The Annals of Probability
1990, Vol. 18, No. 3, 905–930

HOW SAMPLING REVEALS A PROCESS

BY DONALD S. ORNSTEIN[1] AND BENJAMIN WEISS

Stanford University and Hebrew University

A series of observations $\{\xi_1, \xi_2, \xi_3, \dots\}$ is presented to us and at each time n, when we have observed the first n of them, we are called upon to give our guess for what stochastic process produced the data. A universal scheme is given which, for any Bernoulli process (not necessarily independent), gives a sequence of processes that converges in a strong sense (the $\bar{d}$-metric) to the real process. In addition to this main result, many others are given which put it into proper perspective. In particular it is shown that in a certain sense the class of Bernoulli processes is the largest one for which such a universal scheme is possible.

1. Statements of results.

1.1. *Introduction.* A stochastic process $X = \{x_n : 0 \le n < \infty\}$ is determined by the joint distributions of the random variables $\{x_0, \dots, x_k\}$ for k. It is an easy consequence of Birkhoff's ergodic theorem that if a process X is ergodic and stationary, then from almost every sample point $\{\xi_n\}_{n=0}^{\infty}$ of the process one can determine these joint distributions. Indeed, in that case, for fixed k, with probability 1, the empirical distributions on k-tuples determined by $\{\xi_n\}_{n=0}^{\infty}$ will converge to the true distribution and with these finite distributions the original process X can be reconstructed exactly. In brief, with probability 1, a single sampling of an ergodic stationary process suffices to determine the process exactly. The above discussion holds equally well for continuous-time stochastic processes.

A more realistic situation is one in which as time goes on we are presented with more and more observations and we are asked to give some approximation to X based on a finite sampling $(\xi_0, \xi_1, \dots, \xi_n)$ (or $\{\xi_t\}_{0<t<n}$), which improves as n (or t) increases. The measure of distance between stochastic processes that we use is the $\bar{d}$-distance which gives a strong global sense of "closeness."

Our main result is an approximation scheme that works if X is a B-process. We also give a counterexample to show that no scheme can be found for general X. In addition, we prove a variant of the Shannon–McMillan theorem that holds for a single sample point.

1.2. *The $\bar{d}$-distance.* We first want a strong way to measure the distance between stochastic processes. We shall use the $\bar{d}$-distance which, roughly speaking, says that two stationary processes X and Y are close if they can be

Received July 1988; revised June 1989.
[1]Partially supported by National Science Foundation Grant DMS-86-05098.
AMS 1980 *subject classifications.* Primary 60G10; secondary 60F15, 28D05, 28D10.
Key words and phrases. Stationary process, ergodic theory, entropy, Shannon–McMillan theorem, prediction, Bernoulli shifts.

coupled in a stationary fashion so that their outputs x_n and y_n are close with high probability.

If, for example, a fixed stochastic process X is observed through a noisy medium, so that one obtains the process $\bar{X}$, where

$$(*) \qquad \bar{x}_n = x_n + e_n$$

with $\{e_n\}$ some process of errors, then X and $\bar{X}$ will be close in $\bar{d}$ (at least to the extent that the error process is 0 most of the time) since $(*)$ defines a coupling of X with $\bar{X}$ in which $\Pr\{x_0 \neq \bar{x}_0\} = \Pr\{e_0 \neq 0\}$.

More generally, supposing that X and Y are finite-valued, we say that

$$\bar{d}(X, Y) \leq \varepsilon$$

if there is a single stationary process Z of pairs $\{(\bar{x}_n, \bar{y}_n)\}_{n=0}^{\infty}$ such that

(i) $\{\bar{x}_n\}_0^{\infty} \sim \{x_n\}_0^{\infty}$, $\{\bar{y}_n\}_0^{\infty} \sim \{y_n\}_0^{\infty}$;
(ii) $\Pr\{\bar{x}_0 \neq \bar{y}_0\} \leq \varepsilon$.

Here $\sim$ denotes equality for processes, i.e., same joint distributions or same measure on realizations. Now the infimum of the ε's for which a coupling can be found which satisfies (i) and (ii) is taken to be the $\bar{d}$-distance between X and Y. Note that if X and Y are ergodic, then the coupling can be chosen to be ergodic as well. If X and Y take values in $\mathbb{R}$ or in a separable metric space, then (ii) must be modified to read $\Pr\{d(\bar{x}_0, \bar{y}_0) \geq \varepsilon\} \leq \varepsilon$.

For continuous-time processes there is an analogous definition. The $\bar{d}$-distance between stationary processes $(X, \mathbb{R}^+), (Y, \mathbb{R}^+)$ will be the infimum of those ε's such that a stationary coupling $Z = \{(\bar{x}_t, \bar{y}_t);\ t \in \mathbb{R}^+\}$ can be found satisfying

(i) $\{\bar{x}_t\colon t \in \mathbb{R}^+\} \sim \{x_t\colon t \in \mathbb{R}^+\}$, $\{\bar{y}_t\colon t \in \mathbb{R}^+\} \sim \{y_t\colon t \in \mathbb{R}^+\}$;
(ii) $\Pr\{\bar{x}_0 \neq \bar{y}_0\} \leq \varepsilon$. [$\Pr\{d(\bar{x}_0, \bar{y}_0) \geq \varepsilon\} \leq \varepsilon$ if X and Y take values in a separable metric space.]

Another way to define the $\bar{d}$-distance is in terms of generic realizations. The sequence $\{\xi_n\}$ is generic if for each k the empirical distributions of k-tuples in $\{\xi_n\colon 0 \leq n \leq N\}$ converges to the correct distribution. For continuous time, $\{\xi_t\}$ is generic if for each fixed $\alpha_1 < \alpha_2 < \cdots < \alpha_k$ the empirical distribution of $\{\xi_{t+\alpha_1}, \ldots, \xi_{t+\alpha_k}\}$ in $\{\xi_t\colon 0 \leq t \leq t\}$ converges to the right distribution. $\bar{d}(X, Y)$ is the infimum of the ε's such that there are generic realizations $\{\xi_t\}$ for X and $\{\eta_t\}$ for Y such that the set of t's for which $d(\xi_t, \eta_t) > \varepsilon$ has lower density equal to ε. For more details concerning this notion see [6]. Let us just note that to see that $\bar{d}$ satisfies the triangle inequality $\bar{d}(X, Y) + \bar{d}(Y, Z) \geq \bar{d}(X, Z)$ one takes couplings $(\bar{x}_t, \bar{y}_t)$, and $(\bar{\bar{y}}_t, \bar{\bar{z}}_t)$ and then couples *these* independently while identifying $\bar{y}_t$ and $\bar{\bar{y}}_t$.

1.3. *B-processes.* Recall that a process X is *totally ergodic* (TE) if for each d the process $\{x_{nd}\colon n \in \mathbb{N}\}$ is ergodic. This means that there is no finite

periodicity in the process. B-processes in discrete time are defined as follows:

A process is B if and only if it is the limit in $\bar{d}$ of TE *k-step Markov processes.*

Any process has a canonical k-step Markov approximation which gives the same probability to sequences of length less than or equal to k (i.e., we do not let the process look back more than k-steps). We find that

A TE *process is B if and only if it is the $\bar{d}$-limit of its canonical k-step Markov approximations.*

If a totally ergodic process is not B, then it cannot be approximated arbitrarily well by k-step Markov processes. One can state this in a more picturesque way by saying that the long-term behavior cannot be simulated by a machine with finite memory equipped with a random mechanism that simulates a roulette wheel.

A continuous-time process is B if and only if all of its discretizations at times $\{n\Delta\}_0^\infty$ give discrete time B-processes.

A more direct characterization (one that does not discretize time) can be given. Define a "semi-Markov process" to be a finite-state continuous-time process which stays at each state α for a *fixed* length of time t_α, depending on the state, and the jumps to one of the other states with a fixed probability distribution that depends only on the state. If the holding times t_α are irrationally related and it is possible to get from any state to any other, then the resulting process will be aperiodic (and mixing).

Continuous-time B-processes are exactly the $\bar{d}$-limits of such aperiodic "semi-Markov processes" [13].

The usual (and equivalent) definition of a B-process is given in terms of ergodic theory. From this point of view a stationary process is a transformation f or flow f_t on a measure space X and a function P on X. (X is the space of realizations of the process. f or f_t shifts each realization—representing the passage of time—and P evaluates each realization at time 0.) (f, X, P) or (f_t, X, P) is a B-process if and only if (f, X) or (f_t, X) is a Bernoulli shift or flow. This is discussed in [6]. For now we will note that certain dynamical systems such as the motion of a billiard ball on a square table with a convex obstacle or the geodesic flow on a manifold of negative curvature are isomorphic to the Bernoulli flow and our results will therefore apply to observing the time evolution of such systems. Furthermore, as is discussed in [12], it is not unreasonable to believe that most of the systems of physical interest that "look chaotic" are Bernoulli.

1.4. *Existence of a guessing scheme.* Our main result says that for the class of finite-valued B-processes, one can give a method for constructing approximations X_n, based on $(\xi_0, \xi_1, \ldots, \xi_n)$, which will converge in $\bar{d}$ to X. More formally, we suppose that the values the process can take on is a subset

of a separable metric space T. We think of T as the reals or as a finite set. By a *process reconstruction scheme* or more colloquially, a *guessing scheme* S, we mean a family of functions $S^{(n)}$: $T^{n+1} \to$ {stationary stochastic processes}, $n = 1, 2, 3, \ldots$. Thus $S^{(n)}(\xi_0, \xi_1, \ldots, \xi_n)$ is a process and we have the following theorem.

