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 discrete-valued sample $X_1, \dots, X_n$ we wish to test whether it was generated by a process belonging to a family H_0, or it was generated by a process outside H_0. All process distributions are assumed stationary ergodic, and no further probabilistic or parametric assumptions are made. We require the Type I error of the test to be uniformly bounded, while the probability of Type II error has to tend to zero as the sample size increases. 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.

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

1 Introduction

Given a sample $X_1, \dots, X_n$ (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 the probability of Type II error (the test says H_0) goes to 0, as the sample size goes to infinity. 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 distributions, or when it is some parametric family, or when it is the hypothesis of homogeneity or that of independence. It is also worth noting that in many practical situations the Type I and Type II errors may have very different meaning: for example, this is the case when H_0 is interpreted as that a patient has a certain ailment, and the alternative as that he does not. In such cases, it is natural to require a hard guarantee on the error of one type.

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 obvious: there is a consistent test if and only if H_0 is closed, where the topology is that of the parameter space (probabilities of each symbol), see e.g. [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, [9] gives sufficient conditions on two families H_0 and H_1 that consist of stationary ergodic real-valued processes, under which a consistent 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 [10, 11, 12], which address problems of testing identity, independence, estimating the order of a Markov process, and also the change point problem.

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 applicability of the main result, we show that any set of processes which is continuously parametrized by a compact set of parameters, and is closed under taking ergodic decompositions, can be consistently tested against its complement to the set of all stationary ergodic processes. Such parametric families include k-order Markov processes, k-state Hidden Markov processes (for any natural k), and many others.

2 Preliminaries

Let A be a finite alphabet, and denote A^* the set of words (or tuples) $\cup_{i=1}^{\infty} A^i$. For a word B the symbol $|B|$ stands for the length of B. Denote B_i the ith 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 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 word $X \in A^*$ and $\nu(X,B)$ 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 $\mathcal{S}$ 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 $X_1, X_2, \dots$ 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.

A **distributional distance** is defined for a pair of processes ρ_1, ρ_2 as follows [6]:

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

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 a compact, and the set of stationary processes $\mathcal{S}$ is its convex closed subset. (The set $\mathcal{E}$ is not closed.) When talking about closed and open subsets of $\mathcal{S}$ we assume the topology of d. For $H \subset \mathcal{S}$, denote $\operatorname{cl} H$ the closure of H.

Compactness of the set $\mathcal{S}$ is one of the main ingredients of the analysis in this work. Another is that the distance d can be consistently estimated, as is demonstrated in Lemma 1 of section 5 below (see also [12]).

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]): 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 $\psi^\alpha : A^* \to \{0, 1\}$ 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.

Call a test $\psi^\alpha, \alpha \in (0, 1)$ **consistent** 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 $\rho \in \mathcal{E} \backslash H_0$ and every $\alpha \in (0, 1)$.

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

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

Moreover, let

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

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 $C^n(H, \theta) := b^n_{\gamma_n(H,\theta)}(H)$ be a neighbourhood of this radius. Define

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

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

The main result of this work is the following theorem, whose proof is given in section 5.

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

(i) There exists a consistent test for H_0.

(ii) The test $\psi^\alpha_{H_0}$ 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 Applications: testing membership to parametric families

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, [11] gives a consistent test when H_0 has a finite and bounded memory, and [12] 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 $\mathcal{HM}^k =$"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 $\mathcal{HM}^k$ was established in [7]. For the notion of consistency considered here, a consistent test for $\mathcal{M}^k$ was proposed in [10], while for the case of testing $\mathcal{HM}^k$ the result above is apparently new.

Moreover, from the discussion above one can see that the following generalization is valid. Let $H_0 \subset \mathcal{S}$ be a set of processes that is continuously parametrized by a compact set of parameters. If H_0 is closed under taking ergodic decompositions, then there is a consistent test for H_0. In particular, this strengthens the mentioned result of [7], since a stronger notion of consistency is used, as well as a more general class of parametric families is considered.

5 Proofs

The proofs will use the following lemmas.

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. □

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 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:

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

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

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

Summing over all B, for any μ, we get

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

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 $|B|\le k$. Since this holds for any μ, we conclude that

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

Therefore, for any $X_{1..m}\in A^m$, if $\hat{d}(X_{1..m},H) > \varepsilon$ then there is an index $i\le m-k$ 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

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

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

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

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

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

(since the frequencies are non-negative, there is no t_n term here), which, using stationarity of ρ, implies (3). □

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. □

Proof of Theorem 1. The implication *(ii)* ⇒ *(i)* is obvious. We will show *(iii)* ⇒ *(ii)* and *(i)* ⇒ *(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 $n_k, k \in \mathbb{N}$, and a sequence $\rho_k \in \operatorname{cl} H_0$, $k \in \mathbb{N}$, such that

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

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_k converges to a certain $\rho_* \in \operatorname{cl} H_0$. 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

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

Since 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

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

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

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

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

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

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

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 $\rho \in \operatorname{cl} H_0$. Take $\rho \in \operatorname{cl} H_0$ and suppose that $W_\rho(\mathcal{E}\backslash H_0) = \delta > 0$. We have

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

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 T. In particular, it holds for the set $T := \{X_{1..n} : \psi_n^{\delta/2}(X_{1..n}) = 0\}$. Therefore, since $\rho \in \operatorname{cl} H_0$, 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. □

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

[10] 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.

[11] 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.

[12] 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