ENTROPY IS THE ONLY FINITELY OBSERVABLE INVARIANT

DONALD ORNSTEIN AND BENJAMIN WEISS

(Communicated by Anatole Katok)

ABSTRACT. Our main purpose is to present a surprising new characterization of the Shannon entropy of stationary ergodic processes. We will use two basic concepts: isomorphism of stationary processes and a notion of finite observability, and we will see how one is led, inevitably, to Shannon’s entropy. A function \( J \) with values in some metric space, defined on all finite-valued, stationary, ergodic processes is said to be finitely observable (FO) if there is a sequence of functions \( S_n(x_1, x_2, ..., x_n) \) that for all processes \( \mathcal{X} \) converges to \( J(\mathcal{X}) \) for almost every realization \( x_1^\infty \) of \( \mathcal{X} \). It is called an invariant if it returns the same value for isomorphic processes. We show that any finitely observable invariant is necessarily a continuous function of the entropy. Several extensions of this result will also be given.

1. INTRODUCTION

One of the central concepts in probability is the entropy of a process. It was first introduced in information theory by C. Shannon [S] as a measure of the average informational content of stationary stochastic processes and was then carried over by A. Kolmogorov [K] to measure-preserving dynamical systems, and ever since it and various related invariants such as topological entropy have played a major role in the theory of dynamical systems.

Our main purpose is to present a surprising new characterization of this invariant. We shall use two ideas: the first, isomorphism invariants, drawn from abstract ergodic theory and the second, finite observability, taken from statistics, and we will see how one is led, inevitably, to entropy. Since the second idea is less familiar we will begin by explaining it.

We will be interested in what kind of features of a stationary stochastic process \( \mathcal{X} \) can be determined in an effective way by observing more and more values of the process. For this to make sense we restrict ourselves to ergodic processes where a single infinite sequence of outputs will, in general, determine the process uniquely. We will also assume that all of our ergodic processes are infinite; that is, we exclude the purely periodic sequences from our discussion. Furthermore, we will confine our attention to finite-valued processes. If \( J \) is a function of ergodic processes taking values in a metric space (\( \Omega, d \)), then we say

Received May 4, 2006.

2000 Mathematics Subject Classification: Primary 37A35, Secondary 60G10.

Key words and phrases: entropy, finitely observable, isomorphism invariants, finitary isomorphism.
that $J$ is \textit{finitely observable} if there is some sequence of functions $S_n(x_1, x_2, \ldots, x_n)$ that converges to $J(\mathcal{X})$ for almost every realization of the process $\mathcal{X}$, for all ergodic processes. A weaker notion would involve convergence in probability of the functions $S_n$ to $J$ rather than convergence almost everywhere. The particular labels that a process carries play no role in the following and so we may assume that all our processes take values in finite subsets of $\mathbb{Z}$.

Here are some examples of $FO$ functions. If $J(\mathcal{X}) = \mathbb{E}\{X_0\}$ is the expected value of $X_0$ then the basic pointwise ergodic theorem of G. D. Birkhoff implies that $J$ is $FO$ via the estimators

$$S_n(x_1, x_2, \ldots, x_n) = (x_1 + x_2 + \ldots + x_n)/n.$$  

This may easily be generalized as follows. Denote by $\mathcal{P}$ the shift-invariant probability measures on $\mathbb{Z}^\mathbb{Z}$ with support on a finite number of symbols and the topology of convergence in finite distributions. Then to each finite-valued stationary process there will correspond a unique element of $\mathcal{P}$, namely its distribution function $\text{DIST}(\mathcal{X})$. This function is also $FO$ by the same argument, replacing the arithmetic averages of the $x_i$ by the empirical distributions of finite blocks. Next consider the memory length $L(\mathcal{X})$ of a process. This equals the minimal $m$ such that the process is an $m$-Markov process, and $+\infty$ if no such $m$ exists. In [MW] it is shown that this function is $FO$. A better-known example is the Shannon entropy of a process. Here, several different estimators $S_n$ are known to converge to the entropy; cf. [B, Z, OW, OW2, KASW].

