Gretton and Györfi (2010) and references therein.

5 Proofs

The proofs will use the following lemmas.

Lemma 2 (Smooth probabilities of deviation) *Let $m > 2k > 1$, $\rho \in \mathcal{S}$, $H \subset \mathcal{S}$, and $\varepsilon > 0$. Then*

$$\rho(\hat{d}(X_{1..m}, H) \geq \varepsilon) \leq \rho\left(\hat{d}(X_{1..k}, H) \geq \varepsilon - \frac{2k}{m - k + 1} - t_k\right), \tag{2}$$

where t_k is the sum of all the weights of tuples longer than k in the definition of d : $t_k := \sum_{i:|B_i|>n} w_i$, and

$$\rho(\hat{d}(X_{1..m}, H) \leq \varepsilon) \leq \rho\left(\hat{d}(X_{1..k}, H) \leq \frac{m}{m - k + 1}\varepsilon + \frac{2k}{m - k + 1}\right). \tag{3}$$

The meaning of this lemma is as follows. For any word $X_{1..m}$, if it is far away from (or close to) a given distribution μ (in the empirical distributional distance), then some of its shorter subwords $X_{i..i+k}$ is far from (close to) μ too. By stationarity, we may assume that $i = 1$. Therefore, the probability of a δ -ball of samples of a given length is close to the probability of a δ -ball of samples of a smaller size. In other words, for a stationary distribution μ , it cannot happen that a small sample is likely to be close to μ , but a larger sample is likely to be far.

Proof Let B be a tuple such that $|B| < k$, and $X_{1..m} \in A^m$ be any sample of size $m > 1$. The number of occurrences of B in X can be bounded by the number of occurrences of B in subwords of X of length k as follows:

$$\begin{aligned} \#(X_{1..m}, B) &\leq \frac{1}{k - |B| + 1} \sum_{i=1}^{m-k+1} \#(X_{i..i+k-1}, B) + 2k \\ &= \sum_{i=1}^{m-k+1} v(X_{i..i+k-1}, B) + 2k. \end{aligned}$$

Indeed, summing over $i = 1..m - k$ the number of occurrences of B in all $X_{i..i+k-1}$, we count each occurrence of B exactly $k - |B| + 1$ times, except for those that occur

in the first and last k symbols. Dividing by $m - |B| + 1$ and using definition (1), we obtain

$$v(X_{1..m}, B) \leq \frac{1}{m - |B| + 1} \left(\sum_{i=1}^{m-k+1} v(X_{i..i+k-1}, B) + 2k \right). \quad (4)$$

Summing over all B , for any μ , we get

$$\hat{d}(X_{1..m}, \mu) \leq \frac{1}{m - k + 1} \sum_{i=1}^{m-k+1} \hat{d}(X_{i..i+n-1}, \mu) + \frac{2k}{m - k + 1} + t_k, \quad (5)$$

where in the right-hand side, t_k corresponds to all the summands in the left-hand side for which $|B| > k$, where for the rest of the summands, we used $|B| \leq k$. Since this holds for any μ , we conclude that

$$\hat{d}(X_{1..m}, H) \leq \frac{1}{m - k + 1} \left(\sum_{i=1}^{m-k+1} \hat{d}(X_{i..i+k-1}, H) \right) + \frac{2k}{m - k + 1} + t_k.$$

Therefore, for any $X_{1..m} \in A^m$, if $\hat{d}(X_{1..m}, H) > \varepsilon$, then there is an index $i \leq m - k$ such that $\hat{d}(X_{i..i+k-1}, H) > \varepsilon - \frac{2k}{m-k+1} - t_k$. Moreover, we have (by the definition of stationarity)

$$\rho(\hat{d}(X_{i..i+k-1}, H) > \varepsilon') = \rho(\hat{d}(X_{1..k}, H) > \varepsilon'),$$

where $\varepsilon' = \varepsilon - \frac{2k}{m-k+1} - t_k$. So we have

$$\rho(\hat{d}(X_{1..k}, H) \geq \varepsilon') \geq \rho(\hat{d}(X_{1..m}, H) \geq \varepsilon),$$

proving (2). The second statement can be proven similarly; indeed, analogously to (4) we have

$$\begin{aligned} v(X_{1..m}, B) &\geq \frac{1}{m - |B| + 1} \sum_{i=1}^{m-k+1} v(X_{i..i+k-1}, B) - \frac{2k}{m - |B| + 1} \\ &\geq \frac{1}{m - k + 1} \left(\frac{m - k + 1}{m} \sum_{i=1}^{m-k+1} v(X_{i..i+k-1}, B) \right) - \frac{2k}{m}, \end{aligned}$$

where we have used $|B| \geq 1$. Summing over different B , we obtain (similarly to (5))

$$\hat{d}(X_{1..m}, \mu) \geq \frac{1}{m - k + 1} \sum_{i=1}^{m-k+1} \frac{m - k + 1}{m} \hat{d}_k(X_{i..i+n-1}, \mu) - \frac{2k}{m}$$

