An impossibility result for process discrimination

Daniil Ryabko

Discrimination between B-processes is impossible.

Daniil Ryabko *

Abstract

Two series of binary observations $x_1, x_1, \dots$ and $y_1, y_2, \dots$ are presented: at each time $n \in \mathbb{N}$ we are given x_n and y_n. It is assumed that the sequences are generated independently of each other by two B-processes. We are interested in the question of whether the sequences represent a typical realization of two different processes or of the same one. We demonstrate that this is impossible to decide, in the sense that every discrimination procedure is bound to err with non-negligible frequency when presented with sequences from some B-processes. This contrasts earlier positive results on B-processes, in particular those showing that there are consistent $\bar{d}$-distance estimates for this class of processes.

Keywords: Process discrimination, B-processes, stationary ergodic processes, time series, homogeneity testing

1 Introduction

Two series of binary observations $x_1, x_1, \dots$ and $y_1, y_2, \dots$ are presented sequentially. A *discrimination procedure* D is a family of mappings $D_n : X^n \times X^n \to \{0, 1\}$, $n \in \mathbb{N}$, that maps a pair of samples $(x_1, \dots, x_n)$, $(y_1, \dots, y_n)$ into a binary ("yes" or "no") answer: the samples are generated by different distributions, or they are generated by the same distribution.

A discrimination procedure D is *asymptotically correct for a set $\mathcal{C}$ of process distributions* if for any two distributions $\rho_x, \rho_y \in \mathcal{C}$ independently generating the sequences $x_1, x_2, \dots$ and $y_1, y_2, \dots$ correspondingly the expected output converges to the correct answer: the following limit exists and the equality holds

$$\lim_{n\to\infty} \mathbf{E} D_n((x_1, \dots, x_n), (y_1, \dots, y_n)) = \begin{cases} 0 & \text{if } \rho_x = \rho_y \\ 1 & \text{otherwise} \end{cases} .$$

Note that one can consider other notions of asymptotic correctness, for example one can require the output to stabilize on the correct answer with probability 1. The notion of correctness that we consider is perhaps one of the weakest. Clearly, asymptotically correct discriminating procedures exist for many classes of processes, for example for the class of all i.i.d. processes.

*INRIA, Lille, France. daniil@ryabko.net

Ornstein and Weiss [7] and Ornstein and Shields [6] show that consistent estimates of $\bar{d}$-distance for B-processes (see definitions below) exist, while it is impossible to estimate this distance outside this class (i.e. in general for stationary ergodic processes). We show that discrimination between B-processes is impossible, in the sense that any discrimination procedure is bound to err on some processes (the expected answer does not converge to the correct one). This demonstrates that discrimination is harder than distance estimation. This also complements earlier negative results on B-processes, such as [11] that shows that upper and lower divergence rates need not be the same for B-processes, and on stationary ergodic processes, such as [8, 2, 1], that establish negative results concerning prediction and density estimation.

The class of B-processes is sufficiently wide to include, for example, k-order Markov processes and functions of them, but, on the other hand, it is a strict subset of the set of stationary ergodic processes. B-processes play important role in such fields as information theory and ergodic theory [12, 13, 5]. Discrimination procedures for smaller classes of processes, such as the set of i.i.d. processes or various parametric families, exist and are widely studied (see e.g. [3]); some positive results on hypothesis testing for stationary ergodic process can be found in [4, 9, 10].

Next we define the $\bar{d}$ distance and B-processes, mainly following [7] in our formulations. For two finite-valued stationary processes ρ_x and ρ_y the *$\bar{d}$-distance* $\bar{d}(\rho_x, \rho_y)$ is said to be less than ε if there exists a single stationary process ν_{xy} on pairs (x_n, y_n), $n \in \mathbb{N}$, such that x_n, $n \in \mathbb{N}$ are distributed according to ρ_x and y_n are distributed according to ρ_y while

$$\nu_{xy}(x_1 \neq y_1) \leq \varepsilon. \tag{1}$$

The infimum of the ε's for which a coupling can be found such that (1) is satisfied is taken to be the $\bar{d}$-distance between ρ_x and ρ_y.

A process is called a *B-process* (or a Bernoulli process) if it is in the $\bar{d}$-closure of the set of all aperiodic stationary ergodic k-step Markov processes, where $k \in \mathbb{N}$. For more information on $\bar{d}$-distance and B-processes (including a more conventional ergodic-theory definition and its equivalence to the one above) the reader is referred to [5].

2 The main result

The main result of this work is the following theorem; the construction used in the proof is based on the same ideas as the construction used in [8] to demonstrate that consistent prediction for stationary ergodic processes is impossible, and to its modification in [2].

