A criterion for hypothesis testing for stationary processes

Daniil Ryabko

A criterion for hypothesis testing for stationary processes

Daniil Ryabko

INRIA Lille-Nord Europe,
40, Avenue Halley 59650 Villeneuve d'Ascq, France
`daniil@ryabko.net`

Abstract

Given a finite-valued sample $X_1, \ldots, X_n$ we wish to test whether it was generated by a stationary ergodic process belonging to a family H_0, or it was generated by a stationary ergodic process outside H_0. We require the Type I error of the test to be uniformly bounded, while the type II error has to be mande not more than a finite number of times with probability 1. For this notion of consistency we provide necessary and sufficient conditions on the family H_0 for the existence of a consistent test. This criterion is illustrated with applications to testing for a membership to parametric families, generalizing some existing results. In addition, we analyze a stronger notion of consistency, which requires finite-sample guarantees on error of both types, and provide some necessary and some sufficient conditions for the existence of a consistent test. We emphasize that no assumption on the process distributions are made beyond stationarity and ergodicity.

Keywords: Hypothesis testing, stationary processes, ergodic processes, distributional distance.

1 Introduction

Given a sample $X_1, \ldots, X_n$ (where X_i are from a finite alphabet A) that is known to be generated by a stationary ergodic process, we wish to decide whether it was generated by a distribution belonging to a certain family H_0, versus it was generated by a stationary ergodic distribution that does not belong to H_0. 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.

A test is a function that takes a sample and an additional parameter α (the significance level), and gives a binary (possibly incorrect) answer: the sample was generated by a distribution from H_0 or by a stationary ergodic distribution not belonging to H_0. Here we are concerned with characterizing those families H_0 for which consistent tests exist.

We consider the following notion of consistency. Call a test *consistent* if, for any pre-specified *level* $\alpha \in (0, 1)$, any sample size n and any distribution in H_0 *the probability of Type I error (the test says "not H_0") is not greater than α,* while for every stationary ergodic distribution from outside H_0 and every α *Type II error (the test says H_0) is made only a finite number of times (as the sample size goes to infinity) with probability 1.* This notion of consistency represents a classical statistical approach to the problem, and suites well situations where the hypothesis H_0 is considerably more simple than the alternative, for example when H_0 consists of just one distribution, or when it is some parametric family, or when it is the hypothesis of homogeneity or that of independence.

Prior work. There is a vast body of literature on hypothesis testing for i.i.d. (real- or discrete-valued) data (see e.g. [8]). 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 to obtain: there is a consistent test for H_0 (against "i.i.d. but not H_0") if and only if H_0 is closed, where the topology is that of the parameter space (probabilities of each symbol), e.g. see [4]. The consistency being easy to ensure, the prime concern for the case of i.i.d. data is optimality.

There is, however, much less literature on hypothesis testing beyond i.i.d. or parametric models, while the questions of determining whether a consistent test exists (for different notions of consistency and different hypotheses) 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 the alternative), [7] 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. For the same notion of consistency, [10] gives sufficient conditions on two families H_0 and H_1 that consist of stationary ergodic real-valued processes, under which a consistent continuous test exists, extending the results of [5] for i.i.d. data. The latter condition is that H_0 and H_1 are contained in disjoint F_σ sets (countable unions of closed sets), with respect to the topology of weak convergence. For the notion of consistency that we consider, consistent tests for some specific hypotheses, but under the general alternative of stationary ergodic processes, have been proposed in [11, 12, 14], which address problems of testing identity, independence, estimating the order of a Markov process, and also the change point problem. Some impossibility results for testing hypotheses about stationary ergodic processes can be found in [9, 13].

The results. The aim of this work is to provide topological characterizations of the hypotheses for which consistent tests exist, for the case of stationary ergodic distributions. The obtained characterization is rather similar to those mentioned above for the case of i.i.d. data, but is with respect to the topology of distributional distance (or weak convergence). The fact that necessary and sufficient conditions are obtained indicates that this topology is the right one to consider.

A distributional distance between two process distributions is defined as a weighted sum 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 main result

is the following theorem (formalized in the next sections).

Theorem. There exists a consistent test for H_0 if and only if H_0 has probability 1 with respect to ergodic decomposition of every distribution from the closure of H_0.