THEOREM 1. *There is a guessing scheme S such that if X is any B-process, then with probability* 1,

$$(*) \qquad \lim_{n\to\infty} \bar{d}\big(S^{(n)}(\xi_0, \xi_1, \ldots, \xi_n), X\big) = 0.$$

It also turns out that the scheme S that we will construct is *robust* in the following sense: If instead of applying it to the actual sample $\{\xi_n\}_{n=0}^{\infty}$, it is applied to a sequence $\{\tilde{\xi}_n\}_{n=0}^{\infty}$ such that the upper density of the set of n's for which $d(\tilde{\xi}_n, \xi_n) > \delta$ is less than or equal to δ, then

$$\limsup \bar{d}\big(s^{(n)}(\tilde{\xi}_0, \tilde{\xi}_1, \ldots, \tilde{\xi}_n), X\big) \le \delta.$$

For continuous-time processes $\{x_t\}$ a guessing scheme S will be a family of mappings $\{S^{(n)}\}_{n=1}^{\infty}$,

$$S^{(n)}\text{: } T^{[0,n]} \to \{\text{stationary processes with continuous-time parameter}\},$$

where $T^{[0,n]}$ denotes finite-valued measurable maps from $[0, n]$ to T. This constitutes the possible outcomes of observing a continuous process for a period of length n. We have the following theorem.

THEOREM 2. *There is a guessing scheme S such that if X is any continuous parameter B-process, then with probability* 1,

$$\lim_{n\to\infty} \bar{d}\big(S^{(n)}(\xi_t\text{: } 0 \le t \le n), X\big) = 0.$$

The robustness of the scheme S described above is valid for Theorem 2 as well.

1.5. *Description of the guessing scheme.* Since the scheme S is easy to describe we will do so here. Fix some sequence $L(k)$ such that

$$\lim_{k\to\infty} L(k)\Big/A^k = \infty$$

for any positive A. For $L(k) \le n < L(k+1)$, $S^{(n)}(\xi_0, \xi_1, \ldots, \xi_n)$ is defined in two steps:

Step 1: Calculate the empirical distribution of k-blocks in the finite string $(\xi_0, \xi_1, \ldots, \xi_n)$, call it π_k.

Step 2: Define $S^{(n)}(\xi_0, \xi_1, \ldots, \xi_n)$ to be the process obtained by concatenating k-blocks independently with distribution π_k.

This procedure is not very random, since there is a periodicity of size k. To obtain a very random process, we randomize as follows:

Step 2′: Define $S^{(n)}(\xi_0, \xi_1, \ldots, \xi_n)$ to be the process obtained by concatenating k-blocks independently with distribution π_k. Now leave a space between the consecutive k-blocks with some probability p, $0 < p < 1$, and do this in an independent way.

This construction will yield a B-process.

For continuous-time processes we use the same function $L(k)$ and the scheme is defined in two steps similar to what was described above. The first step is the same, for $L(k) \leq t < L(k+1)$.

Step 1 (Continuous time): Calculate the empirical distribution of $[0, k]$-blocks in $(\xi_s\colon 0 \leq s \leq t)$ that begin at an integer value of t and call it π_k.

Step 2: Define $S^{(t)}((\xi_s\colon 0 \leq s \leq t))$ to be the process obtained by concatenating independently k-blocks with the distribution π_k.

As above we need to modify step 2 to get a B-process.

Step 2′: Define $S^{(t)}((\xi_s\colon 0 \leq s \leq t))$ to be the process obtained by concatenating independently k-blocks with the distribution π_k. Now pick an irrational number α and probability p, $0 < p < 1$, and leave a space of length α between consecutive k-blocks independently and with probability p.

This scheme is robust in the sense that a (small) change in π_k leads to a corresponding (small) change in $S^{(t)}$. (The change should be measured in the $\bar{d}$-sense. See the discussion before Proposition 5, Section 2.)

1.6. *A sharp form of the Shannon–McMillan theorem.* Our scheme is based on a strong form of the Shannon–McMillan theorem (Theorem 2.2 below). We could base our scheme on a special case of Theorem 2.2, the case of independent random variables, by using the characterization of B-processes given in the last paragraph of Section 1.3. We believe, however, that Theorem 2.2 is of independent interest.

THEOREM. *Let $\{x_n\}_1^\infty$ be a stationary, finite-valued ergodic process of entropy h. Then for a.e. realization $\{\xi_n\}_1^\infty$ of $\{x_n\}_1^\infty$, and all N large enough and $M > 2^{hN}$ we have: $\{\xi_n\}_1^M$ puts a measure on N-blocks (via their frequency) and after ignoring a collection of N-blocks of measure less than ε, the remaining N-blocks have measure between $2^{-(h+\varepsilon)N}$ and $2^{-(h-\varepsilon)N}$.*

This statement for $\{\xi_i\}_1^\infty$ instead of $\{\xi_i\}_1^M$ would, because of the ergodic theorem, be the usual Shannon–McMillan theorem. Our theorem says that if M is big enough to fit in the right number of N-blocks, then it works.

Paul Shields pointed out to us that as soon as we have any scheme to guess the entropy, then we can pick, for each n, the k such that the entropy of the empirical distribution of k blocks in $(\xi_0, \ldots, \xi_n)$ is closest to the guessed entropy and proceeds as in Section 1.4. However, this only works in discrete time and for processes with values in a finite alphabet. For other such schemes see [1] and [16].

It is worth pointing out that other features of a process may be guessed by universal schemes. For example, in [7] it is shown that as more and more of the past of a process is observed, $(\xi_{-n}, \ldots, \xi_{-2}, \xi_{-1})$ we can make a series of guesses as to the distribution of ξ_0 given $(\xi_{-n}, \ldots, \xi_{-2}, \xi_{-1})$ which will converge with probability 1 to the correct conditional distribution. Reversing time, this gives a universal scheme for reconstructing the past of a zero-entropy process, from current observations.

1.7. *The non-B case.* In general, one cannot hope to guess the long-term behavior from finite information. Consider, for example, the circle $\{z \in \mathbb{C}: |z| = 1\} = K$ and a fixed partition $\mathscr{P}$ into two sets determined by $\operatorname{Im} z \geq 0$, $\operatorname{Im} z < 0$. For each $\alpha \in [0, 1]$ one gets a two-valued process $(\mathscr{P}, R_\alpha)$, from this partition with the mapping $z \mapsto e^{2\pi i\alpha} z$, the rotation by α. For any N and ε, there is some $\delta > 0$ such that if $|\alpha - \beta| < \delta$, then the N-block distribution of $(\mathscr{P}, R_\alpha)$ is within ε of the N-block distribution of $(\mathscr{P}, R_\beta)$. Nonetheless the processes for distinct irrational α, β remain a fixed distance apart in $\bar{d}$. This latter point may be verified by taking any two points $z, w \in K$ and observing their orbit under $R_\alpha \times R_\beta$ in $K \times K$. There are similar examples in positive entropy cases and even for K-processes (those processes that satisfy 0–1 laws, i.e., have a trivial remote future). Indeed, in [8] one finds for any N, an uncountable family of processes that have the same distribution on N-blocks but are a distance at least $1/4$ apart in $\bar{d}$.

1.8. *Comparing two realizations.* One consequence of the existence of our guessing scheme is that given two processes X and Y, at least one of which is Bernoulli, then we have a method for computing the $\bar{d}$-distance between them.

On the other hand, if neither X nor Y is Bernoulli, then one can still assert that our guesses will eventually be greater than or equal to $\bar{d}(X, Y)$.

For non-Bernoulli processes, not only can we not give a universal scheme for reconstructing the process in a sequential fashion—but we cannot even estimate the distance between processes. We shall prove that there is no scheme S which, as a function of

$$(\xi_0, \xi_1, \ldots, \xi_n; \eta_0, \eta_1, \ldots, \eta_n)$$

will converge with probability 1 to the $\bar{d}$-distance between X and Y. In fact even the following weaker requirement cannot be satisfied for processes $X^{(i)}, Y^{(i)}$ with

$$\lim_{i \to \infty} \bar{d}(X^{(i)}, Y^{(i)}) = 0,$$

$$\limsup_{i \to \infty} \limsup_{n \to \infty} S\left(\left(\xi_0^{(i)}, \xi_1^{(i)}, \ldots, \xi_n^{(i)}, \eta_0^{(i)}, \eta_1^{(i)}, \ldots, \eta_n^{(i)}\right)\right) \leq \tfrac{1}{4}$$

with probability 1, while if

$$\lim_{i\to\infty} \bar{d}(X^{(i)}, Y^{(i)}) = 1,$$

then

$$\liminf_{i\to\infty} \liminf_{n\to\infty} S\big(\big(\xi_0^{(i)}, \xi_1^{(i)}, \ldots, \xi_n^{(i)}, \eta_0^{(i)}, \eta_1^{(i)}, \ldots, \eta_n^{(i)}\big)\big) \geq \tfrac{3}{4}.$$

1.9. *Further results.* (a) The Bernoulli theory tells us that if two B-processes have the same entropy and are $\bar{d}$-close, then the $\bar{d}$-joining can be taken to be an isomorphism. This is called α-congruence and is discussed fully in [12]. Let us give an example of this type of result. Say we are observing a Bernoulli flow on a compact manifold, M, then the $\bar{d}$-joining between the flow and its reconstruction can be given a more specific form. We define a viewer as a random device, independent of the flow, with the property that if we look at a point x in M through the viewer, then we see another point, $\bar{x}$, in M and $\bar{x}$ depends only on x and the state of the viewer. We say that the viewer is ε-reliable if $\Pr\{|x - \bar{x}| > \varepsilon\} < \varepsilon$. Our strengthened result says that, given ε, if we observe long enough, then the reconstructed process could be simulated exactly (as a stochastic process) by observing the original flow through an ε-reliable viewer. (There need be no redundancy in this description in the sense that the orbit we are looking at and the states of the viewer can be reconstructed from the path that we see.)

(b) In Section 5, where we discuss in detail the stability of our reconstruction scheme, it will also be shown that if the reconstruction scheme is applied to a non-B-process, then it will oscillate in $\bar{d}$ by an amount that depends upon the distance of the non-B-process to the class of B-processes.

(c) In the concluding section we shall discuss briefly some aspects of sampling a nonstationary process. In this case the ergodic theorem does not apply so that even observing an entire realization $\{\xi_n\}_1^\infty$ may not teach us a great deal about the process. This was shown, in a very beautiful example, by Blackwell [2]. We discuss some positive results connected with the example as well as some other problems and directions for further research.