Theorem 1. *There is no asymptotically correct discrimination procedure for the set of all B-processes.*

Proof. We will assume that such a procedure D exists and will construct a B-process ρ such that if both sequences x_i and y_i, $i \in \mathbb{N}$ are generated by ρ then

$\mathbf{E}D_n$ diverges; this contradiction will prove the theorem.

The scheme of the proof is as follows. On Step 1 we construct a sequence of B-processes ρ_{2k}, ρ_{u2k+1}, and ρ_{u2k+1}, where $k = 0, 1, \ldots$. On Step 2 we construct a B-process ρ. On Step 3 we show that two independent runs of the process ρ have a property that (with high probability) they first behave like two runs of a single process ρ_0, then like two runs of two different processes ρ_{u1} and ρ_{d1}, then like two runs of a single process ρ_2, and so on, thereby showing that the test D diverges and obtaining the desired contradiction.

Each processes that we construct has the form of a stationary Markov chain with a countably infinite set of states, with a (deterministic) function mapping each state to $\{0, 1\}$. In other words, the constructions are based on partially observable Markov processes, where the observed variables are from $\{0, 1\}$ and the (non-observable) states are from the set $\mathbb{N}$.

The construction of each of the processes ρ_{2k}, ρ_{u2k+1}, ρ_{d2k+1}, for $k > 0$, and of the process ρ is broken into two parts: first it is given in terms of a chain on which Markov property is violated, and then, in order to define the initial distribution on this chain, and to show that the resulting process is stationary ergodic (and a B-process), we show that this chain is equivalent to a Markov chain, which has a stationary distribution with positive probabilities of all states.

Assume that there exists an asymptotically correct discriminating procedure D. Fix some $\varepsilon > 0$ and $\delta \in [1/2, 1)$, to be defined on Step 3.

Step 1. We will construct the sequence of process ρ_{2k}, ρ_{u2k+1}, and ρ_{u2k+1}, where $k = 0, 1, \ldots$. Step 1 consists of sub-steps 1.0, on which ρ_0 is constructed, 1.1, which constructs ρ_{u1} and ρ_{d1}, followed by 1.2 with ρ_2, and the step 1.k with ρ_{u3} and ρ_{d3}, ρ_4 and a general scheme of constructing the rest of the sequence.

Step 1.0. Construct the process ρ_0 as follows. A Markov chain m_0 is defined on the set $\mathbb{N}$ of states. From each state $i \in \mathbb{N}$ the chain passes to the state 0 with probability δ and to the state $i + 1$ with probability $1 - \delta$. With transition probabilities so defined, the chain possesses a unique stationary distribution M_0 on the set $\mathbb{N}$ of states such that $M_0(i) > 0$ for all $i \in \mathbb{N}$ (see e.g. [14]). Take this distribution as the initial distribution over the states.

The function f_0 maps the states to the output alphabet $\{0, 1\}$ as follows: $f_0(i) = 1$ for every $i \in \mathbb{N}$. Let s_t be the state of the chain at time t. The process ρ_0 is defined as $\rho_0 = f_0(s_t)$ for $t \in \mathbb{N}$. As a result of this definition, the process ρ_0 simply outputs 1 with probability 1 on every time step (however by using different functions f we will have less trivial processes in the sequel). Clearly, the constructed process is stationary ergodic and a B-process. So, we have defined the chain m_0 (and the process ρ_0) up to a parameter δ.

Step 1.1. We begin with the process ρ_0 and the chain m_0 of the previous step. Since the test D is asymptotically correct we will have

$$\mathbf{E}_{\rho_0 \times \rho_0} D_{t_0}((x_1, \ldots, x_{t_0}), (y_1, \ldots, y_{t_0})) < \varepsilon,$$

for some t_0, where both samples x_i and y_i are generated by ρ_0 (that is, both samples consist of 1s only). Let k_0 be such an index that the chain m_0 starting from the state 0 with probability 1 does not reach the state $k_0 - 1$ by time t_0 (we can take $k_0 = t_0 + 2$).

Construct two processes ρ_{u1} and ρ_{d1} as follows. They are also based on the Markov chain m_0, but the functions f are different. The function $f_{u1} : \mathbb{N} \to \{0, 1\}$ is defined as follows: $f_{u1}(i) = f_0(i) = 1$ for $i \le k_0$ and $f_{u1}(i) = 0$ for $i > k_0$. The function f_{d1} is identically 1 ($f_{d1}(i) = 1$, $i \in \mathbb{N}$). The processes ρ_{u1} and ρ_{d1} are defined as $\rho_{u1} = f_{u1}(s_t)$ and $\rho_{d1} = f_{d1}(s_t)$ for $t \in \mathbb{N}$. Thus the process ρ_{d1} will again produce only 1s, but the process ρ_{u1} will occasionally produce 0s. It is easy to check that the processes ρ_{u1} and ρ_{u1} are B-processes (cf. Step 2c below, where it is shown that the process ρ is a B-process).