The test that we construct to establish this result is based on empirical estimates of distributional distance. For a given level α, it takes the largest ε-neighbourhood of the closure of H_0 that has probability not greater than $1 - \alpha$ with respect to every ergodic process in it, and outputs 0 if the sample falls into this neighbourhood, and 1 otherwise.

To illustrate the applicability of the main result, we show that the families of k-order Markov processes and k-state Hidden Markov processes (for any natural k), satisfy the conditions of the theorem, and therefore there exists a consistent test for membership to these families.

It should be emphasized that the results of this work concern what is possible in principle; finding an efficient testing procedure for each specific hypothesis for which we can demonstrate existence of a consistent test is a different problem.

2 Preliminaries

Let A be a finite alphabet, and denote A^* the set of words (or tuples) $\cup_{i=1}^{\infty} A^i$ and A^{∞} the set of all one-way infinite sequences. For a word $B \in A^*$ the symbol $|B|$ stands for the length of B. Distributions, or (stochastic) processes, are measures on the space $(A^{\infty}, \mathcal{F}_{A^{\infty}})$, where $\mathcal{F}_{A^{\infty}}$ is the Borel sigma-algebra of A^{∞}. Denote $\#(X, B)$ the number of occurrences of a word $B \in A^*$ in a word $X \in A^*$ and $\nu(X, B)$ its frequency:

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

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

We use the abbreviation $X_{1..k}$ for $X_1, \ldots, X_k$. 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 $\mathcal{S}$ the set of all stationary processes on A^{∞}. A stationary process ρ is called *(stationary) ergodic* if the frequency of occurrence of each word B in a sequence $X_1, X_2, \ldots$ generated by ρ tends to its a priori (or limiting) probability a.s.: $\rho(\lim_{n \to \infty} \nu(X_{1..n}, B) = \rho(X_{1..|B|} = B)) = 1$. By virtue of the ergodic theorem (e.g. [3]), 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. [4]). Denote $\mathcal{E}$ the set of all stationary ergodic processes.

Definition 1 (distributional distance)**.** *The distributional distance is defined for a pair of processes ρ_1, ρ_2 as follows [6]:*

$$d(\rho_1, \rho_2) = \sum_{k=1}^{\infty} w_k |\rho_1(X_{1..|B_k|} = B_k) - \rho_2(X_{1..|B_k|} = B_k)|,$$

where $w_k = 2^{-k}$ and B_k, $k \in \mathbb{N}$ range through the set A^* of all words in length-lexicographical order (the weights and ordering are fixed for the sake of concreteness only).

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's 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 [6] for more details and proofs of these facts.

When talking about closed and open subsets of $\mathcal{S}$ we assume the topology of d. 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.

Considering the Borel (with respect to the metric d) sigma-algebra $\mathcal{F}_\mathcal{S}$ on the set $\mathcal{S}$, we obtain a standard probability space $(\mathcal{S}, \mathcal{F}_\mathcal{S})$. An important tool that will be used in the analysis is **ergodic decomposition** of stationary processes (see e.g. [6, 3]): which we recall here. 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 $B \in \mathcal{F}_{A^\infty}$. The *support* of a stationary distribution ρ is the minimal closed set $U \subset \mathcal{S}$ such that $W_\rho(U) = 1$.

A **test** is a function $\psi^\alpha : A^* \rightarrow \{0, 1\}$ that takes as input a sample and a parameter $\alpha \in (0, 1)$, and outputs a binary answer, where the answer 0 is interpreted as "the sample was generated by a distribution that belongs to H_0", and the answer 1 as "the sample was generated by a stationary ergodic distribution that does not belong to H_0." 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_0 is false.

Definition 2 (consistency)**.** *Call a test $\psi^\alpha, \alpha \in (0, 1)$ consistent as a test of H_0 against H_1 if:*

  (i) The probability of Type I error is always bounded by α: $\rho\{X \in A^n : \psi^\alpha(X) = 1\} \leq \alpha$ for every $\rho \in H_0$, every $n \in \mathbb{N}$ and every $\alpha \in (0, 1)$, and

  (ii) Type II error is made not more than a finite number of times with probability 1: $\rho(\lim_{n \to \infty} \psi^\alpha(X_{1..n}) = 1) = 1$ for every $\rho \in H_1$ and every $\alpha \in (0, 1)$.