2. Finite-valued discrete time processes. We begin with a few lemmas that have the flavor of an individual Shannon–McMillan theorem.

LEMMA 1. *For any finite-valued ergodic process $\{x_n\}_1^\infty$, with entropy h, given $\delta > 0$ there is some N_0, such that for all $N \geq N_0$ there is a collection D_N of N-blocks such that*

(i) $|D_N| \leq 2^{(h+\delta)N}$;

(ii) *for a.e. realization $\{\xi_n\}_1^\infty$, for n sufficiently large and $n \geq N \geq N_0$ a $(1 - \delta)$-fraction of the N-blocks of $\{\xi_i\}_1^n$ belong to D_N.*

REMARKS. 1. For any fixed N, the result with n depending on N follows immediately from the standard Shannon–McMillan theorem and the individ-

ual ergodic theorem. The more elaborate argument is required to get that for n large enough (ii) will hold for *all* N in the range (N_0, n).

2. By "a $(1-\delta)$-fraction of the N-blocks in $(\xi_1, \xi_2, \ldots, \xi_n)$" we usually mean a $(1-\delta)$-fraction of *all* of the N-blocks $\xi_i, \xi_{i+1}, \ldots, \xi_{i+N-1}$, $1 \le i \le n-N+1$. However, the proof is valid even if one considers only the N-blocks that end at ξ_{kN}'s, $(\xi_1, \ldots, \xi_N), (\xi_{N+1}, \ldots, \xi_{2N}), \ldots$, or any of the other $(N-1)$ fixed ways of dividing $(1, \ldots, n)$ into disjoint N-blocks.

PROOF OF LEMMA 1. 1. Apply the Shannon–McMillan theorem to find some K sufficiently large so that there is a collection of K-blocks, $\mathscr{C} \subset \{1, 2, \ldots, A\}^K$ (A = number of values that $\{x_n\}$ takes on) satisfying

(i) $|\mathscr{C}| \le 2^{(h+\delta/2)K}$;
(ii) $\Pr(\bigcup \mathscr{C}) \ge 1 - \eta^2$,

where η will be determined below.

2. For any $N \ge N_0$ define D_N to be those N-blocks such that a $(1-3\eta)$-fraction of their K-blocks belong to $\mathscr{C}$.

To count the number of elements in D_N, we first remark that by considering the K different ways of dividing an N-block into disjoint K-blocks we have that for any $w \in D_N$ we can find *disjoint* K-blocks from $\mathscr{C}$ that cover $(1-3\eta)N$ of w. The number of such patterns is at most $\binom{N}{3\eta N}$. The K-blocks can be filled in at most $|\mathscr{C}|^{N/K}$ different ways, while for the gaps we have at most $A^{3\eta N}$ different possibilities, all together

$$|D_N| \le \binom{N}{3\eta N} A^{3\eta N} \left(2^{(h+\delta/2)K}\right)^{N/K} \le 2^{(h+\delta)N}$$

for a suitable choice of η and N_0.

3. Now for a.e. $\{\xi_i\}_1^\infty$, for all n sufficiently large, the individual ergodic theorem ensures us that

$$\frac{\left|\{1 \le j \le n-K+1 : \xi_j \xi_{j+1} \cdots \xi_{j+K-1} \in \mathscr{C}\}\right|}{n} \ge 1 - 2\eta^2.$$

Finally, the L_1-Tchebycheff or Markov inequality implies that a $(1-2\eta)$-fraction of the N-blocks of $\{\xi_i\}_1^n$ are $1-3\eta$ covered by $\mathscr{C}$ and thus belong to D_N. Since we could have taken $\eta < \delta/2$ this together with (2) establishes the lemma. □

THEOREM 2. *Let $\{x_n\}_1^\infty$ be a stationary finite-valued ergodic process whose entropy is h. Then for a.e. realization $\{\xi_i\}_1^\infty$ we have: given $\varepsilon > 0$ if N is sufficiently large and $M > 2^{hN}$, then we cannot cover a ε-fraction of the N-blocks in $\{\xi_i\}_i^M$ by fewer than $2^{(h-\varepsilon)N}$ distinct kinds of N-blocks and we can cover a $(1-\varepsilon)$-fraction of the N-blocks in $\{\xi_i\}_1^M$ by fewer than $2^{(h+\varepsilon)N}$ distinct kinds of N-blocks.*

PROOF. 1. Fix, as before, A equal to the number of different values the process may assume. Fix some $\varepsilon > 0$, and an auxiliary $\delta > 0$ which will be specified later. Apply Lemma 1 with this δ to find an N_0 and suppose that N is larger than N_0. Now if n is sufficiently large, Lemma 1 gives us half of the conclusion of the theorem. We shall analyze the way in which the second half can fail—i.e., those realizations $\{\xi_i\}_1^n$ which can be covered by fewer than $2^{(h-\varepsilon)N}$ different N-blocks. We shall simply count the number of different such realizations and get a number decidedly smaller than 2^{hn}. Since the typical probability of the events of the form $(x_i = \xi_i: 1 \le i \le n)$ is 2^{-hn} these anomalous n-realizations occupy a very small piece of the total probability space. Then an application of the Borel–Cantelli lemma will conclude the argument. Here are the details.

2. Recall D_N—the collection of N-blocks that numbered at most $2^{(h+\delta)N}$—from Lemma 1. We fix one subset C of D_N with fewer than $2^{(h-\varepsilon)N}$ elements. The number of *different* such subsets is at most

$$(*) \qquad |D_N|^{2^{(h-\varepsilon)N}} \le 2^{(h+\delta)N \cdot 2^{(h-\varepsilon)N}}$$

Next notice that if an ε-fraction of the N-blocks in $(\xi_1, \xi_2, \ldots, \xi_n)$ belongs to C, then the same is true of one of the N different ways of dividing $(1, 2, \ldots, n)$ into disjoint consecutive N-blocks. Thus the number of different ways of fixing places where disjoint N-blocks belong to C is at most $N2^{n/N}$.

3. The part of the n-blocks not covered by these special N-blocks from C divides into two parts. On one of them Lemma 1 implies that the N-blocks are chosen from D_N while what is left fills at most a δ-fraction of $(1, 2, \ldots, n)$. The number of ways of dividing these two parts is at most

$$(**) \qquad \sum_{j=0}^{\delta} \binom{n/N}{\delta n/N} \le 2^{[-\delta \log \delta - (1-\delta)\log(1-\delta)]n/N} \le 2^{-2[\delta \log \delta]n/N},$$

while the number of ways of filling in a fixed pattern is bounded by

$$(A^{\delta n})\left(\{2^{(h+\delta)N}\}^{(1-\varepsilon)n/N}\right)\left(\{2^{(h-\varepsilon)N}\}^{\varepsilon n/N}\right)$$

$$\le 2^{n(h-\varepsilon^2+\delta(\log A+(1-\varepsilon)))}.$$

Now notice that by the choice of $n \ge 2^{hN}$, $(*)$ is bounded by

$$2^{n(h+\delta)N2^{-\varepsilon N}}$$

and thus all together by choosing δ small enough and N large enough we can assert that the number of different anomalous n-blocks is at most

$$2^{n(h-\varepsilon^2/2)}.$$

4. For each n, consider now the collection of n-blocks $(\xi_1, \ldots, \xi_n)$ whose probability is less than

$$2^{-(h-\varepsilon^2/4)n}$$

and call these *ordinary* blocks. The Shannon–McMillan–Breiman theorem says that for a.e. $(\xi_i)_1^\infty$ for n large enough the block corresponding to $(\xi_1, \ldots, \xi_n)$

is ordinary. The result of (3) gives now that the probability of the event that the n-block is ordinary and constitutes a failure for the assertion of the theorem for $\varepsilon > 0$, is at most $2^{-(\varepsilon^2/4)n}$. By the Borel–Cantelli lemma a.e. $(\xi_i)_1^\infty$ belongs to this event only finitely many times and thus with probability 1 for n large enough the theorem holds for this ε. Repeat the argument for a sequence of ε's that tend to 0, and collect together the exceptional null sets to complete the proof of the theorem. □

When we begin to observe a process we do not know the value of its entropy. However, the qualitative nature of the above theorem allows us to do the following. Fix some sequence of integers $L(k)$ that grows faster than any exponential,

$$\lim_{k\to\infty} L(k)^{-1}C^k = 0 \quad \text{for all } C > 0.$$

When n is in the range $L(N) \le n < L(N+1)$ try to estimate the entropy of the process by finding a covering of at least half of the sequence of observations $(\xi_1, \ldots, \xi_n)$ by as few N-blocks as possible, and calculate

$$\frac{1}{N} \times \log(\text{minimum number of } N\text{-blocks needed to cover one half of } (\xi_1, \ldots, \xi_n)).$$

This will converge with probability 1 to the entropy of the process.

Note that the restriction to ergodic processes was only made to give an easier formulation. In fact even if a process is not ergodic, with probability 1 its realizations are typical for one of the ergodic components of the process and thus the theorem is true as stated with the understanding that h is now the entropy of the ergodic component for which $(\xi_i)_1^\infty$ is typical.

LEMMA 3. *If μ is a distribution on N-blocks (in a finite alphabet with A elements) that satisfies*:

(i) *for a collection $\mathscr{C}$ of N-blocks with $|\mathscr{C}| \le 2^{(h+\delta)N}$, $\mu(\bigcup \mathscr{C}) \ge 1 - \delta$;*

(ii) *for any collection $\overline{\mathscr{C}}$ with $|\overline{\mathscr{C}}| \le 2^{(h-\delta)N}$ we have $\mu(\bigcup \overline{\mathscr{C}}) < \delta$, then if $\{\mu_j = \mu(\mathscr{C}_j)\}$, where C_j are the distinct N-blocks,*

$$\left|\left(-\sum \mu_j \log \mu_j\right) - Nh\right| \le (3\delta + 2\delta \log A)N + 2\delta \log \frac{1}{\delta} + 2\delta hN.$$

In particular, the process obtained by concatenating N-blocks independently with distribution μ has an entropy that tends to h as $\delta \to 0$ and $N \to \infty$.