Recall that two processes $\mathcal{X}$ and $\mathcal{X}'$ are isomorphic if there is a stationary invertible coding going from one to the other. More formally, let us denote the bi-infinite sequence $\ldots x_{-2}, x_{-1}, x_0, x_1, x_2, \ldots$ by $x_{\infty}^{-\infty}$, and the shift by $\sigma$ where

$$(\sigma x)_n = x_{n+1} \text{ for all } n$$

A coding from $\mathcal{X}$ to $\mathcal{X}'$ is a mapping $\phi$ defined on the sequences $x_{\infty}^{-\infty}$ with values in $\mathcal{X}'$, which maps the probability distribution of the $\mathcal{X}$ random variables to that of the $\mathcal{X}'$ random variables. It is stationary if almost surely $\phi \sigma = \sigma' \phi$, where $\sigma'$ is the shift on $\mathcal{X}'$. Finally, it is invertible if it is almost surely one-to-one. In this case it is not hard to see that the inverse mapping, where defined, will yield a stationary coding from $\mathcal{X}'$ to $\mathcal{X}$.

While the definition of the entropy of a process was given by C. Shannon [S] it was the great insight of A. Kolmogorov [K] that it is in fact an isomorphism invariant. This enabled him to solve an outstanding problem in ergodic theory; namely, he proved that independent processes with differing entropies are not isomorphic. Since that time entropy has turned out to be fundamental in many areas of ergodic theory. It is perhaps somewhat surprising that no new invariants of that kind were discovered and it is our purpose to explain why this is so. The first version of our main result is the following:

**Theorem 1.** If $J$ is a finitely observable function, defined on all ergodic finite-valued processes, that is an isomorphism invariant, then $J$ is a continuous function of the entropy.
A stronger version replaces isomorphism by a more restricted notion of finitary isomorphism. These are isomorphisms where the codings, in both directions, depend only on a finite (but variable) number of the variables. In the natural product topology on the space of finite sequences these are codings that are continuous after the removal of a null set. About ten years after Kolmogorov’s result D. Ornstein [O] showed the converse; namely, independent processes with the same entropy are isomorphic. This was strengthened to finitary isomorphism by M. Keane and M. Smorodinsky [KS], and is a strictly stronger notion than isomorphism, since there are many examples of processes that are isomorphic but not finitarily isomorphic.

It is rather remarkable that such a qualitative description leads one inevitably to the entropy. Before discussing some extensions of this and some open problems, we will briefly sketch a proof of the fact that under the hypotheses of the theorem $J$ is a function of the entropy (ignoring the continuity requirement). Assuming the contrary means that there is a finitely observable function $J$ and two processes $\mathcal{X}, \mathcal{Y}$ with equal entropy such that $J(\mathcal{X}) \neq J(\mathcal{Y})$. Furthermore, since the values that $J$ assumes lie in a metric space, there is some positive $d_0$ such that $d(J(\mathcal{X}), J(\mathcal{Y})) > 3d_0$. By considering the balls with radii $d_0$ around these values we can deduce from the finite observability of $J$ the existence of a sequence of functions $S_n(x_1, \ldots, x_n)$ with three values $-1, 0, 1$, such that for any process isomorphic to $\mathcal{X}$ the values of $S_n$ will eventually be 1, for any process isomorphic to $\mathcal{Y}$ the values of $S_n$ will eventually be $-1$, while for no ergodic process will we be able to see with positive probability both of these values $-1, 1$ infinitely often.

An explicit construction in §2 will falsify this possibility. In zero entropy such a construction was given in [W] and a closely related one was carried out in [OW] to show that the process function defined by $B(\mathcal{X}) = 1$ in case $\mathcal{X}$ is Bernoulli, i.e., isomorphic to an independent process, and $B(\mathcal{X}) = 0$ otherwise, is not finitely observable.

In many contexts one has some prior information about the nature of the process so that it is natural to ask about process functions that are defined only for a restricted class of processes. Recall, for example the class $\mathcal{K}$ of finite-valued processes with a trivial tail field. These are closed under isomorphism and include all Bernoulli processes. For this class the analogue of our main theorem continues to hold. In fact, we can establish such an analogue for any class that includes the Bernoulli processes. However, if we explicitly exclude these from $\mathcal{K}$, then our proof breaks down and we do not know whether or not there are invariants defined on the non-Bernoulli $\mathcal{K}$ processes that are finitely observable and not functions of the entropy.

2. The Construction

We will carry out the key construction under weaker assumptions on the purported sequence of functions $S_n$ from which we will derive a contradiction. Fix a finite alphabet $A$, and a value for the entropy $h < \log |A|$. We assume that the functions $S_n$ are defined on $A^n$ and take values in some metric space $\Omega$, and that
for every ergodic process $\mathcal{X}$ with values in $A$, $S_n(x^n)_{\mathcal{X}}$ converges in probability to a point $f(\mathcal{X})$ in $\Omega$ that depends only on the finitary isomorphism class of $\mathcal{X}$. We are using here the notation $x^n = (x_1, x_2, ..., x_n)$. Furthermore we suppose that there are two $A$-valued processes $\mathcal{X}_i$, $i = 0, 1$, with the same entropy $h$ such that $f(\mathcal{X}_0) \neq f(\mathcal{X}_1)$.