4

3 Main results

The test constructed below is based on *empirical estimates of the distributional distance d*:

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

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

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

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

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

Moreover, let

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

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

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

We will often omit the subscript H_0 from $\psi_{H_0}^{\alpha}$ when it can cause no confusion.

The main result of this work is the following theorem, whose proof is given in section 6.

Theorem 1. *Let $H_0 \subset \mathcal{E}$. The following statements are equivalent:*

(i) *There exists a consistent test for H_0 against $\mathcal{E} \backslash H_0$.*

(ii) *The test $\psi_{H_0}^{\alpha}$ is consistent.*

(iii) *The set H_0 has probability 1 with respect to ergodic decomposition of every ρ in the closure of H_0: $W_{\rho}(H_0) = 1$ for each $\rho \in \operatorname{cl} H_0$.*

4 Examples

The first simple illustration of Theorem 1 above is identity testing, or goodness of fit: testing whether a distribution generating the sample obeys a certain given law, versus it does not. Let $\rho \in \mathcal{E}$, $H_0 = \{\rho\}$. Since H_0 is closed, Theorem 1 implies that there is a consistent test for H_0. 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. [8]), and Markov chains [2]. For stationary ergodic processes, [12] gives a consistent test when H_0 has a finite and bounded memory, and [14] for the general case.

Another example is bounding the order of a Markov or a Hidden Markov process. Theorem 1 implies that for any given $k \in \mathbb{N}$ there is a consistent test of the hypothesis $\mathcal{M}^k=$ "the process is Markov of order not greater than k" (against $\mathcal{E} \backslash \mathcal{M}^k$). Moreover, there is a consistent test of $\mathcal{H}\mathcal{M}^k=$"the process is given by a Hidden Markov process with not more than k states." Indeed, in both cases (k-order Markov, Hidden Markov with not more than k states), the hypothesis H_0 is a parametric family, with a compact set of parameters, and a continuous function mapping parameters to processes (that is, to the space $\mathcal{S}$). Weierstrass theorem then implies that the image of such a compact parameter set is closed (and compact). Moreover, in both cases 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 [1], see also [2]. For a weaker notion of consistency (the test has to stabilize on the correct answer eventually, with probability 1) the existence of a consistent test for $\mathcal{H}\mathcal{M}^k$ was established in [7]. For the notion of consistency considered here, a consistent test for $\mathcal{M}^k$ was proposed in [11], while for the case of testing $\mathcal{H}\mathcal{M}^k$ the result above is apparently new.

5 Uniform testing

Finally, let us consider a stronger notion of hypothesis testing, that requires uniform speed of convergence for errors of either type.

A test φ is called **uniformly consistent** if for every α there is an $n_\alpha \in \mathbb{N}$ such that for every $n \geq n_\alpha$ the probability of error on a sample of size n is less than α: $\rho(X \in A^n : \varphi(X) = i) < \alpha$ for every $\rho \in H_{1-i}$ and every $i \in \{0, 1\}$.

For $H_0, H_1 \subset \mathcal{S}$, the **uniform test** φ_{H_0,H_1} is constructed as follows. For each $n \in \mathbb{N}$ let

$$\varphi_{H_0,H_1}(X_{1..n}) := \begin{cases} 0 & \text{if } \hat{d}(X_{1..n}, \mathrm{cl}\, H_0 \cap \mathcal{E}) < \hat{d}(X_{1..n}, \mathrm{cl}\, H_1 \cap \mathcal{E}), \\ 1 & \text{otherwise.} \end{cases} \quad (2)$$

Theorem 2 (uniform testing). *Let $H_0 \subset \mathcal{S}$ and $H_1 \subset \mathcal{S}$. If $W_\rho(H_i) = 1$ for every $\rho \in \mathrm{cl}\, H_i$ then the test φ_{H_0,H_1} is uniformly consistent. Conversely, if there exists a uniformly consistent test for H_0 against H_1 then $W_\rho(H_{1-i}) = 0$ for any $\rho \in clH_i$.*

The proof is given in the next section.