PROOF. 1. Let $\hat{\mathscr{C}}$ denote the C_j's whose measure satisfies

$$\mu(C_j) \ge 2^{-(h-\delta)N};$$

there are no more than $2^{(h-\delta)N}$ such atoms so that by (ii)

$$\mu\left(\bigcup \hat{\mathscr{C}}\right) < \delta.$$

Next let $\tilde{\mathscr{C}}$ denote those C_j's in $\mathscr{C}$ that satisfy

$$\mu(C_j) < 2^{-(h+2\delta)N};$$

then the total measure of $\tilde{\mathscr{C}}$ satisfies by (i)

$$\mu\Big(\bigcup \tilde{\mathscr{C}}\Big) < 2^{-\delta N} < \delta$$

if N is large enough, as we may assume. Thus the atoms $\mathscr{C}_j$ with measure μ_j in the range $(2^{-(h+2\delta)N}, 2^{-(h-\delta)N})$ fill up at least $(1 - 2\delta)$ of the distribution. This gives the main term for $\Sigma \mu_j \log \mu_j$. On the part with measure at most 2δ, using the fact that there are at most A^N different $\mathscr{C}_j$'s and maximizing the possible contribution, we get the remainder of the term.

2. The second statement follows immediately from the explicit estimate for $-\Sigma \mu_j \log \mu_j$ since for independent concatenations this, suitably normalized, gives the entropy of the process. □

A finite-valued process $\{x_n\}_0^\infty$ is said to be finitely determined (FD) if given $\varepsilon > 0$ there are a δ and an N, such that if $\{y_n\}_0^\infty$ is any process satisfying

$$(1) \qquad \big|h\big(\{y_n\}_0^\infty\big) - h\big(\{x_n\}_0^\infty\big)\big| < \delta,$$

$$(2) \qquad d\big(\mathrm{dist}(y_n)_1^N, \mathrm{dist}(x_n)_1^N\big) < \delta,$$

then $\bar{d}(\{y_n\}_0^\infty, \{x_n\}_0^\infty) < \varepsilon$.

To be completely explicit, (2) means the following: Denote by Π_x^N, Π_y^N the distribution of the $\{x_n\}, \{y_n\}$-processes on N-strings. That is to say,

$$\Pi_X^N(a_1, \ldots, a_N) = \Pr(x_1 = a_1, x_2 = a_2, \ldots, x_N = a_N),$$

$$\Pi_Y^N(a_1, \ldots, a_N) = \Pr(y_1 = a_1, y_2 = a_2, \ldots, y_N = a_N).$$

Then the distance on distributions in (2) is

$$d\big(\Pi_X^N, \Pi_Y^N\big) = \sum_{(a_1, \ldots, a_n)} \big|\Pi_X^N(a_1, \ldots, a_N) - \Pi_Y^N(1_1, \ldots, a_N)\big|.$$

The $\bar{d}$-metric on processes was explained in the introduction to Section 1.3. The result we need from the isomorphism theory of Bernoulli shifts is [6]:

THEOREM. *A process is B if and only if it is finitely determined.*

THEOREM 4. *Let $L(k)$ be any function such that $\lim_{k\to\infty} L(k)/A^k = \infty$ for any positive A. Let the scheme S be defined by the following algorithm: $S(\xi_0, \ldots, \xi_n)$ is for $L(N) \le n < L(N+1)$, the independent concatenation of N-blocks with the distribution on N-blocks, π, being given by their frequency in $(\xi_0, \ldots, \xi_n)$. Then if $\{x_n\}_0^\infty$ is a B-process we have*

$$(*) \qquad \lim_{n\to\infty} \bar{d}\big(S(\xi_0, \ldots, \xi_n), \{x_n\}_0^\infty\big) = 0$$

for a.e. realization $\{\xi_n\}_0^\infty$ of the process.

REMARK. If we leave a space between the N-blocks we are concatenating, and do this independently and with probability p, $0 < p < 1$, then $(*)$ will still hold and $S(\xi_0, \ldots, \xi_n)$ will be a B-process. This follows easily from the isomorphism theory (one can very easily check "very weak Bernoulli").

PROOF. Because $\{x_n\}_0^\infty$ is finitely determined we need only show that the entropy of $S(\xi_0, \ldots, \xi_n)$ converges to the entropy of $\{x_n\}_0^\infty$ (this follows from Theorem 2 and Lemma 3) and the distribution of K-blocks in $S(\xi_0, \ldots, \xi_n)$ converges to the distribution of K-blocks in $\{x_n\}_0^\infty$ for each fixed K. To see this, note that the distribution of K-blocks in the N-blocks with the distribution π is the same (except for end effects) as the distribution of K-blocks in $\xi_0, \ldots, \xi_n$. The ergodic theorem implies that the latter converges to the distribution of K-blocks in $\{x_n\}_0^\infty$. □

REMARKS. The $\bar{d}$-limit of B-processes is a B-process (see [6]) and therefore the converse of the theorem is true as well. Namely, if for even one realization we have $(*)$ holding $\{x_n\}_0^\infty$ is a B-process. This is true even if we do not leave a random space as in the remark above. This follows because the independent concatenation is basically a B-process except for the possible occurrence of a finite rotation factor. The paper [14] explains how a limit of block independent processes is Bernoulli. The main point is that because the block size is getting bigger and converging in $\bar{d}$ to a fixed process the rotation factor possibility is ruled out in the limit.

It will be useful to define the *$\bar{d}$-distance for finite sequences of random variables*. The definition is as follows: $\bar{d}(\{x_i\}_1^n, \{y_1\}_1^n) = \inf\{c$: there is a joint distribution on $(x_1, y_1), \ldots, (x_n, y_n)$ with $(1/n)\sum_{i=1}^n \Pr\{x \neq y\} < c\}$ or $(1/n)\sum_{i=1}^n \Pr\{d(x, y) > c\} < c$. The connection with the $\bar{d}$-distance that we have defined on processes is as follows:

$$\bar{d}\left(\{x_i\}_1^\infty, \{y_i\}_1^\infty\right) = \lim_{n_j \to \infty} \bar{d}\left(\{x_i\}_1^{n_j}, \{y_i\}_1^{n_j}\right)$$

for any sequence of n_j's going to ∞. The nontrivial direction of this statement is to show how good (i.e., small c) joinings of the finite distributions lead to good joinings of the processes. This can be done as follows. First, get some weak* limit of the finite joinings to give a joining of the entire process. The resulting process will not be stationary in general; however, its marginals are. Now shift the distribution and averages over longer and longer and longer intervals and go to a weak* limit of some subsequence. This will produce a stationary joining and the distribution of (x_1, y_1) will now reflect the average joint distribution so that if for all n_j we had $\bar{d}(\{x_i\}_1^{n_j}, \{y_i\}_1^{n_j}) < c$ we would get $\Pr(x \neq y) < c$ as required.

PROPOSITION 5. *If $\{x_n\}_0^\infty$ is Bernoulli, then $(*)$ for a.e. realization $\{\xi_n\}_1^\infty$ of $\{x_n\}_1^\infty$ the empirical distributions of the N-blocks in n-blocks with $L(N) \leq n < L(N+1)$ tends in $\bar{d}$ to the true distribution of the N-blocks.*

PROOF. If n is large, then $S(\xi_1, \ldots, \xi_n) = S = \{s_i\}_1^\infty$ will be $\bar{d}$ close to $X = \{x_i\}_1^\infty$. Join S and X so as to realize their $\bar{d}$-distance. Use this joining to pick matching realizations, $\{\alpha_i\}_1^\infty$ of $\{s_i\}_1^\infty$ and $\{\beta_i\}_1^\infty$ of $\{x_i\}_1^\infty$, and use the frequencies of $(\{\alpha_i\}_{jN+1}^{(j+1)N}, \{\beta_k\}_{jN+1}^{(j+1)N})$ to match the "empirical N-block distribution" in S and the distribution of N-blocks in $\{x_n\}$ starting at multiples of N. This last distribution is the true distribution of N-blocks in X because $\{x_n\}$ is totally ergodic (being a B-process). □

This last proposition enables us to give a scheme which will accept the outputs of two processes, one of which is Bernoulli, and to give a sequence of numbers which will converge almost surely to the $\bar{d}$-distance between the processes.

THEOREM 6. *There is a sequence of real-valued functions defined on pairs of outputs of finite-valued processes $\bar{S}^{(n)}$ such that for any two processes $\{x_n\}_1^\infty, \{y_n\}_1^\infty$ where $\{y_n\}$ is B and a.e. realizations $\{\xi_n\}_1^\infty, \{\eta_n\}_1^\infty$ we have that*

$$\lim \bar{S}^{(n)}(\xi_1, \ldots, \xi_n, \eta_1, \ldots, \eta_n) = \bar{d}\big(\{x_n\}_1^\infty, \{y_n\}_1^\infty\big).$$

One defines $\bar{S}^{(n)}$ as follows: For $L(N) \le n < L(N+1)$ one calculates Π_ξ^N, Π_η^N to be the empirical distribution of N-blocks in $(\xi_1, \ldots, \xi_n), (\eta_1, \ldots, \eta_n)$, respectively. Then

$$\bar{S}^{(n)}(\xi_1, \ldots, \xi_n, \eta_1, \ldots, \eta_n) = \bar{d}\big(\Pi_\xi^N, \Pi_\eta^N\big).$$

PROOF. By the definition of $\bar{d}$ we know that for a.e. $\{\xi_n\}_1^\infty$ there is some generic realization of $\{y_n\}_1^\infty, \{\overline{\eta}_n\}_1^\infty$ such that the density of the set of n where $\xi_n \neq \overline{\eta}_n$ is less than or equal to $\bar{d}(\{x_n\}_1^\infty, \{y_n\}_1^\infty) = \beta$. This implies $\bar{d}(\Pi_\xi^N, \Pi_{\overline{\eta}}^N) \to \beta$.