(since the frequencies are nonnegative, there is no term t_n here), which, using stationarity of ρ , implies (3). \square

Lemma 3 Let $\rho_k \in \mathcal{S}$, $k \in \mathbb{N}$, be a sequence of processes that converges to a process ρ_* . Then, for any $T \in A^*$ and $\varepsilon > 0$, if $\rho_k(T) > \varepsilon$ for infinitely many indices k , then $\rho_*(T) \geq \varepsilon$.

Proof The statement follows from the fact that $\rho(T)$ is continuous as a function of ρ . \square

Proof of Theorem 1 To establish the first statement of Theorem 1, we have to show that the family of tests ψ^α is consistent. By construction, for any $\rho \in \text{cl } H_0 \cap \mathcal{E}$, we have $\rho(\psi^\alpha(X_{1..n}) = 1) \leq \alpha$.

To prove the consistency of ψ , it remains to show that

$$\xi \left(\lim_{n \rightarrow \infty} \psi^\alpha(X_{1..n}) = 1 \right)$$

for any $\xi \in H_1$ and $\alpha > 0$. To do this, fix any $\xi \in H_1$ and let

$$\Delta := d(\xi, \text{cl } H_0) := \inf_{\rho \in \text{cl } H_0 \cap \mathcal{E}} d(\xi, \rho).$$

Since $\xi \notin \text{cl } H_0$, we have $\Delta > 0$. Suppose that there exists an $\alpha > 0$ such that, for infinitely many n , some samples from the $\Delta/2$ -neighborhood of n -samples around ξ are sorted as H_0 by ψ , that is, $C^n(\text{cl } H_0 \cap \mathcal{E}, 1 - \alpha) \cap b_{\Delta/2}^n(\xi) \neq \emptyset$. Then, for these n , we have $\gamma_n(\text{cl } H_0 \cap \mathcal{E}, 1 - \alpha) \geq \Delta/2$.

This means that there exists an increasing sequence $n_k, k \in \mathbb{N}$, and a sequence $\rho_k \in \text{cl } H_0, k \in \mathbb{N}$, such that

$$\rho_k(b_{\Delta/2}^{n_k}(\text{cl } H_0 \cap \mathcal{E})) < 1 - \alpha.$$

Since the set $\text{cl } H_0$ is compact (as a closed subset of a compact set \mathcal{S}), we may assume (passing to a subsequence, if necessary) that ρ_k converges to a certain $\rho_* \in \text{cl } H_0$. Using Lemma 2, (3), for every m large enough to satisfy $\frac{n_m}{n_m - n_k + 1} \delta/4 + \frac{n_k}{n_m - n_k + 1} < \delta/2$, we have

$$\rho_m(b_{\Delta/4}^{n_k}(\text{cl } H_0 \cap \mathcal{E})) < 1 - \alpha.$$

Since this holds for infinitely many m , using Lemma 3 (with $T = b_{\Delta/4}^{n_k}(\text{cl } H_0 \cap \mathcal{E})$), we conclude that

$$\rho_*(b_{\Delta/4}^{n_k}(\text{cl } H_0 \cap \mathcal{E})) \leq 1 - \alpha.$$

Since the latter inequality holds for infinitely many indices k , we also have

$$\rho_* \left(\limsup_{n \rightarrow \infty} \hat{d}(X_{1..n}, \text{cl } H_0 \cap \mathcal{E}) > \Delta/4 \right) > 0.$$

However, we must have $\rho_*(\lim_{n \rightarrow \infty} \hat{d}(X_{1..n}, \text{cl } H_0 \cap \mathcal{E}) = 0) = 1$ for every $\rho_* \in \text{cl } H_0$: indeed, for $\rho_* \in \text{cl } H_0 \cap \mathcal{E}$, this follows from Lemma 1 and, for $\rho_* \in \text{cl } H_0 \setminus \mathcal{E}$, from Lemma 1, ergodic decomposition, and the conditions of the theorem ($W_\rho(H_0) = 1$ for $\rho \in \text{cl } H_0$).