6 Proofs

The proof of the main results will use the following lemmas.

6

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

$$\lim_{k \to \infty} \hat{d}(X_{1..k}, \xi) = d(\rho, \xi) \ \rho\text{-}a.s.$$

The proof is based on the fact that the frequency of each word converges to its expectation. For each δ we can find a time by which the first $K(\delta)$ frequencies will have converged up to δ, where $K(\delta)$ is such that the cumulative weight of the rest of the frequencies is smaller than δ too.

Proof. For any $\varepsilon > 0$ find such an index J that $\sum_{i=J}^{\infty} w_i < \varepsilon/2$. For each j we have $\lim_{k \to \infty} \nu(X_{1..k}, B_j) = \rho(B_j)$ a.s., so that $|\nu(X_{1..k}, B_j) - \rho(B_j)| < \varepsilon/(2Jw_j)$ from some k on; denote K_j this k. Let $K = \max_{j<J} K_j$ (K depends on the realization $X_1, X_2, \dots$). Thus, for $k > K$ we have

$$|\hat{d}(X_{1..k}, \xi) - d(\rho, \xi)| = \left| \sum_{i=1}^{\infty} w_i \big(|\nu(X_{1..k}, B_i) - \xi(B_i)| - |\rho(B_i) - \xi(B_i)| \big) \right|$$

$$\leq \sum_{i=1}^{\infty} w_i |\nu(X_{1..k}, B_i) - \rho(B_i)| \leq \sum_{i=1}^{J} w_i |\nu(X_{1..k}, B_i) - \rho_X(B_i)| + \varepsilon/2$$

$$\leq \sum_{i=1}^{J} w_i \varepsilon/(2Jw_i) + \varepsilon/2 = \varepsilon,$$

which proves the statement. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ $\square$

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

$$\rho(\hat{d}(X_{1..m}, H) \geq \varepsilon) \leq 2\varepsilon'^{-1} \rho(\hat{d}(X_{1..k}, H) \geq \varepsilon'), \qquad (3)$$

where $\varepsilon' := \varepsilon - \frac{2k}{m-k+1} - t_k$ with t_k being the sum of all the weights of tuples longer than k in the definition of d: $t_k := \sum_{i:|B_i|>k} w_i$. Further,

$$\rho(\hat{d}(X_{1..m}, H) \leq \varepsilon) \leq 2\rho \left(\hat{d}(X_{1..k}, H) \leq \frac{m}{m-k+1} 2\varepsilon + \frac{4k}{m-k+1} \right). \qquad (4)$$

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 $X_{i..i+k}$ are far from (close to) μ too. 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:

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

$$= \sum_{i=1}^{m-k+1} \nu(X_{i..i+k-1}, B) + 2k.$$

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 the definition (1), we obtain

$$\nu(X_{1..m}, B) \leq \frac{1}{m - |B| + 1} \left(\sum_{i=1}^{m-k+1} \nu(X_{i..i+k-1}, B)| + 2k \right). \tag{5}$$

Summing over all B, for any μ, we get

$$\hat{d}(X_{1..m}, \mu) \leq \frac{1}{m - k + 1} \sum_{i=1}^{m-k+1} \hat{d}(X_{i..i+n-1}, \mu) + \frac{2k}{m - k + 1} + t_k, \tag{6}$$

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

$$\hat{d}(X_{1..m}, H) \leq \frac{1}{m - k + 1} \left(\sum_{i=1}^{m-k+1} \hat{d}(X_{i..i+k-1}, H) \right) + \frac{2k}{m - k + 1} + t_k. \tag{7}$$

Note that the $\hat{d}(X_{i..i+k-1}, H) \in [0, 1]$. Therefore, for the average in the r.h.s. of (7) to be larger than ε', at least $\varepsilon'/2(m - k + 1)$ summands have to be larger than $\varepsilon'/2$.