With these assumptions we shall construct an ergodic $A$-valued process $\mathcal{Y}$, such that $S_n(y^n)_{\mathcal{Y}}$ will fail to converge in probability. The process $\mathcal{Y}$ will be specified by giving better and better approximations to its finite distributions. We will need a criterion, formulated in terms of these finite distributions, that will guarantee the ergodicity of the limit process. To explain the criterion we need some notation for the empirical distributions of $k$-blocks in $n$-strings. Both terms, blocks and strings, mean the same thing; we use both to avoid monotony.

Let $b \in A^k$ be a fixed $k$-block and $u \in A^n$ an $n$-string; then define
\[ D(b \mid u) = |\{1 \leq i \leq n - k + 1 : u[i, ..., i + k - 1] = b\}|/(n - k + 1). \]

Fix some $A$-valued ergodic process and let $P$ denote the probability functions that the process determines on $A^n$. It is an immediate consequence of the ergodic theorem that for every $k \in \mathbb{N}$, and every $\varepsilon > 0$, we have some $n = n(k, \varepsilon)$ such that
\[ P|u \in S^n : |P(b) - D(b \mid u)| < \varepsilon, \text{ for all } b \in A^k > 1 - \varepsilon. \]

In this last equation the limiting probabilities $P(b)$ appear. Since they may not be known a priori it is better to recast the equation as follows. There exist subsets $G_n \subset A^n$ that satisfy $P(G_n) > 1 - \varepsilon$ and
\[ |D(b \mid u) - D(b \mid v)| < \varepsilon, \text{ for all } b \in A^k \text{ and for all } u, v \in G_n \]

This last condition can be turned into a criterion for ergodicity as is formulated in the next proposition which is readily derived from the mean ergodic theorem.

**Proposition 1.** If for some sequence $\varepsilon_k$ tending to zero there are $n_k$’s and subsets $G_{n_k} \subset A^{n_k}$ that satisfy $P(G_{n_k}) > 1 - \varepsilon_k$ and
\[ |D(b \mid u) - D(b \mid v)| < \varepsilon_k, \text{ for all } b \in A^k \text{ and for all } u, v \in G_{n_k} \]
then the process that $P$ defines is ergodic.

The successive stages in the construction will be carried out using the following proposition. In the proposition we will need a technical condition that concerns the topological support of a process. Recall that the topological support of a process $\mathcal{X}$ on the finite alphabet $A$ is the closed, shift-invariant subset $\mathcal{X} \subset A^\mathbb{Z}$ that is obtained after omitting all finite cylinder sets with zero probability. We will denote by $h_{top}(\mathcal{X})$ the topological entropy of $(\mathcal{X}, \sigma)$. This is just the exponential growth rate of the blocks of length $n$ that have a positive probability. Clearly, this is an upper bound for the Shannon entropy of a process.

**Proposition 2.** Given two ergodic processes $\mathcal{X}, \mathcal{Y}$ on the finite alphabet $A$ with equal entropy $h$ and $h_{top}(\mathcal{X}) < \log |A|$, a finite integer $L$, and a positive number...
δ, there is a process \( Z \) on the same alphabet \( A \) that is finitarily isomorphic to \( X \), whose \( L \)-block distributions are within \( \delta \) of those of the \( Y \) process and with \( h_{top}(Z) < \log |A| \).

The proposition with isomorphism replacing finitary isomorphism can be established by some minor modifications in the proof of Krieger’s theorem on the existence of finite generators for finite entropy systems. The proof would be somewhat easier if we wouldn’t demand that the \( Z \)-process have the same alphabet, but this is essential for our iterative construction. The method that we will use to construct finitary isomorphisms goes back to the pioneering paper of Monroy and Russo [MR] and was heavily exploited in [KS]. For the proof we will need the following technical lemma whose proof we postpone:

**Lemma 1.** Given an ergodic process \( \mathcal{X} \) with entropy \( h \), for any positive \( \alpha \), if \( M \) is sufficiently large, then we can find for each \( m > M \) a collection of words \( \mathcal{H}_m \subset A^m \) with \( |\mathcal{H}_m| < 2^{(h + \alpha)m} \) and for almost every realization of the \( \mathcal{X} \)-process, \( x_{-\infty}^\infty \), and any parsing of \( x_{-\infty}^\infty \) into blocks all of whose lengths are at least \( M \), the blocks in the parsing that do not belong to the appropriate \( \mathcal{H}_m \) occupy a set of positions whose upper density is less than \( \alpha \).