We can also assume by Proposition 5 that $\Pi_{\overline{\eta}}^N$ tends to the distribution of N-blocks in $\{y_n\}_1^\infty$, and that the above also holds for $\{\Pi_\eta^N\}_1^\infty$. Therefore, $\bar{d}(\Pi_\eta^N, \Pi_{\overline{\eta}}^N) \to 0$ and $\bar{d}(\Pi_\xi^N, \Pi_\eta^N) \to \beta$. □

3. General processes and flows. In the preceding section we formulated a single scheme S which worked no matter how large the alphabet was. It is this feature that we shall now exploit to extend the results there to more general processes. We begin with a discussion of the $\bar{d}$-distance. We now have processes $\{x_n\}_0^\infty, \{y_n\}_0^\infty$ with values in a separable metric space (M, d). We will say that the $\bar{d}$-distance between the processes is less than C if there is a joining of the two processes $\{(x_n, y_n)\}_0^\infty$ as a stationary process and

$$(*) \qquad \Pr(d(x_0, y_0) > C) < C.$$

If M happens to be compact, then the closeness in measure between x_0 and y_0

can be replaced by other, perhaps more natural, measures of closeness such as

$$\text{Expectation}(d(x_0, y_0)).$$

However, for unbounded spaces the notions are not equivalent and it is easier to work with (*). The $\bar{d}$-distance is defined to be the infimum of the C's for which there is a joining with (*) holding.

Let us see what happens if we apply the scheme S from Theorem 5 to a general process. There is no formal problem—we simply take for n in the interval $(L(N), L(N+1))$ the empirical N-block distribution in $(\xi_1, \ldots, \xi_n)$ and concatenate independently. If the distribution of x_0 is continuous, then typically all of the n-blocks will be different; however, the recipe still makes perfectly good sense. Now we claim that we will have convergence in $\bar{d}$ just as before! This arises from the fact that x_0 is essentially finite-valued in the sense that given $\varepsilon > 0$, one can find a finite number of sets $U_1, \ldots, U_a \subset M$, each of diameter at most ε and

$$\Pr\left\{x_0 \notin \bigcup_{i=1}^{a} U_i\right\} < \varepsilon.$$

Consider the finite-valued random variable $\bar{x}_n$ which has the value i if $x_n \in U_i$ and ∞ if $x_n \notin \bigcup_{i=1}^{a} U_i$. We think of ∞ as being an error in the joining even if both sides have that value.

We now see that the process we have constructed by S after being coarse-grained by this partition, is the same as what S would have constructed as in Section 2, from the coarse-grained observations $\{\bar{\xi}_i\}_1^\infty$, in other words S commutes with coarse-graining. The effect of ignoring ∞ is at most 2ε, and thus the convergence of Theorem 4 guarantees the convergence of this generalized procedure.

THEOREM 1. *The guessing scheme of Section* 2, *Theorem* 4, *will converge in* $\bar{d}$ *for a general B-process with values in a separable metric space.*

PROOF. The property that we need now is that a factor of a B-process is B (see [6]). Thus the coarse-grained process is B and Theorem 4 applies in the way we explained above. □

In the same way that we deduced Theorem 2.6 from the proof of Theorem 2.4 we can also establish its analog for any process (always with values in a separable metric space).

THEOREM 2. *There is a sequence of real-valued functions* $\bar{S}^{(n)}$ *defined on pairs of outputs of stationary stochastic processes such that for any two processes* $\{x_n\}_1^\infty, \{y_n\}_1^\infty$, *where* $\{y_n\}$ *is B and a.e. realization* $\{\xi_n\}_1^\infty, \{\eta_n\}_1^\infty$, *we have that*

$$\lim_{n\to\infty} \bar{S}^{(n)}(\xi_1, \ldots, \xi_n, \eta_1, \ldots, \eta_n) = \bar{d}\left(\{x_n\}_1^\infty, \{y_n\}_1^\infty\right).$$

Now that we have established the convergence of our schemes for general B-processes we make the extension to stochastic processes in continuous time. A stationary continuous parameter stochastic process $\{x_t\}_{t\geq 0}$ is B if and only if every derived discrete-time process is B. It suffices for this that the discrete-time process defined by

$$(*) \qquad \bar{x}_n = \{x_t\colon n \leq t < n-1\}$$

is a B-process. Here if the values of x_t lie in (M, d), the values of $\bar{x}_n$ lie in the space of measurable functions on $[0, 1]$ with values in (M, d). We metrize this space with the metric of convergence in measure. One defines the $\bar{d}$-distance between continuous-time processes in a fashion completely analogous to discrete processes. We shall need the following simple lemma which relates this $\bar{d}$-distance to the $\bar{d}$-distance between the discretization $\bar{x}_n, \bar{y}_n$ defined by $(*)$.

LEMMA 3. *If* $\bar{d}(\{\bar{x}_n\}_0^\infty, \{\bar{y}_n\}_0^\infty) < \varepsilon$, *then also* $\bar{d}(\{x_t\}_{t\leq 0}, \{y_t\}_{t\leq 0}) < \varepsilon$.

PROOF. Take some joining of $\{\bar{x}_n\}, \{\bar{y}_n\}$ that realizes the $\bar{d}$-distance. This joining gives a probability distribution on the space of pairs $(\xi_n, \eta_n)_{n=0}^\infty$ which we would like to use to give a $\bar{d}$-joining of the original processes. Identifying $(\xi_n, \eta_n)_{n=0}^\infty$ in a natural way with the pair of functions $(\xi_t, \eta_t)_{t\geq 0}$ gives us a measure on the correct space, and its restriction to $(\xi_t), (\eta_t)$ is correct. The only problem is that it is not stationary with respect to all of $\mathbb{R}_+$. To solve that problem, one simply averages that distribution over all shifts in $0, 1$. Since the restriction to the coordinate functions are shift-invariant for all $t \geq 0$ this yields a joining of the required type. Finally, it is easy to check that the distance in probability between x_0 and y_0 now becomes what the distance between $\bar{x}_0$ and $\bar{y}_0$ was. □

With this lemma in mind our next step is to extend the scheme S of Theorem 1 to continuous time by simply taking for $n \in (L(N), L(N+1))$ the empirical distribution of N-blocks in $\{\xi_t\}_{t=0}^n$ where we only look at n-blocks that start at an integer value of t (an N-block is now $\{\xi_t\}$, $i \leq t \leq i+N$, for some integer i) and then concatenating independently N-blocks with that distribution. This gives a continuous-time process which is just like what we would have gotten had we applied our scheme to $\{\xi_t\}$. Because of Theorem 1 and the lemma we get that these independent concatenations now-converge in $\bar{d}$ to the original flow if it was B.

We have proved the following theorem.

THEOREM 4. *With the modifications above we have that if* $\{x_t\}_{t\geq 0}$ *is a* B*-process, then for a.e. realization* $\{\xi_t\}_{t\geq 0}$, $S((\xi_t\colon 0 \leq t \leq T))$ *converges in* $\bar{d}$ *to* $\{x_t\}_{t\geq 0}$.

We leave the obvious modifications of Theorem 2 to this context to the reader and go on to further generalizations of the main theorem to other index sets. Let ν range now over the elements of $\mathbb{Z}^d$ and consider a process $\{x_\nu\}_{\nu\in\mathbb{Z}^d}$

which is stationary with respect to all shifts, i.e., for any $\nu_0 \in \mathbb{Z}^d$ the processes $\{x_\nu\}, \{x_{\nu+\nu_0}\}$ have the same distribution. Once again the definition of $\bar{d}$ carries over, and the same kinds of questions can be raised as before where now one makes the observations over some finite box $\{v\colon |v_i| \le n,\ 1 \le i \le d\} = B_n$, i.e., $\{\xi_\nu\colon \nu \in B_n\}$, and tries to learn something about the process as a whole from such finite observations.

For such processes there is an entropy theory and a Shannon–McMillan theorem (cf. [5], [3], [15] and [4]) and also a theory of B-processes and finitely determined processes which is completely analogous to the theory for Z-processes (cf. [5] and in even greater generality [9] and [11]). In order to extend Theorem 1 to this setting we need to go through the analogs of Lemmas 1–3 of Section 2. We shall not do this in detail and content ourselves with pointing out that the arguments used in proving those lemmas had an essentially combinatorial flavor and made very little use of the precise structure of the interval $\{1, \dots, n\}$ and N-blocks (except for the initial and final segment); such arguments extend to tilings of big cubes in $\mathbb{Z}^n$ by smaller cubes.

Having made these observations, it is routine to carry over to these processes the results of Section 2. The extension to $\mathbb{R}^n$ parameter processes presents no further obstacles. We refrain from wearying the reader with the statements of the results that he or she can easily supply. In the same spirit in the following sections we formulate results for discrete time but their obvious analogs hold for continuous-time processes as well.

4. Can one discriminate between two processes? We address here the general problem of discriminating between distinct processes. The situation we envision is that we are observing two sequences of outputs, $(\xi_1, \dots, \xi_n, \dots), (\eta_1, \dots, \eta_n, \dots)$ and we would like to know if they represent typical outputs of the same underlying process or of distinct underlying process. If one output sequence is Bernoulli, then by applying our scheme of Section 2 we can determine the issue, as we saw in Theorem 6 there. Here we shall show that outside of $\mathscr{B}$, the class of Bernoulli-processes, even this simple "yes–no" question of "same–different" cannot be answered in an effective way. We shall, at first, for simplicity, describe a construction which will assume the existence of a scheme S that for any two processes $\{x_n\}_1^\infty, \{y_n\}_1^\infty$ and a.e. pair of realizations $(\xi_n)_1^\infty, (\eta_n)_1^\infty$ will converge to the d-distance between the processes. With that assumption we will construct a *single* process that has the property that when S is applied to an a.e. pair of realizations we will get as an upper limit a number that is bounded away from 0! Ex post facto, this is a contradiction to the properties of such a scheme S and then what we have done is show that such an S cannot exist. However, by a minor modification, we can carry out the construction knowing only that S has the requisite property for Bernoulli-processes. As we have already emphasized such an S exists. Thus the construction below can be interpreted as producing a process which in spite of ergodicity will behave, with respect to any S that works for Bernoulli, as though it were many processes in the $\bar{d}$ sense.