Using stationarity, we can conclude

$$\rho\left(\hat{d}(X_{1..k}, H) \geq \varepsilon' \right) \geq \varepsilon'/2\rho\left(\hat{d}(X_{1..m}, H) \geq \varepsilon \right),$$

proving (3). The second statement can be proven similarly; indeed, analogously to (5) we have

$$\nu(X_{1..m}, B) \geq \frac{1}{m - |B| + 1} \sum_{i=1}^{m-k+1} \nu(X_{i..i+k-1}, B) - \frac{2k}{m - |B| + 1}$$

$$\geq \frac{1}{m - k + 1} \left(\frac{m - k + 1}{m} \sum_{i=1}^{m-k+1} \nu(X_{i..i+k-1}, B) \right) - \frac{2k}{m},$$

8

where we have used $|B| \geq 1$. Summing over different B, we obtain (similar to (6)),

$$\hat{d}(X_{1..m}, \mu) \geq \frac{1}{m-k+1} \sum_{i=1}^{m-k+1} \frac{m-k+1}{m} \hat{d}_k(X_{i..i+n-1}, \mu) - \frac{2k}{m} \qquad (8)$$

(since the frequencies are non-negative, there is no t_n term here). For the average in (8) to be smaller than ε, at least half of the summands must be smaller than 2ε. Using stationarity of ρ, this implies (4). $\qquad \square$

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. $\qquad \square$

Proof of Theorem 1. The implication *(ii)* $\Rightarrow$ *(i)* is obvious. We will show *(iii)* $\Rightarrow$ *(ii)* and *(i)* $\Rightarrow$ *(iii)*. To establish the former, we have to show that the family of tests ψ^α is consistent. By construction, for any $\rho \in \operatorname{cl} H_0 \cap \mathcal{E}$ we have $\rho(\psi^\alpha(X_{1..n}) = 1) \leq \alpha$.

To prove the consistency of ψ, it remains to show that $\xi(\psi^\alpha(X_{1..n}) = 0) \to 0$ a.s. for any $\xi \in \mathcal{E} \backslash H_0$ and $\alpha > 0$. To do this, fix any $\xi \in \mathcal{E} \backslash H_0$ and let $\Delta := d(\xi, \operatorname{cl} H_0) := \inf_{\rho \in \operatorname{cl} H_0 \cap \mathcal{E}} d(\xi, \rho)$. Since $\operatorname{cl} H_0$ is closed, we have $\Delta > 0$. Suppose that there exists an $\alpha > 0$, such that, for infinitely many n, some samples from the $\Delta/2$-neighbourhood of n-samples around ξ are sorted as H_0 by ψ, that is, $C^n(\operatorname{cl} H_0 \cap \mathcal{E}, 1 - \alpha) \cap b^n_{\Delta/2}(\xi) \neq \varnothing$. Then for these n we have $\gamma_n(\operatorname{cl} H_0 \cap \mathcal{E}, 1 - \alpha) \geq \Delta/2$.

This means that there exists an increasing sequence $n_m, m \in \mathbb{N}$, and a sequence $\rho_m \in \operatorname{cl} H_0$, $m \in \mathbb{N}$, such that

$$\rho_m(\hat{d}(X_{1..n_m}, \operatorname{cl} H_0 \cap \mathcal{E}) > \Delta/2) > \alpha.$$

Using Lemma 2, (3) (with $\rho = \rho_m$, $m = n_m$, $k = n_k$, and $H = \operatorname{cl} H_0$), and taking k large enough to have $t_{n_k} < \Delta/4$, for every m large enough to have $\frac{2n_k}{n_m - n_k + 1} < \Delta/4$, we obtain

$$8\Delta^{-1} \rho_m \left(\hat{d}(X_{1..n_k}, \operatorname{cl} H_0) \geq \Delta/4 \right) \geq \rho_m \left(\hat{d}(X_{1..n_m}, \operatorname{cl} H_0) \geq \Delta/2 \right) > \alpha. \qquad (9)$$

Thus,

$$\rho_m(b^{n_k}_{\Delta/4}(\operatorname{cl} H_0 \cap \mathcal{E})) < 1 - \alpha\Delta/8. \qquad (10)$$

Since the set $\operatorname{cl} H_0$ is compact (as a closed subset of a compact set $\mathcal{S}$), we may assume (passing to a subsequence, if necessary) that ρ_m converges to a certain $\rho_* \in \operatorname{cl} H_0$. Since (10) this holds for infinitely many m, using Lemma 3 (with $T = b^{n_k}_{\Delta/4}(\operatorname{cl} H_0 \cap \mathcal{E})$) we conclude that

$$\rho_*(b^{n_k}_{\Delta/4}(\operatorname{cl} H_0 \cap \mathcal{E})) \leq 1 - \Delta\alpha/8.$$