Step 1.2. Being run on two samples generated by ρ_{u1} and ρ_{d1} the test D_n on the first t_0 steps with high probability (that is, at least if both processes start at the state 0) produces many 0s, since on these first k_0 states all the functions f, f_{u1} and f_{d1} coincide. However since the processes are different and the test is asymptotically correct (by assumption), the test starts producing 1s, until by a certain time step t_1 almost all answers are 1s. Next we will construct the process ρ_2 by "gluing" together ρ_{u1} and ρ_{d1} and continuing them in such a way that, being run on two samples produced by ρ_2 the test first produces 0s (as if the samples were drawn from ρ_0), then, with probability close to 1/2 it will produce many 1s (as if the samples were from ρ_{u1} and ρ_{d1}) and then again 0s.

The process ρ_2 is constructed in two steps. On step 1.2a the construction is given in the form of a chain on which Markov property is violated, and on step 1.2b we show that the construction is equivalent to a Markov chain. The stationary distribution on this chain will be used to finish the construction of ρ_2.

Step 1.2a. Let $t_1 > t_0$ be such a time index that

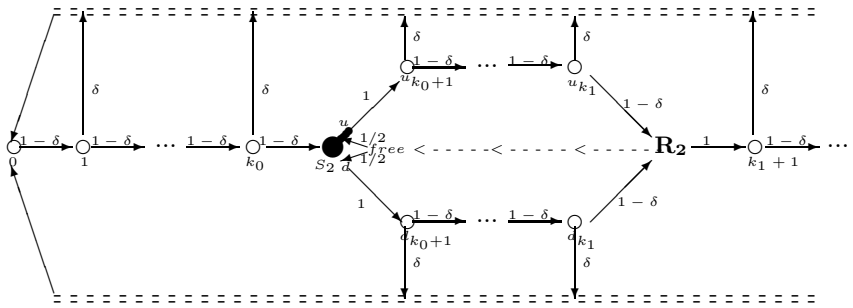
$$\mathbf{E}_{\rho_{u1} \times \rho_{d1}} D_k((x_1, \dots, x_{t_1}), (y_1, \dots, y_{t_1})) > 1 - \varepsilon,$$

where the samples x_i and y_i are generated by ρ_{u1} and ρ_{d1} correspondingly (the samples are generated independently; that is, the process are based on two independent copies of the Markov chain m). Let k_1 be such an index that the chain m starting from the state 0 with probability 1 does not reach the state $k_1 - 1$ by time t_1.

Construct the process ρ_2 as follows (see fig. 1). It is based on a chain m_2 on which Markov assumption is violated. The transition probabilities on states $0, \dots, k_0$ are the same as for the Markov chain m (from each state return to 0 with probability δ or go to the next state with probability $1 - \delta$).

There are two "special" states: the "switch" S_2 and the "reset" R_2. From the state k_0 the chain passes with probability $1 - \delta$ to the "switch" state S_2. The switch S_2 can itself have 3 values: u, d or $free$. If S_2 has the value u then from S_2 the chain passes to the state u_{k_0+1}, while if $S_2 = d$ the chain goes to d_{k_0+1}, with probability 1. In these cases S_2 does not change its value. If $S_2 = free$ then S_2 takes the value u or d with equal probabilities and passes either to u_{k_0+1} or to d_{k_0+1} accordingly. If the chain reaches the state R_2 then the value of S_2 is set to $free$. For now assume that the initial value of S_2 is $free$. In other words, the first transition from S_2 is random (either to u_{k_0+1} or to d_{k_0+1} with equal probabilities) and then this decision is remembered until the "reset" state R_2 is visited.

Figure 1: The processes m_2 and ρ_2. The states are depicted as circles, the arrows symbolize transition probabilities: from every state the process returns to 0 with probability δ or goes to the next state with probability $1-\delta$. The function f_2 is 1 on all states except $u_{k0+1}, \ldots, u_{k1}$ where it is 0; f_2 applied to the states output by m_2 defines ρ_2.