The construction will be carried out by describing a sequence of distributions Π_n on blocks of length l_n. This is done by giving a list of words of length l_n in the alphabet $\{0, 1\}$ and assigning to each such word equal probability. This gives a joint distribution on the $\{0, 1\}$-valued random variables $\{x_1, \ldots, x_{l_n}\}$. The construction is carried out in such a way that it is manifest that for any fixed k, the average of the distributions of $\{x_i, \ldots, x_{i+k-1}\}$ as i ranges over $\{1, 2, \ldots, l_n - k + 1\}$ converges to a limit as n tends to ∞. These distributions fit together to form a stationary process, simply because they are obtained by averaging over longer and longer intervals.

Parenthetically, we should remark that a process could also be constructed in a formal way out of the description below via the method of "cutting and stacking" well known to ergodic theorists. While this method is perhaps more natural in the geometrical framework of a space and a measure-preserving transformation, the probabilistic method described above is simpler and more natural in our present context.

CONSTRUCTION. Start with four specific blocks of length 128 given as follows:

$$\begin{aligned} a &= 0101\ldots01 &&= (01)^{64}, \\ b &= 000011110000\ldots &&= (0^4 1^4)^{16}, \\ c &= \ldots &&= (0^{16} 1^{16})^4, \\ d &= \ldots &&= (0^{64} 1^{64})^1. \end{aligned}$$

Build now two processes one, say X_1, obtained by concatenating a and b independently and the other X_2 obtained by concatenating c and d independently. We claim that the $\bar{d}$-distance between the two processes is at least $2/5$ since one cannot match better any string of a's and b's with a string of c's and d's.

When the purported guessing scheme is applied to typical outputs of the two processes we eventually get the response that they are at least $1/4$ apart in $\bar{d}$. Thus there is some r_1, and sets of outputs of X_1, X_2 of length r_1, say R_1, R_2, such that each R_1, R_2 has probability at least $99/100$ and for *each* $w \in R_1$, $99/100$ of the elements $u \in R_2$, will give $S^{(r_1)}(w, u) \geq 1/4$, and similarly for each $u \in R_2$, $99/100$ of the elements w of R_1 will give the same. Notice that we cannot say $S^{(r_1)}(w, u) \geq 1/4$ for all pairs $w \in R_1$, $u \in R_2$ since large subsets in product measure spaces need not contain large rectangles.

Now the ergodic theorem applied to each process separately gives us some l_1 so that most blocks of x_i of length l_1 are $49/50$ covered by R_i. We take for the blocks of length l_1 of our construction all of these, taking care to have an equal number of each type. Thus we have the blocks of length l_1,

$$A_1, A_2, \ldots, A_{k_1}; B_1, B_2, \ldots, B_k$$

such that with high probability an r_1-block w selected from the A's and an r_1-block u selected from the B's will cause $S^{(r_1)}(w, u)$ to give a reading greater than $1/4$. This completes step 1 of the construction.

Step 2: Here we will merge together the blocks of the two kinds of A's and B's in order to ensure that the ultimate process will be ergodic. We will also take this opportunity to build many different types of blocks which mutually are at a fixed distance apart in $\bar{d}$. This is needed for the next stage in which we will simulate many different processes. If the number of different processes does not grow, then we will not be able to maintain a fixed distance apart in $\bar{d}$ during the construction.

Take $a_i = 100^i$, $1 \le i \le 10^3$, and $N_2 = 100^{10^4}$. For each i, first form all possible concatenations of a_i-blocks from among $A_1, \ldots, A_k$, and separately all possible concatenations of a_i-blocks from among the B_j's, and then form all possible concatenations of these larger building blocks taken N_2/a_i times. The number of blocks of length $N_2 l_1 = l_2$ that we get in this way is

$$(2k^{a_i})^{N_2/a_i} = 2^{N_2/a_i} k^{N_2}$$

and these blocks form the ith group of blocks constructed in step 2.

Notice that blocks of different groups have the property that any reasonable initial segment of one compared with a final segment of the other have A's lining up against B's at least half the time, so that they disagree at least $1/8$ of time. (If the number of processes would remain fixed, then this fraction would eventually go to 0.) Also notice that instead of four groups that we started out with we have now 1000 groups to work with in the next stage. Observe that with very high probability each block from every group contains many repetitions of *all* the blocks from the previous stage. This will ensure the ergodicity of the resulting process.

$$\begin{array}{ccc} \vdots & \vdots & \vdots \end{array}$$

Step $2n + 1$: At stage $2n$ we will have constructed a collection of blocks of length l_{2n}, divided into $f(n)$ groups $G_1^{2n}, \ldots, G_{f(n)}^{2n}$ (each group G_i^{2n} contains at least $2n$ elements) with the property that any two l_{2n}-blocks from distinct groups will differ in at least $\alpha_n l_{2n}$ places and moreover any initial segment of length k greater than $\varepsilon_{2n} l_{2n}$ from one group differs in at least $\alpha_n k$ places from any final segment of length k of any block from a different group. As we proceed we must make certain that α_n is bounded away from 0, and to do this we will need $\varepsilon_{2n} \to 0$.

For each group G_i^{2n} we form a process $X_i^{(n)}$ by concatenating independently the blocks from G_i^{2n}. These processes will satisfy

$$\bar{d}\left(X_i^{(n)}, X_j^{(n)}\right) \ge \alpha_n - 3\varepsilon_n$$

by our assumption, and therefore the scheme S will yield that with probability tending to 1 if w_i, w_j are representative l-blocks of the processes $X_i^{(n)}, X_j^{(n)}$,

then

$$S^{(l)}(w_i, w_j) \geq \alpha_n - 4\varepsilon_n.$$

As before this gives for any desired $\delta > 0$ and r_n and classes $R_i^{(n)}$, $1 \leq i \leq f(n)$, of blocks of length r_n from the processes $X_i^{(n)}$ such that

(i) $\Pr(R_i^{(n)}) \geq 1 - \delta$;
(ii) for *all* $U_i \in R_i$, $j \neq i$,

$$\Pr\{u \in R_j^{(n)}: S^{(rn)}(u_i, u) \geq \alpha_n - 4\varepsilon_n\} \geq 1 - 2\delta.$$

Apply now the ergodic theorem to each of these $f(n)$-processes to get some l_{2n+1} so that most blocks of length l_{2n+1} from the $X_i^{(n)}$-process have at least a $(1 - 2\delta)$-fraction of their r_{2n}-blocks belonging to $R_i^{(n)}$ and having distribution $(1 - 2s)$ close to dist $R_i^{(n)}$. The new groups $G_i^{(2n+1)}$, $1 \leq i \leq f(n)$, consist precisely of those l_{2n+1}-blocks of the $X_i^{(n)}$-process that have this last-mentioned property. It is also easy to arrange at this point that all of these groups have the same number of elements.

The key feature that we get from this part of the construction is that in our ultimate process, when we pick two R_{2n}-blocks u, u' at random, then with high probability they come from different processes and will cause $S^{(r_{2n})}(n, n')$ to be sure to be at least $\alpha - 4\varepsilon_n$. To be sure, we still have the property that blocks from different groups are at least $\alpha_{2n} - 4\varepsilon_{2n} = \alpha_{2n+1}$ apart, and the new ε_{2n+1} can be made as small as we please by having l_{2n+1} long enough.

Step $2n + 2$: Once again we want to merge together the blocks from different groups to get ergodicity. Define now

$$a_i = M^i, \qquad 1 \leq i \leq f(n+1), \qquad N_{n+1} = M^{2^{n+1}},$$

where M is taken to be larger than $\max_i |G_i^{(2n+1)}|^2$, and $f(n+1)$ will be specified below.

As before, form $G_i^{(2n+2)}$ in two steps:

(i) First form all possible concatenations of $a_i l_{2n+1}$-blocks chosen from within each group $G_j^{(2n+1)}$ to get $f(n)|G_j^{(2n+1)}|^{a_i}$ distinct $l_{2n+1}a_i$-blocks.

(ii) Now concatenate N_{n+1}/a_i times in some cyclic order the elements from (i).

These new blocks, with high probability, contain the blocks of the previous level with equal probability. This justifies our remark above that the joint distributions defined by assigning equal probability to all possible l_n-blocks converge to a limiting distribution. To check the disagreement between blocks from different groups we now use the fact that because of the great differences in the periodicity between the appearances of the old groups in the newly formed groups the loss in α_{2n+1} is now proportional to $1/f(n)$, so ε_{2n+1} can be taken to be as small as we please by enlarging M if necessary. It is easy to see how to choose $f(n+1)$ so that limit $\alpha_n = \alpha > 0$.

Continuing in this way, we complete the definition of our stationary process which can easily be seen to be ergodic. It has the property that if two r_n-blocks are chosen at random from the process, then with high probability they will come from different processes and S will say that they are from processes that are at least $\frac{1}{2}\alpha$ apart in $\bar{d}$.

REMARK 1. The infinite processes produced at the even stages could be modified to make them Bernoulli. This implies that if S is a scheme that works for Bernoulli-processes, then there is a process such that with probability 1 if we choose two outputs independently, then S will infinitely often say that the outputs come from a process greater than $\alpha > 0$ apart. α could be taken close to $1/2$ and close to 1 if we consider processes on more than two states.

The modification needed for Bernoulliness is to insert between any two blocks an extra space with probability $1/2$ and independently. The result is a $l_n + 1$-step Markov process (l_n being the length of the block being concatenated), which is easily seen to be mixing and therefore Bernoulli.