9

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

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

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

This contradiction shows that for every α there are not more than finitely many n for which $C^n(\mathrm{cl}\, H_0 \cap \mathcal{E}, 1 - \alpha) \cap b_{\Delta/2}^n(\xi) \neq \varnothing$. To finish the proof of the implication, it remains to note that, as follows from Lemma 1,

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

To establish the implication $(i) \Rightarrow (iii)$, we assume that there exists a consistent test φ for H_0, and we will show that $W_\rho(\mathcal{E} \backslash H_0) = 0$ for every $\rho \in \mathrm{cl}\, H_0$. Take $\rho \in \mathrm{cl}\, H_0$ and suppose that $W_\rho(\mathcal{E} \backslash H_0) = \delta > 0$. We have

$$\limsup_{n\to\infty} \int_{\mathcal{E} \backslash H_0} dW_\rho(\mu)\mu(\psi_n^{\delta/2} = 0) \leq \int_{\mathcal{E} \backslash H_0} \limsup_{n\to\infty} dW_\rho(\mu)\mu(\psi_n^{\delta/2} = 0) = 0,$$

where the inequality follows from Fatou's lemma (the functions under integral are all bounded by 1), and the equality from the consistency of ψ. Thus, from some n on we will have $\int_{\mathcal{E} \backslash H_0} dW_\rho\mu(\psi_n^{\delta/2} = 0) < 1/4$ 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 μ. In particular, it holds for the set $T := \{X_{1..n} : \psi_n^{\delta/2}(X_{1..n}) = 0\}$. Therefore, since $\rho \in \mathrm{cl}\, H_0$, 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, $W_\rho(H_0) = 1$, and Theorem 1 is proven. □

Proof of Theorem 2. To prove the first statement of the theorem, we will show that the test φ_{H_0, H_1} is a uniformly consistent test for $\mathrm{cl}\, H_0 \cap \mathcal{E}$ against $\mathrm{cl}\, H_1 \cap \mathcal{E}$ (and hence for H_0 against H_1), under the conditions of the theorem. Suppose that, on the contrary, for some $\alpha > 0$ for every $n' \in \mathbb{N}$ there is a process $\rho \in \mathrm{cl}\, H_0$ such that $\rho(\varphi(X_{1..n}) = 1) > \alpha$ for some $n > n'$. Define

$$\Delta := d(\mathrm{cl}\, H_0, \mathrm{cl}\, H_1) := \inf_{\rho_0 \in \mathrm{cl}\, H_0 \cap \mathcal{E}, \rho_1 \in \mathrm{cl}\, H_1 \cap \mathcal{E}} d(\rho_0, \rho_1),$$

which is positive since $\mathrm{cl}\, H_0$ and $\mathrm{cl}\, H_1$ are closed and disjoint. We have

$$\alpha < \rho(\varphi(X_{1..n}) = 1)$$
$$\leq \rho(\hat{d}(X_{1..n}, H_0) \geq \Delta/2 \text{ or } \hat{d}(X_{1..n}, H_1) < \Delta/2)$$
$$\leq \rho(\hat{d}(X_{1..n}, H_0) \geq \Delta/2) + \rho(\hat{d}(X_{1..n}, H_1) < \Delta/2). \quad (11)$$

This implies that either $\rho(\hat{d}(X_{1..n}, \mathrm{cl}\, H_0) \geq \Delta/2) > \alpha/2$ or $\rho(\hat{d}(X_{1..n}, \mathrm{cl}\, H_1) < \Delta/2) > \alpha/2$, so that, by assumption, at least one of these inequalities holds for infinitely many $n \in \mathbb{N}$ for some sequence $\rho_n \in H_0$. Suppose that it is the first one, that is, there is an increasing sequence n_i, $i \in \mathbb{N}$ and a sequence $\rho_i \in \mathrm{cl}\, H_0$, $i \in \mathbb{N}$ such that

$$\rho_i(\hat{d}(X_{1..n_i}, \mathrm{cl}\, H_0) \geq \Delta/2) > \alpha/2 \text{ for all } i \in \mathbb{N}. \tag{12}$$

The set $\mathcal{S}$ is compact, hence so is its closed subset $\mathrm{cl}\, H_0$. Therefore, the sequence ρ_i, $i \in \mathbb{N}$ must contain a subsequence that converges to a certain process $\rho_* \in \mathrm{cl}\, H_0$. Passing to a subsequence if necessary, we may assume that this convergent subsequence is the sequence ρ_i, $i \in \mathbb{N}$ itself.