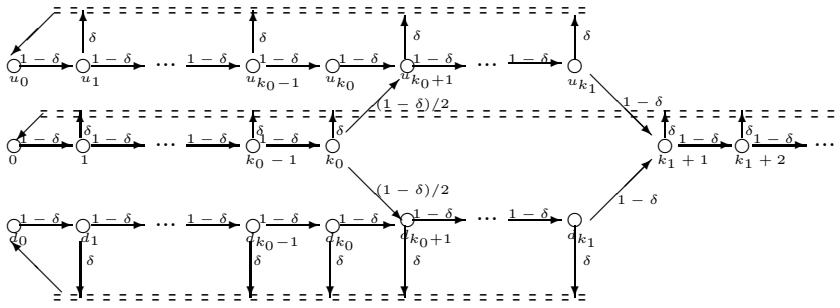


The rest of the transitions are as follows. From each state u_i, $k_0 \le i \le k_1$ the chain passes to the state 0 with probability δ and to the next state u_{i+1} with probability $1-\delta$. From the state u_{k_1} the process goes with probability δ to 0 and with probability $1-\delta$ to the "reset" state R_2. The same with states d_i: for $k_0 < i \le k_1$ the process returns to 0 with probability δ or goes to the next state d_{i+1} with probability $1-\delta$, where the next state for d_{k_1} is the "reset" state R_2. From R_2 the process goes with probability 1 to the state k_1+1 where from the chain continues ad infinitum: to the state 0 with probability δ or to the next state k_1+2 etc. with probability $1-\delta$.

The function f_2 is defined as follows: $f_2(i) = 1$ for $0 \le i \le k_0$ and $i > k_1$ (before the switch and after the reset); $f_2(u_i) = 0$ for all i, $k_0 < i \le k_1$ and $f_2(d_i) = 1$ for all i, $k_0 < i \le k_1$. The function f_2 is undefined on S_2 and R_2, therefore there is no output on these states (we also assume that passing through S_2 and R_2 does not increment time). As before, the process ρ_2 is defined as $\rho_2 = f_2(s_t)$ where s_t is the state of m_2 at time t, omitting the states S_2 and R_2. The resulting process s illustrated on fig. 1.

Step 1.2b. To define the initial distribution on the states of the process m_2, we first show that it is equivalent to a Markov chain. Indeed, construct the Markov chain m_2' as follows (see fig. 2). This chain has states $0, \ldots, k_0$, $k_1+1, \ldots$ and also $u_0, \ldots, u_{k_0}, u_{k_0+1}, \ldots, u_{k_1}$ and $d_0, \ldots, d_{k_0}, d_{k_0+1}, \ldots, d_{k_1}$. Transitions from the states 0 to $k_0 - 1$ are defined in the same way as for all the chains described before. From the state k_0 the chain passes with probability $(1-\delta)/2$ to the state u_{k_0+1} and with probability $(1-\delta)/2$ to the state d_{k_0+1}, while with probability δ it returns to 0; thus the state k_0 corresponds to the $free$ state of the switch S_2. From the states u_i, $i = 0, \ldots, k_1$ the chain passes with probability $1-\delta$ to the next state u_{i+1}, where the next state for u_{k_1} is $k+1$ and with probability δ returns to the state u_0 (and not to the state 0). Transitions for the state $d_0, \ldots, d_{k_1-1}$ are defined analogously. Thus the state

Figure 2: The process m'_2. The function f_2 is 1 everywhere except the states $d_{k_0+1}, \ldots, d_{k_1}$ where it is 0.



u_{k_0} corresponds to the state u of the switch S_2 and the state d_{k_0} — to the state d of the switch. Transitions for the states $k+1, k+2, \ldots$ are defined as before: with probability δ to 0 and with probability $1-\delta$ to the next state. Thus the state k_1+1 corresponds to the reset R_2. Clearly, the chain m'_2 as defined possesses a unique stationary distribution M_2 over the set of states and $M_2(i) > 0$ for every state i; take this distribution as its initial distribution. The resulting process m'_2 is stationary ergodic. The initial distribution M_2 can be used to define the initial distribution on the original (non-Markov) process m_2, with the initial value of the switch S_2 random: if m'_2 starts at a state u_i where $i \le k_0$ this corresponds to the process m'_2 starting at the state i with S_2 set to u, and analogously for d. If m'_2 starts at a state j where $j \le k_0$ or $j > k_1$ this corresponds to the process m_2 starting at the state j with S_2 set to $free$; the initial probabilities for S_2 and R_2 themselves are 0. Thus, we have defined (in two equivalent ways) the process ρ_2, which is stationary ergodic, and Bernoulli.