REMARK 2. The process that we construct in Remark 1 that confuses S can be made K. (It obviously has positive entropy.) One way to achieve this is the following: Whenever we concatenate blocks we first enlarge each block by adding f_n extra terms, e_n of them to the beginning and $f_n - e_n$ to the end, f_n will be fixed but e_n will be random and will take values between 0 and f_n with equal probability and independently of everything else. If $f_n \to \infty$ but $\sum^{\infty} f_n/l_n < \infty$, then it is easy to see that the resulting process is K (we need to show that for any fixed k, the output of the process between time 0 and k is independent of the distant past). Pick n so the distribution of k-sequences in most l_n-sequences look like the unconditioned distribution. Then $(0, k)$ will lie in some n-block and knowledge of the distant past will not tell us the e_n at the beginning of the n-block. This means that we are sampling the k-sequences in our l_n-sequence uniformly. Details of this kind of construction can be found in [8].

REMARK 3. The construction as it stands gives an example with the property that

$$\limsup S(\eta_1, \ldots, \eta_n, \xi_1, \ldots, \xi_n) \geq \beta > 0$$

for a.e. pair $(\eta_n)_1^{\infty}, (\xi_n)_1^{\infty}$ of realizations of the process. In order to get a stronger result, namely that along a sequence of n's of positive upper density one has $\lim \ldots \geq \beta$, one has to modify the construction at even n's. Then one takes $l_{n+1} \gg M_{n+1}$ large enough so that with high probability most indices $i \in \{1, l_{n+1}\}$ are such that $(\xi_i, \xi_{i+1}, \ldots)$ falls in the set described above (that had probability greater than $1 - \delta_n$).

5. Non-*B*-processes. (The results in this section are written in the notation of discrete time but hold also for continuous time.) The scheme we gave in Section 2 for guessing the process was shown to converge in $\bar{d}$ to the process if and only if the process itself was B. It is natural to ask what would happen if $\{x_n\}_1^\infty$ is a stochastic process that is not B but is close to a B-process. It might be, for example, that we are observing a B-process $\{y_n\}_1^\infty$ to which some noise has been added, producing $\{x_n\}_1^\infty$ with $\bar{d}(\{x_n\}_1^\infty, \{y_n\}_1^\infty) < \varepsilon$. What we shall see is that our scheme is robust in the sense that although $S(\xi_1, \ldots, \xi_n)$ will not converge in $\bar{d}$ to $\{x_n\}$ it will eventually get into an ε-ball (in the $\bar{d}$-metric) surrounding $\{x_n\}_1^\infty$. In fact a sharper theorem is true, namely, the distance in $\bar{d}$ between $S^{(n)}(\xi_1, \xi_2, \ldots, \xi_n)$ and the fixed B-process $\{y_n\}$ converges to the $\bar{d}$-distance between $\{x_n\}$ and $\{y_n\}$.

THEOREM 1. *If $\{y_n\}_1^\infty$ is a B-process, then for a.e. realization $(\xi_n)_1^\infty$ of $\{x_n\}_1^\infty$ we have that*

$$\lim \bar{d}\big(S(\xi_1, \ldots, \xi_n), \{y_n\}_1^\infty\big) = \bar{d}\big(\{x_n\}_1^\infty, \{y_n\}_1^\infty\big),$$

where S is the scheme of Section 2, Theorem 4.

PROOF. As we saw in Section 2, the finite distributions of $S(\xi_1, \ldots, \xi_n)$ converge to the finite distributions of the $\{x_n\}_1^\infty$ process. For that weak convergence one only needs ergodicity. Because of that, for any fixed k and n large, one can read from a $\bar{d}$-joining between $S(\xi_1, \ldots, \xi_n)$ and $\{y_n\}_1^\infty$ a $\bar{d}$-joining between $(x_1, \ldots, x_k)$ and $(y_1, \ldots, y_k)$. Going to the limit gives us the following general fact:

$$\bar{d}\big(\{x_n\}_1^\infty, \{y_n\}_1^\infty\big) \le \liminf_{n\to\infty} \bar{d}\big(S(\xi_1, \ldots, \xi_n), \{y_i\}_1^\infty\big)$$

for a.e. realization $(\xi_i)_1^\infty$, and any process $\{y_i\}_1^\infty$.

Conversely, for almost every $(\xi_n)_0^\infty$ we can find a realization $\{\eta_n\}_0^\infty$ for Y that satisfies two properties

(i) $$\bar{d}\big(S^{(n)}((\eta_0, \ldots, \eta_n)), Y\big) \to 0,$$

(ii) $$\frac{1}{n}\sum_0^n |\eta_i - \xi_i| \to \bar{d}(Y, X).$$

By the nature of our scheme (ii) implies $\bar{d}(S^{(n)}(\eta_0, \ldots, \eta_n), S^{(n)}(\xi_0, \ldots, \xi_n)) \to \bar{d}(Y, X)$ which together with (i) proves the theorem.

We can also establish a converse to Theorem 1. Namely, if we know that the distance between $\{x_n\}$ and any process that is either B or the direct product of B and a rotation is greater than or equal to some constant β, then for a.e. realization $(\xi_n)_1^\infty$ the scheme $S(\xi_1, \ldots, \xi_n)$ will oscillate by at least β in the sense that the diameter of the set $\{S(\xi_1, \ldots, \xi_n),\ n_0 \le n < n_1\}$ cannot become smaller than β for each n_0 as $n_1 - n_0$ tends to ∞.

Here is the proof. As in the proof of Theorem 1 we have the following fact:

$$\bar{d}(\{x_n\}_0^\infty, \{y_n\}_0^\infty) \leq \liminf_{n\to\infty} \bar{d}(S(\xi_0, \ldots, \xi_n), \{y_i\}_0^\infty).$$

This fact is now applied with $\{y_n\}_1^\infty$ equal to $S(\xi_1, \ldots, \xi_{n_0})$ which is the direct product of a B and a rotation. A slight variation of the definition of the scheme S will produce B-processes for every value of n without changing any of the earlier results. What we have in mind is simply to add a random spacer with probability $1/2$ after each block in the independent concatenation. A more intrinsic way of doing this would be to keep at every stage the distribution of both N-blocks and $N-1$-blocks and concatenate them independently where one first chooses independently with probability $1/2$ which of the two types of blocks to place and then fills that in independently with the empirical distributions. None of these variants affects finite distributions or entropy as $N \to \infty$ and they all give processes that are certainly B. (By the B theory it is enough to check "very weak Bernoulli" and this is very easy.) The scheme $\hat{S}$ in the next theorem is one of these above variants. We will denote the class of B-processes by $\mathscr{B}$.

THEOREM 2. *If $\bar{d}(\{x_n\}_1^\infty, \mathscr{B}) > \beta$, then for a.e. realization $\{\xi_n\}_1^\infty$ for all n_0 there is some $n_1 > n_0$ with*

$$\bar{d}\big(\hat{S}(\xi_1, \ldots, \xi_{n_0}), \hat{S}(\xi_1, \ldots, \xi_{n_1})\big) \geq \beta.$$

This pair of theorems that shows how the behavior of S on a particular output depends upon the distance between the underlying process and the class of B-shifts raises the following questions: Can one determine in any effective way, i.e., via some sequence of functions $f_n(\xi_n, \ldots, \xi_n)$, what the distance is between the underlying process and the class of B-processes? More precisely, we would like ideally

$$\lim_{n\to\infty} f_n(\xi_1, \ldots, \xi_n) = \bar{d}(\{x_n\}_1^\infty, \mathscr{B})$$

for a.e. realization $(\xi_n)_1^\infty$ of the process $\{x_n\}_1^\infty$.

It turns out that this is impossible and indeed much weaker desiderata cannot be fulfilled. Suppose, for example, that f_n is a sequence of functions that have the property that for a.e. realization (ξ_n) of a B-process

$$\lim_{n\to\infty} f_n(\xi_1, \ldots, \xi_n) = 0.$$

Then we will construct a zero-entropy process $\{y_n\}_1^\infty$ such that, almost surely, its realizations $(\eta_n)_1^\infty$ will satisfy

$$\liminf_{n\to\infty} f_n(\eta_1, \ldots, \eta_n) = 0.$$

Thus not only can we not get a good estimate for the distance to the class of B-processes, we cannot even tell that we do not belong to the class.

Begin by considering the independent $[0, 1]$-valued process with $\Pr(x_n = 0) = P(x_n = 1) = 1/2$. If m_1 is sufficiently large, then for a set of strings

$V_1 = \{(\xi_1, \ldots, \xi_n)\}$ of length m_1, with probability $1 - 1/100$, $f_{m_1}(\xi_1, \ldots, \xi_{m_1})$ will be less than $1/100$.

For n_1 sufficiently large, there is now a set of strings of length n_1, say W_1, such that if $(\xi_1, \ldots, \xi_{n_1}) \in W_1$, for a set of $i \in (1, n_1 - m_1)$ of relative frequency at least $1 - 2/100$ we have that

$$(\xi_i, \xi_{i+1}, \ldots, \xi_{i+m_1-1}) \in V_1.$$

We will begin by having this set of words W_1 for the n_1-blocks of our construction. In order to drop the entropy, we form n_1^2-blocks by repeating each block n_1 times. Next we form a B-process by concatenating independently the $(1 + n_1^2)$-blocks and the set of $(n_1^2 + 2)$-blocks formed by adding one and two 2's, respectively, to each block in our earlier set. [This is to make sure that the $(n_1^2 + 1)$-step Markov process so formed is mixing and hence B.] Once again since we have a B-process $\{Z_n\}_1^\infty$, if m_2 is sufficiently large there is a set of strings of length m_2, say V_2, such that for all $(\xi_1, \ldots, \xi_n)_{m_2} \in V_2$ we have that $f_{m_2}(\xi_1, \ldots, \xi_{m_2}) < 10^{-4}$. Applying the ergodic theorem as before, there is an n_2 and a set W_2 of strings of length n_2 such that for every $(\xi_1, \ldots, \xi_{m_2}) \in W_2$ for a set of, i.e., $(1, n_2 - m_2)$ of relative frequency at least $1 - 2/10^4$ we have that

$$(\xi_i, \xi_{i+1}, \ldots, \xi_{i+m_2-1}) \in V_2.$$

To drop the entropy even more, we repeat each word in W_2 n_2 times and get our set of n_2^2-blocks for the second step in the construction. Continuing in this way, we get our $\{y_n\}_1^\infty$.