Using Lemma 2, (3) (with $\rho = \rho_{n_m}$, $m = n_m$, $k = n_k$, and $H = \mathrm{cl}\, H_0$), and taking k large enough to have $t_{n_k} < \Delta/4$, for every m large enough to have $\frac{2n_k}{n_m - n_k + 1} < \Delta/4$, we obtain

$$8\Delta^{-1}\rho_{n_m}\left(\hat{d}(X_{1..n_k}, \mathrm{cl}\, H_0) \geq \Delta/4\right) \geq \rho_{n_m}\left(\hat{d}(X_{1..n_m}, \mathrm{cl}\, H_0) \geq \Delta/2\right) > \alpha/2. \tag{13}$$

That is, we have shown that for any large enough index n_k the inequality $\rho_{n_m}(\hat{d}(X_{1..n_k}, \mathrm{cl}\, H_0) \geq \Delta/4) > \Delta\alpha/16$ holds for infinitely many indices n_m. From this and Lemma 3 with $T = T_k := \{X : \hat{d}(X_{1..n_k}, \mathrm{cl}\, H_0) \geq \Delta/4\}$ we conclude that $\rho_*(T_k) > \Delta\alpha/16$. The latter holds for infinitely many k; that is, $\rho_*(\hat{d}(X_{1..n_k}, \mathrm{cl}\, H_0) \geq \Delta/4) > \Delta\alpha/16$ infinitely often. Therefore,

$$\rho_*(\limsup_{n\to\infty} d(X_{1..n}, \mathrm{cl}\, H_0) \geq \Delta/4) > 0.$$

However, we must have

$$\rho_*(\lim_{n\to\infty} d(X_{1..n}, \mathrm{cl}\, H_0) = 0) = 1$$

for every $\rho_* \in \mathrm{cl}\, H_0$: indeed, for $\rho_* \in \mathrm{cl}\, H_0 \cap \mathcal{E}$ it follows from Lemma 1, and for $\rho_* \in \mathrm{cl}\, H_0 \backslash \mathcal{E}$ from Lemma 1, ergodic decomposition and the conditions of the theorem.

Thus, we have arrived at a contradiction that shows that $\rho_n(\hat{d}(X_{1..n}, \mathrm{cl}\, H_0) > \Delta/2) > \alpha/2$ cannot hold for infinitely many $n \in \mathbb{N}$ for any sequence of $\rho_n \in \mathrm{cl}\, H_0$. Analogously, we can show that $\rho_n(\hat{d}(X_{1..n}, \mathrm{cl}\, H_1) < \Delta/2) > \alpha/2$ cannot hold for infinitely many $n \in \mathbb{N}$ for any sequence of $\rho_n \in \mathrm{cl}\, H_0$. Indeed, using Lemma 2, equation (4), we can show that $\rho_{n_m}(\hat{d}(X_{1..n_m}, \mathrm{cl}\, H_1) \leq \Delta/2) > \alpha/2$ for a large enough n_m implies $\rho_{n_m}(\hat{d}(X_{1..n_k}, \mathrm{cl}\, H_1) \leq 3\Delta/4) > \alpha/4$ for a smaller n_k. Therefore, if we assume that $\rho_n(\hat{d}(X_{1..n}, \mathrm{cl}\, H_1) < \Delta/2) > \alpha/4$ for infinitely many $n \in \mathbb{N}$ for some sequence of $\rho_n \in \mathrm{cl}\, H_0$, then we will also find a ρ_* for which $\rho_*(\hat{d}(X_{1..n}, \mathrm{cl}\, H_1) \leq 3\Delta/4) > \alpha/4$ for infinitely many n, which, using Lemma 1 and ergodic decomposition, can be shown to contradict the fact that $\rho_*(\lim_{n\to\infty} d(X_{1..n}, \mathrm{cl}\, H_1) \geq \Delta) = 1$.