Step 1.k. As before, we can continue the construction of the processes ρ_{u3} and ρ_{d3}, that start with a segment of ρ_2. Let $t_2 > t_1$ be a time index such that

$$\mathbf{E}_{\rho_2 \times \rho_2} D_{t_2} < \varepsilon,$$

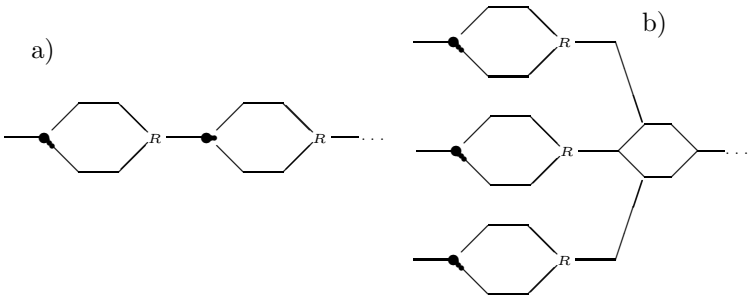
where both samples are generated by ρ_2. Let k_2 be such an index that when starting from the state 0 the process m_2 with probability 1 does not reach $k_2 - 1$ by time t_2 (equivalently: the process m'_2 does not reach $k_2 - 1$ when starting from either 0, u_0 or d_0). The processes ρ_{u3} and ρ_{d3} are based on the same process m_2 as ρ_2. The functions f_{u3} and f_{d3} coincide with f_2 on all states up to the state k_2 (including the states u_i and d_i, $k_0 < i \le k_1$). After k_2 the function f_{u3} outputs 0s while f_{d3} outputs 1s: $f_{u3}(i) = 0$, $f_{d3}(i) = 1$ for $i > k_2$.

Furthermore, we find a time $t_3 > t_2$ by which we have $\mathbf{E}_{\rho_{u3} \times \rho_{d3}} D_{t_3} > 1 - \varepsilon$, where the samples are generated by ρ_{u3} and ρ_{d3}, which is possible since D is consistent. Next, find an index k_3 such that the process m_2 does not reach $k_3 - 1$ with probability 1 if the processes ρ_{u3} and ρ_{d3} are used to produce two independent sequences. We then construct the process ρ_4 based on a (non-Markovian) process m_4 by "gluing" together ρ_{u3} and ρ_{d3} after the step k_3 with

a switch S_4 and a reset R_4 exactly as was done when constructing the process ρ_2. The process m_4 is illustrated on fig. 3 a).

The process m_4 can be shown to be equivalent to a Markov chain m'_4 (cf. fig. 3 b)), which consists of 3 copies of the process m_2 (which was shown to be equivalent to a Markov chain) truncated at step k_2 and linked as follows. From the step k_2 the process passes with probability $1/2$ to the state u_{k_2+1} with probability $(1-\delta)/2$ and to the state d_{k_2+1} with the same probability, while with probability δ it goes to the origin 0. From states u_i, $k_2 < i \leq k_3$ the process returns to u_0 with probability δ or goes to the next state $u_i + 1$ with probability $1-\delta$, and analogously for d_i. For the states u_{k_3} and d_{k_3} the next state is $k_3 + 1$, from which the process returns to the state 0 with probability δ or continues to the next state with probability $1-\delta$. The definition of f_4 is analogous to the previous definitions of f_i.

Figure 3: The processes m_4 and m'_4



Proceding this way we can construct the processes ρ_{2j}, ρ_{u2j+1} and ρ_{dj2+1}, $j \in \mathbb{N}$ choosing the time steps $t_j > t_{j-1}$ so that the test converges to 0 by t_j being run on two samples produced by ρ_j for even j, and converges to 1 by t_j being run on samples produced by ρ_{uj} and ρ_{dj} for odd j:

$$\mathbf{E}_{\rho_{2j}\times\rho_{2j}} D_{t_{2j}} < \varepsilon \tag{2}$$

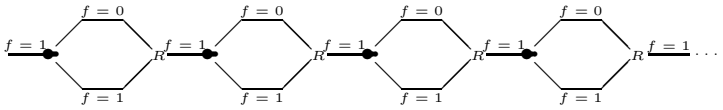
and

$$\mathbf{E}_{\rho_{2j+1}\times\rho_{2j+1}} D_{t_{2j+1}} > (1-\varepsilon). \tag{3}$$

For each j the number k_j is selected in a such a way that the state $k_j - 1$ is not reached (with probability 1) by the time t_j when starting from the state 0.

Step 2. Having defined k_j, $j \in \mathbb{N}$ we can define the process ρ, illustrated on fig. 4. The construction of the process is described in Step 2a up to an initial distribution on the states, while on Step 2b we show that the process is equivalent to a Markov chain, which we use to define the initial distribution for ρ. On Step 2c we show that the process ρ is a B-process. *Step 2a.* The construction is based on the process m_ρ that has states $0, \ldots, k_0$, $k_{2j+1}+1, \ldots, k_{2(j+1)}$, $u_{k_{2j}+1}, \ldots, u_{k_{2j+1}}$ and $d_{k_{2j}+1}, \ldots, d_{k_{2j+1}}$ for $j \in \mathbb{N}$, along with switch states S_{2j} and reset states R_{2j}. Each switch S_{2j} diverts the process to the state $u_{k_{2j}+1}$

Figure 4: The processes m_ρ and ρ. The states are on horizontal lines. The function f that defines the process ρ takes value 0 on the states on the upper lines and 1 on the rest of the states.