6. Nonstationary processes and concluding remarks. Up to now we have considered only stationary stochastic processes where the subtlety of the questions lies in the fact that we would like to draw conclusions based on finite observations. For nonstationary processes the ergodic theorem is no longer available and already it is not clear that observing the infinite string of outputs $(\xi_0, \ldots, \xi_n, \ldots)$ teaches us a good deal about the process. Now we place a natural restriction on the kind of functions of infinite strings $(\xi_n)_0^\infty$ and require that they be Borel functions in the natural Borel structure. Blackwell has given a beautiful example [2] that shows that one cannot even distinguish between two quite disparate families of processes by Borel functions. Let P_0 be the class of 0–1-valued processes such that

$$\text{limit in probability of } x_n = 0,$$

$$\text{i.e., } \lim_{n \to \infty} P(x_n = 0) = 1,$$

and P_1 the class where $\lim_{n \to \infty} P(x_n = 1) = 1$. We think of P_0, P_1 as being measures on $X = \{0, 1\}^N$. Then Blackwell shows that there is no Borel partition of X into sets $A_0 \cup A_1$ such that for every measure $\mu \in P_i$ we have $\mu(A_i) = 1$, and this in spite of the easily verifiable fact that for any fixed pair $\mu_i \in P_i$, $i = 0, 1$, we do have a partition of X into Borel sets $A_0 \cup A_1$ s.t.

$$\mu_0(A_0) = \mu_1(A_1) = 1.$$

Our contribution to this circle of ideas consists in pointing out that this phenomenon of not being able to distinguish two families depends upon the noncompactness of the families P_0, P_1 or putting it another way is a truly infinite phenomenon. Observe that if two families could be separated by a single pair of sets A_0, A_1, then convex combinations of the elements in each family would also be separated by A_0, A_1. We will prove the following theorem.

THEOREM 1. *If P_0, P_1 are compact convex sets of measures on $X = \{0,1\}^N$ such that for each $\mu_i \in P_i$, $i = 0, 1$, there is a partition of X into sets A_0, A_1 with $(*)$ $\mu_0(A_0) = \mu_0(A_1) = 1$, then there exists a single Borel partition (A_0, A_1) such that $(*)$ holds for all $\mu_i \in P_i$, $i = 0, 1$.*

This is essentially equivalent to a finite result where X is replaced by a finite set and $(*)$ is replaced by an approximate kind of separation, say $\geq 1 - \varepsilon$ instead of $= 1$. In that case the compactness of the P_i is no longer relevant and this is the sense in which Blackwell's phenomenon is a feature of the infinite.

It is a little more transparent to begin with the case in which P_1 consists of a single probability measure, say λ. Fix an $\varepsilon > 0$ and let $\Phi_\varepsilon = \{0 \leq \phi \leq 1: \phi \text{ is continuous and } \int \phi \, d\lambda \leq \varepsilon\}$. We set our first goal to find a $\phi \in \Phi_\varepsilon$ such that

$$\int \phi \, d\mu \geq 1 - \varepsilon \quad \text{for all } \mu \in P_0.$$

For such a ϕ, $A_\varepsilon = \{x: \phi_\varepsilon(x) \geq 1/2\}$ will satisfy $\lambda(A_\varepsilon) \leq 2\varepsilon$, and $\mu(A_\varepsilon) \geq 1 - 2\varepsilon$ for all $\mu \in P_0$.

Then

$$A = \bigcap_{n=1}^{\infty} \bigcup_{k \leq n} A_{2^{-k}}$$

will satisfy $\lambda(A) = 0$, $\mu(A) = 1$ for all $\mu \in P_0$ and $(A, X \setminus A)$ will give us the required partition.

Now since $\mu \perp \lambda$ for all $\mu \in P_0$, for each fixed μ we can find a continuous function $\phi \in \Phi_\varepsilon$ with $\int \phi \, d\mu \geq 1 - \varepsilon/2$. We find a single such ϕ by the following procedure. Consider elements of Φ_ε to define continuous functions on P_0 by setting for $\hat{\varphi} \in \phi_\varepsilon$

$$\hat{\varphi}(\mu) = \int \hat{\varphi} \, d\mu,$$

and denote by $\hat{\Phi}_\varepsilon$ the subset of $C(P_0)$ that we get in this way. Let $\Psi \subset C(P_0)$ denote those functions on P_0 that are everywhere greater than or equal to $1 - \varepsilon$. If $\Psi \cap \hat{\Phi}_\varepsilon$ would be empty by the Hahn–Banach theorem (note that Ψ has an interior point) there would be a linear functional on $C(P_0)$, i.e., a measure on P_0, m, such that

$$\int_{P_0} \psi(\mu) \, dm(\mu) \geq a \geq \int_{P_0} \hat{\phi}(\mu) \, dm(\mu), \quad \text{all } \psi \in \Psi, \hat{\phi} \in \hat{\Phi}_\varepsilon.$$

Since dm is bounded from below on Ψ it is a positive linear functional; hence we can normalize it so as to be a probability measure. Then since $1 - \varepsilon \in \Psi$ we have $a \leq 1 - \varepsilon$ and thus

$$(*) \qquad \int \hat{\phi}(\mu)\, dm(\mu) \leq 1 - \varepsilon \quad \text{for all } \hat{\phi} \in \hat{\Phi}_\varepsilon.$$

But if $\bar{\mu}$ denotes the barycenter of m in P_0 (recall that P_0 was compact convex), then $(*)$ becomes

$$\int_X \phi(x)\, d\bar{\mu}(x) \leq 1 - \varepsilon \quad \text{for all } \phi \in \Phi_\varepsilon,$$

which contradicts the fact that for $\bar{\mu} \in P_0$ there is some $\phi \in \Phi_\varepsilon$ with

$$\int_X \phi(x)\, d\bar{\mu}(x) \geq 1 - \varepsilon/2.$$

Thus $\Psi \cap \hat{\Phi}_\varepsilon$ is nonempty and we have the required continuous function ϕ and we have completed the proof of Theorem 1, in case P_1 reduces to a single point. For the general case one can use this special case and now repeat the argument above using for Φ_ε the set

$$\left\{0 \leq \phi \leq 1\colon \phi \text{ continuous such that } \int_X \phi(x)\, d\lambda(x) \leq \in \text{ for all } \lambda \in P_1\right\}.$$

The special case we have just established shows that for every fixed $\mu \in P_0$ there is some $\phi \in \Phi_\varepsilon$ with

$$\int \phi\, d\mu \geq 1 - \varepsilon/1$$

and the argument continues as before. This concludes the proof of Theorem 1. Note that the set obtained is a nice set in the sense that not only is it Borel, it can even be made to be a G_δ.

REMARKS. 1. By examples similar to the ones in Sections 4 and 5 one can show that no scheme of the type that we have been considering can distinguish between K-automorphisms and Bernoulli-processes. We do not know of any general discrimination between types of processes that can be done by finite-valued schemes.

2. Ziv [16] has given another method of estimating the entropy of a process from a sequence of observations. It has a simpler algorithmic form than our method but does not give the type of information that we needed and so we were not able to use it.

3. The results we have presented show what is possible in principle. It is worth investigating to what extent these ideas can lead to "practical" algorithms. We hope to return to this issue in the near future.

REFERENCES

[1] BAILEY, D. H. (1976). Sequential schemes for classifying and predicting ergodic processes. Ph.D. dissertation, Stanford Univ.

[2] BLACKWELL, D. (1980). There are no Borel SPLITs. *Ann. Probab.* **8** 1189–1190.

[3] CONZE, J. P. (1972). Entropie d'un groupe Abélian de transformations. *Z. Wahrsch. Verw. Gebiete* **25** 11–30.

[4] FELDMAN, J. (1980). r-entropy, equipartition and Ornstein's isomorphism theorem in R^n. *Israel J. Math.* **36** 321–343.

[5] KATZNELSON, Y. and WEISS, B. (1972). Commuting measure-preserving transformations. *Israel J. Math.* **12** 161–173.

[6] ORNSTEIN, D. S. (1974). *Ergodic Theory, Randomness and Dynamical Systems*. Yale Univ. Press, New Haven, Conn.

[7] ORNSTEIN, D. S. (1978). Guessing the next output of a stationary process. *Israel J. Math.* **30** 292–296.

[8] ORNSTEIN, D. S. and SHIELDS, P. C. (1973). An uncountable family of K-automorphisms. *Adv. in Math.* **10** 63–88.

[9] ORNSTEIN, D. S. and WEISS, B. (1980). Ergodic theory of amenable group actions. I. The Rohlin lemma. *Bull. Amer. Math. Soc.* (*N.S.*) **2** 161–165.

[10] ORNSTEIN, D. S. and WEISS, B. (1983). The Shannon–McMillan–Breiman theorem for a class of amenable groups. *Israel J. Math.* **44** 53–60.

[11] ORNSTEIN, D. S. and WEISS, B. (1987). Ergodic theory of amenable groups. *J. d'Analyse Math.* **48** 1–141.

[12] ORNSTEIN, D. S. and WEISS, B. (1989). Statistical properties of chaotic systems. Preprint.

[13] PARK, K. (1980). A flow built under a step function with a multi-step Markov partition on a base. Ph.D. dissertation, Stanford Univ.

[14] SHIELDS, P. C. (1979). Almost block independence. *Z. Wahrsch. Verw. Gebiete* **49** 119–123.

[15] THOUSVENOT, J.-P. (1972). Convergence en moyenne de l'information pour l'action de Z^2. *Z. Wahrsch. Verw. Gebiete* **24** 135–137.

[16] ZIV, J. (1978). Coding theorems for individual sequences. *IEEE Trans. Inform. Theory* **24** 405–412.

DEPARTMENT OF MATHEMATICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305

DEPARTMENT OF MATHEMATICS
HEBREW UNIVERSITY
JERUSALEM
ISRAEL