Thus, returning to (11), we have shown that from some n on there is no $\rho \in \mathrm{cl}\, H_0$ for which $\rho(\varphi = 1) > \alpha$ holds true. The statement for $\rho \in \mathrm{cl}\, H_1$ can be proven analogously, thereby finishing the proof of the first statement.

11

To prove the second statement of the theorem, we assume that there exists a uniformly consistent test φ for H_0 against H_1, and we will show that $W_\rho(H_{1-i}) = 0$ for every $\rho \in \mathrm{cl}\, H_i$. Indeed, let $\rho \in \mathrm{cl}\, H_0$, that is, suppose that there is a sequence $\xi_i \in H_0, i \in \mathbb{N}$ such that $\xi_i \to \rho$. Assume $W_\rho(H_1) = \delta > 0$ and take $\alpha := \delta/2$. Since the test φ is uniformly consistent, there is an $N \in \mathbb{N}$ such that for every $n > N$ we have

$$\rho(\varphi(X_{1..n} = 0)) \leq \int_{H_1} \varphi(X_{1..n} = 0)dW_\rho + \int_{\mathcal{E} \setminus H_1} \varphi(X_{1..n} = 0)dW_\rho$$
$$\leq \delta\alpha + 1 - \delta \leq 1 - \delta/2.$$

Recall that, for $T \in A^*$, $\mu(T)$ is a continuous function in μ. In particular, this holds for the set $T = \{X \in A^n : \varphi(X) = 0\}$, for any given $n \in \mathbb{N}$. Therefore, for every $n > N$ and for every i large enough, $\rho_i(\varphi(X_{1..n}) = 0) < 1 - \delta/2$ implies also $\xi_i(\varphi(X_{1..n}) = 0) < 1 - \delta/2$ which contradicts $\xi_i \in H_0$. This contradiction shows $W_\rho(H_1) = 0$ for every $\rho \in \mathrm{cl}\, H_0$. The case $\rho \in \mathrm{cl}\, H_1$ is analogous. $\qquad\square$

References

[1] T. Anderson, L. Goodman. Statistical Inference about Markov Chains, Ann. Math. Statist. Vol. 28(1), pp. 89-110, 1957.

[2] P. Billingsley, Statistical methods in Markov chains, Ann. Math. Statist. Vol. 32(1), pp. 12-40, 1961.

[3] P. Billingsley, Ergodic theory and information. Wiley, New York, 1965.

[4] I. Csiszár, P. Shields, Notes on Information Theory and Statistics: A tutorial, Foundations and Trends in Communications and Information Theory (1), pp. 1–111. 2004.

[5] A. Dembo, Y. Peres. A topological criterion for hypothesis testing. Ann. Math. Stat. Vol. 22, pp. 106-117, 1994.

[6] R. Gray. Probability, Random Processes, and Ergodic Properties. Springer Verlag, 1988.

[7] J.C. Kieffer, Strongly consistent code-based identification and order estimation for constrained finite-state model classes, IEEE Transactions on Information Theory, Vol. 39(3), pp. 893-902, 1993.

[8] E. Lehmann, Testing Statistical Hypotheses, 2nd edition, John Wiley & Sons, New York, 1986.

[9] Morvai G., Weiss B. (2005) On classifying processes. *Bernoulli*, vol. 11, no. 3, pp. 523–532.

[10] A. Nobel, Hypothesis testing for families of ergodic processes. Bernoulli, vol. 12(2), pp. 251-269, 2006.

[11] 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.

[12] 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, v.359, pp.440-448, 2006.

[13] D. Ryabko, An impossibility result for process discrimination. In Proceedings of IEEE International Symposium on Information Theory (ISIT'09), pp. 1734-1738, Seoul, South Korea, 2009.

[14] D. Ryabko, B. Ryabko. On Hypotheses Testing for Ergodic Processes In Proceedings of IEEE Information Theory Workshop (ITW'08), Porto, Portugal, pp. 281-283, 2008. see also http://arxiv.org/abs/0804.0510