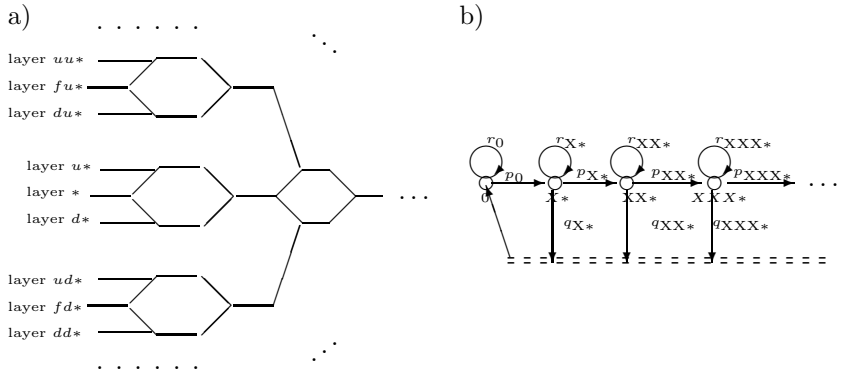


if the switch has value u and to $d_{k_{2j}+1}$ if it has the value d. If the switch has the value $free$ it assumes one of the values u or d with equal probabilities. The reset R_{2j} sets S_{2j} to $free$. From each state that is neither a reset nor a switch, the process goes to the next state with probability $1-\delta$ and returns to the state 0 with probability δ (cf. Step $1k$). The function f is defined as 1 everywhere except for the states u_j (for all $j \in \mathbb{N}$ for which u_j is defined) on which f takes the value 0. The process ρ is defined at time t as $f(s_t)$, where s_t is the state of m_ρ at time t.

Step 2b. As before, we can show that the process m_ρ is equivalent to a Markov chain m'_ρ. Indeed, the corresponding Markov chain can be constructed inductively, by replacing each switch-reset pair with two branches of a tree (see fig. 5a) as follows. To get rid of the switch S_2 and reset R_2 we proceed as in constructing the process ρ_2 on Step 1.2. That is, we introduce two copies $u_0, \dots, u_{k_0}$ and $d_0, \dots, d_{k_0}$ of the states $0, \dots, k_0$. Call each of these copies a layer, the upper one we call layer $u*$, the lower one $d*$, and the central one (states $0, \dots, k_0, k_1+1, \dots, k_2, \dots$) we call $*$. Change the transition probabilities of the states u_i and d_i, $i = k_0, \dots, k_1$ as in the Step 1.2, that is, the return state is not 0 but u_0 for u_i and d_0 for d_i. To get rid of the switch S_4 and reset R_4, introduce two more copies of all the preceding states (three layers each), one copy corresponding to the state u of the switch S_4 and the other to the state d of the switch S_4, while the central copy corresponds to the state $free$ (cf. the construction of the process ρ_4 and fig 3 b). Proceed analogously with the rest of the switches. We will have infinitely many layers in the process, each corresponding to some combination of the states of first n switches, for $n \in \mathbb{N}$. Let the layer $*$ correspond to all switches at the state $free$, the layer $u*$ to the first switch at u and the rest $free$, layer d to the first switch d and the rest $free$, layer $uu*$ to the first two switches u and the rest free, layer $fu*$ to the first switch $free$, second u and the rest free, and so on (cf. fig 5a); thus u, d, or f mean the the corresponding switch in the sequence has the value u, d, or $free$, and the symbol $*$ means that all the rest of the switches are set to $free$.

Let us show that the Markov chain m'_ρ has a stationary distribution over the states which assigns a positive probability to every state. To do this, consider a simpler Markov chain μ'_ρ, constructed by grouping the states of m'_ρ as follows: unite all the states $0, \dots, k_0$ of m_ρ into a state 0 of μ'_ρ. Unite all the states of m_ρ which are between k_0 and k_2, including the layers $u*, d*$, into the state $X*$ of μ'_ρ. That is, the state $X*$ corresponds to all the states that are past the first

Figure 5: a) The process m'_ρ, expanded. There are infinitely many layers above and below, corresponding to different combinations of switches. b) The process μ'_ρ: the states of the process m'_ρ are combined.