This contradiction shows that for every α , there are no more than finitely many n for which $C^n(\text{cl } H_0 \cap \mathcal{E}, 1 - \alpha) \cap b_{\Delta/2}^n(\xi) \neq \emptyset$. To finish the proof of the first statement, it remains to note that, as follows from Lemma 1,

$$\begin{aligned} & \xi \{X_1, X_2, \dots : X_{1..n} \in b_{\Delta/2}^n(\xi) \text{ from some } n \text{ on}\} \\ & \geq \xi \left(\lim_{n \rightarrow \infty} \hat{d}(X_{1..n}, \xi) = 0 \right) = 1. \end{aligned}$$

To establish the second statement of Theorem 1, we assume that there exists a consistent test ϕ for H_0 against H_1 , and we will show that $W_\rho(H_1) = 0$ for every $\rho \in \text{cl } H_0$. Take $\rho \in \text{cl } H_0$ and suppose that $W_\rho(H_1) = \delta > 0$. We have

$$\limsup_{n \rightarrow \infty} \int_{H_1} dW_\rho(\mu) \mu(\psi_n^{\delta/2} = 0) \leq \int_{H_1} dW_\rho(\mu) \limsup_{n \rightarrow \infty} \mu(\psi_n^{\delta/2} = 0) = 0,$$

where the inequality follows from Fatou's lemma (the functions under the integral are all bounded by 1), and the equality from the consistency of ψ . Thus, from some n on, we will have $\int_{H_1} dW_\rho(\mu) \mu(\psi_n^{\delta/2} = 0) < 1/4$, so that $\rho(\psi_n^{\delta/2} = 0) < 1 - 3\delta/4$. For any set $T \in A^n$, the function $\mu(T)$ is continuous as a function of T . In particular, this holds for the set $T := \{X_{1..n} : \psi_n^{\delta/2}(X_{1..n}) = 0\}$. Therefore, since $\rho \in \text{cl } H_0$, for any n large enough, we can find a $\rho' \in H_0$ such that $\rho'(\psi_n^{\delta/2} = 0) < 1 - 3\delta/4$, which contradicts the consistency of ψ . Thus, $W_\rho(H_1) = 0$, and Theorem 1 is proven. \square

Acknowledgements Some preliminary results of this work have appeared in (Ryabko 2010). This research was partially supported by the French Ministry of Higher Education and Research, Nord-Pas-de-Calais Regional Council and FEDER through CPER 2007-2013, ANR projects EXPLO-RA (ANR-08-COSI-004) and Lampada (ANR-09-EMER-007), by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 231495 (project ComplACS) and by Pascal-2.

References

- Anderson T, Goodman L (1957) Statistical inference about Markov chains. *Ann Math Stat* 28(1):89–110
- Biau G, Györfi L (2005) On the asymptotic properties of a nonparametric l_1 -test of homogeneity. *IEEE Trans Inf Theory* 51:3965–3973
- Billingsley P (1961) Statistical inference about Markov chains. *Ann Math Stat* 32(1):12–40
- Billingsley P (1965) *Ergodic theory and information*. Wiley, New York
- Csiszar I (1967) Information-type measures of difference of probability distributions and indirect observations. *Studia Sci Math Hung* 2:299–318
- Csiszar I, Shields P (2000) The consistency of the BIC Markov order estimator. *Ann Stat* 28(6):1601–1619
- Csiszar I, Shields P (2004) Notes on information theory and statistics. In: *Foundations and trends in communications and information theory*.
- Gray R (1988) *Probability, random processes, and ergodic properties*. Springer, Berlin
- Gretton A, Györfi L (2010) Consistent nonparametric tests of independence. *J Mach Learn Res* 11:1391–1423
- Kendall M, Stuart A (1961) *The advanced theory of statistics. Inference and relationship, vol 2*, London (1961)
- Kieffer J (1993) Strongly consistent code-based identification and order estimation for constrained finite-state model classes. *IEEE Trans Inf Theory* 39(3):893–902
- Lehmann E (1986) *Testing statistical hypotheses*, 2nd edn. Wiley, New York
- Morvai G, Weiss B (2005) On classifying processes. *Bernoulli* 11(3):523–532
- Ryabko B, Astola J (2006) Universal codes as a basis for time series testing. *Stat Methodol* 3:375–397
- Ryabko B, Astola J, Gammernan A (2006) Application of Kolmogorov complexity and universal codes to identity testing and nonparametric testing of serial independence for time series. *Theor Comput Sci* 359:440–448

- Ryabko D (2010) Clustering processes. In: Proc the 27th international conference on machine learning (ICML 2000), Haifa, Israel, pp 919–926.
- Ryabko D (2010) Discrimination between B -processes is impossible. *J Theor Probab* 23(2):565–575
- Ryabko D (2010) Testing composite hypotheses about discrete-valued stationary processes. In: Proc. IEEE information theory workshop (ITW'10), IEEE, Cairo, Egypt, pp 291–295
- Ryabko D, Ryabko B (2010) Nonparametric statistical inference for ergodic B processes. *IEEE Trans Inf Theory* 56(3):1430–1435