switch (S_2) but before the second (S_4). Unite all the states of m'_ρ which are between k_2 and k_4, including the layers layers X_1X_2* f m'_ρ, where X_1 and X_2 take any value in u, d, or f, into the state $XX*$ of μ'_ρ, and so on: the states between k_{2j} and $k_{2(j+1)}$, including the layers $X_1..X_j*$ (each X_i is either u, d, or f) are grouped into the state $X..X*$ (j symbols X) of μ'_ρ. In other words, the state $X..X*$ (j symbols X) corresponds to the process m'_ρ passing the switch S_{2j} but not reaching the next switch. From each of the resulting states i the chain procedes to the next state with probability p_i, remains in the state i with probability r_i, and returns to the state 0 with probability q_i. The values of p_i, q_i, and r_i are taken from the groups of states of the original Markov chain m_ρ, for example $p_0 = (1-\delta)^{k_0}$. Since $\delta \geq 1/2$, it is easy to see that for every state i we have $q_i \geq 2p_i$. Therefore, the Markov chain μ'_ρ possesses a stationary distribution with positive probabilities of all states, as can be seen, for example, by checking the conditions given in [14, p. 582]. From this it follows that the original Markov chain m'_ρ is recurrent and the state 0 has a positive probability, which implies (since all the other states are connected with the state 0) that all states have positive probability. Finally, the stationary distribution on the states of m'_ρ defines a stationary distribution on the states of m_ρ, and a distribution on the values of the switches. We take this distribution as the initial one. Clearly, the process ρ is stationary ergodic.

Step 2c. To show that ρ constructed is a B-processes, observe that it is obtained by applying the function f to the states of a chain m_ρ. Since this chain has infinitely many states, for each k we can find a state n_k such that the sum of probabilities (in the initial stationary distribution) of all of the states that follow the state n_k is not greater than 2^{-k}. Let g_k be a function that coincides with f on all states up to n_k and is 0 for all states that follow n_k. The process μ_k obtained by applying g_k to the states of m_ρ is equivalent (that is, the distributions are the same) to that which would be obtained by applying

g_k to a n_k+1-order Markov chain constructed by replacing all the states of m'_ρ greater than n_k by a single state, from which the chain passes to the state 0 with probability δ and with probability $1-\delta$ remains in this state. Therefore, each g_k is a function of an aperiodic stationary ergodic n_k-state Markov chain, and hence it is a B-process (see e.g. [5]). Moreover, ρ is a limit in $\bar{d}$ distance of the processes μ_k; indeed, if we couple ρ with μ_k in an obvious way by requiring that the underlying Markov chains always take state transitions together, then the probability of observing a different output is not greater than the probability that the chains are in one of the states that follow the state n_k, which is bounded by 2^{-k}. Since the set of B-processes is closed in the $\bar{d}$-distance, the process ρ is a B-process.

Step 3. Finally, we will show that the expected output of test D diverges if the test is run on two indpendent samples produced by ρ.

Let M_{2j} denote the initial state distribution of the process m_{2j}, $j \in \mathbb{N}$, and M that of the process m_ρ. Since each of the processes $m_2 j$, $j \in \mathbb{N}$ and the process m_ρ from each state returns to the state 0 with probability δ, the limiting (and hence initial) probability of this state is δ: $M_{2j}(0) = M(0) = \delta$. By construction, if the process m_ρ starts at the state 0 then up to time k_{2j} it behaves exactly as ρ_{2j} that has started at state 0. In symbols, we have

$$E_{\rho\times\rho}(D_{t_{2j}}|s_0^x = 0, s_0^y = 0) = E_{\rho_{2j}\times\rho_{2j}}(D_{t_{2j}}|s_0^x = 0, s_0^y = 0) \tag{4}$$

for $j \in \mathbb{N}$, where s_0^x and s_0^y denote the states of the processes generating the samples x and y correspondingly.

For each of the considered processes, the probability to start from the state 0 is high enough to ignore the behaviour of the tests if the processes start in other states. More formally, we use the following simple decomposition

$$\mathbf{E}(D_{t_j}) = \delta^2\mathbf{E}(D_{t_j}|s_0^x = 0, s_0^y = 0) + (1-\delta^2)\mathbf{E}(D_{t_j}|s_0^x \neq 0 \text{ or } s_0^y \neq 0), \tag{5}$$

(4), and (2) we have

$$\begin{aligned}\mathbf{E}_{\rho\times\rho}(D_{t_{2j}}) &\leq \delta^2\mathbf{E}_{\rho\times\rho}(D_{t_{2j}}|s_0^x = 0, s_0^y = 0) + (1-\delta^2)\\ &= \delta^2\mathbf{E}_{\rho_{2j}\times\rho_{2j}}(D_{t_{2j}}|s_0^x = 0, s_0^y = 0) + (1-\delta^2)\\ &\leq \mathbf{E}_{\rho_{2j}\times\rho_{2j}} + (1-\delta^2) < \varepsilon + (1-\delta^2).\end{aligned} \tag{6}$$

For odd indices, if the process ρ starts at the state 0 then (from the definition of t_{2j+1}) by the time t_{2j+1} it will not reach reset R_{2j}; therefore, the value of the switch S_{2j} either did not change up to the time t_{2j+1}, or has changed once from $free$ to either u or d. Since the definition of m_ρ is symmetric with respect to the values u and d of each switch, the probability that two samples $x_1, \ldots, x_{t_{2j+1}}$ and $y_1, \ldots, y_{t_{2j+1}}$ generated independently by (two runs of) the process ρ produced different values of the switch S_{2j} when passing through it for the first time is $1/2$. In other words, with probability $1/2$ two samples generated by ρ starting at the state 0 will look by the time t_{2j+1} as two samples generated by ρ_{u2j+1}

and ρ_{d2j+1} that has started at state 0. Thus

$$E_{\rho\times\rho}(D_{t_{2j+1}}|s_0^x=0,s_0^y=0) \geq \frac{1}{2}E_{\rho_{u2j+1}\times\rho_{d2j+1}}(D_{t_{2j+1}}|s_0^x=0,s_0^y=0) \quad (7)$$

for $j \in \mathbb{N}$. Using this, (5), and (3) we obtain

$$\begin{aligned}\mathbf{E}_{\rho\times\rho}(D_{t_{2j+1}}) &\geq \delta^2\mathbf{E}_{\rho\times\rho}(D_{t_{2j+1}}|s_0^x=0,s_0^y=0)\\ &\geq \frac{1}{2}\delta^2\mathbf{E}_{\rho_{2j+1}\times\rho_{2j+1}}(D_{t_{2j+1}}|s_0^x=0,s_0^y=0)\\ &\geq \frac{1}{2}\left(\mathbf{E}_{\rho_{2j+1}\times\rho_{2j+1}}(D_{t_{2j+1}})-(1-\delta^2)\right) > \frac{1}{2}(\delta^2-\varepsilon). \quad (8)\end{aligned}$$

Taking δ large and ε small (e.g. $\delta = 0.9$ and $\varepsilon = 0.1$), we can make the bound (6) close to 0 and the bound (8) close to $1/2$, and the expected output of the test will cross these values infinitely often. Therefore, we have shown that the expected output of the test D diverges on two independent runs of the process ρ, contradicting the consistency of D. □

References

[1] Adams, T.M., Nobel, A.B. (1998). On density estimation from ergodic processes, The Annals of Probability vol. 26, no. 2, pp. 794–804.

[2] Gyorfi L., Morvai G., Yakowitz S. (1998), Limits to consistent on-line forecasting for ergodic time series, IEEE Transactions on Information Theory vol. 44 , no. 2, pp. 886–892.

[3] Lehmann, E. L. (1986). Testing Statistical Hypotheses. Springer.

[4] Nobel, A.B. (2006), Hypothesis testing for families of ergodic processes. Bernoulli, vol. 12 no. 2, pp. 251–269.

[5] Ornstein, D. S. (1974). *Ergodic theory, randomness, and dynamical systems.* Yale Mathematical Monographs 5, Yale Univ. Press, New Haven, CT.

[6] Ornstein, D. S., Shields, P.(1994). *The $\bar{d}$-recognition of processes,* Advances in Mathematics, vol. 104, pp. 182–224.

[7] Ornstein, D. S. and Weiss, B.(1990). *How Sampling Reveals a Process.* Annals of Probability vol. 18 no. 3, pp. 905–930.

[8] Ryabko, B.(1988). *Prediction of random sequences and universal coding.* Problems of Information Transmission, vol. 24, pp. 87–96.

[9] Ryabko, B., Astola, J., Gammerman, A. (2006). Application of Kolmogorov complexity and universal codes to identity testing and nonparametric testing of serial independence for time series. Theoretical Computer Science, vol. 359, pp. 440–448.

[10] Ryabko, D., Ryabko, B. (2008). On Hypotheses Testing for Ergodic Processes. In Proceedings of IEEE Information Theory Workshop (ITW'08), Porto, Portugal, pp. 281–283.

[11] Shields, P.(1993). *Two divergence-rate counterexamples,* Journal of Theoretical Probability, vol. 6, pp. 521–545.

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

[13] Shields, P. (1996) *The Ergodic Theory of Discrete Sample Paths.* AMS Bookstore.

[14] Shiryaev, A. (1996). *Probability,* second edition. Springer.