**Proof of Proposition 2.** We will describe the coding from the \( \mathcal{X} \)- to the \( \mathcal{Z} \)-process at the same time that we construct the \( \mathcal{Z} \)-process itself in the following fashion. For a typical realization \( x_{-\infty}^\infty \) of the \( \mathcal{X} \)-process we will give a rule for writing down a corresponding point of the \( \mathcal{Z} \)-process \( z_{-\infty}^\infty \). This rule will not make any use of the specific indices of the \( x_i \) but only of their linear ordering and thus will define a proper shift-invariant coding. The finitary character will be apparent from the construction, as will the injectivity, and the distribution on the \( z_i \)’s will be taken to be the one induced by this coding from the \( \mathcal{X} \)-process. This will suffice to establish that \( \mathcal{Z} \) is finitarily isomorphic to \( \mathcal{X} \). Naturally we will have to show how the other requirements will be met.

**Markers.** A single block \( e \) of length \( l \) with \( P(X_l^1 = e) > 0 \) will be used as a marker. For typical realizations of the \( \mathcal{X} \)-process this \( e \) will occur infinitely many times in both directions and so placing a comma after each \( x_j \) with \( x_{j+1}^j = e \) will parse typical sequences into finite blocks. If \( i < j \) are successive such indices we will now give a rule for \( z_j^i \) that depends only on \( x_j^i \). Thus our coding deals separately with each block that occurs between successive commas. It will be important that there be a minimal size for these blocks, and indeed this can be guaranteed to be any preassigned number \( M \).

Observe that our system \( (\hat{X}, \sigma) \) is infinite and thus has nonperiodic points. If \( e \), a block of length \( l \), defines a sufficiently small neighborhood of a nonperiodic point then successive entries to \( e \) are separated by at least \( M \) time instants, and since \( \hat{X} \) is the topological support of the \( \mathcal{X} \)-process, the probability of visiting \( e \) is positive. Thus using such an \( e \) for the parsing satisfies our requirements. The length of \( e \) might be longer than \( M \) but that possibility has no effect on what we are doing. The actual minimal size for \( M \) will be specified later on but it will definitely be much larger than \( L \) so that in calculating the empirical distribution
of $L$-blocks in the $\mathcal{Z}$-process we will be able to concentrate on what happens inside the $z_i^j$-blocks.

In order for the coding to be injective we will need to be able to recover this parsing into blocks from the $\mathcal{Z}$-sequences. To this end we will use a special word at the beginning of each $z_i^j$ that occurs only there. Fix two symbols from our alphabet $A$, say $\{0, 1\}$, and for some $M_0$ (that will eventually be chosen to be much smaller than $M$) let $f = 10^{2M_0}1$, that is to say, a 1 followed by $2M_0$ zeros and then a 1. At the start of each block of the $\mathcal{Z}$-process corresponding to the parsing, $z_i^j$, there will be a copy of $f$ while within each block we will place a 1 at distances that are less than $2M_0$ apart. This copy of $f$ serves as a marker and replaces the $e$ which as we have already said might be much longer than $M_0$. This will guarantee that the original parsing is uniquely recoverable from the corresponding $\mathcal{Z}$-sequence.

To achieve our main goal of having the $L$-block distributions of $\mathcal{Z}$ be close to those of $\mathcal{Y}$ we will encode the $x_i^j$-blocks by $z_i^j$-blocks whose empirical $L$-block distribution is close to that of the $\mathcal{Y}$-process. To this end we will need to have a sufficiently large supply of distinct blocks of this type. We proceed to describe how to find a sufficient supply of these.

**Words with good $L$-block distribution.** Fix a small $\alpha > 0$ and apply the ergodic theorem and the Shannon–McMillan theorem to the $\mathcal{Y}$-processes in order to find an $R$ large enough so that there is a collection $\mathcal{G} \subset A^R$ that satisfies

- $|\mathcal{G}| > 2^R(h-\alpha)$,
- for each $u \in \mathcal{G}$ and each $b \in A^L$:

$$|P(Y_i^L = b) - D(b | u)| < \delta/10.$$ 

We can economize on notation and suppose that $M_0 = R$. For any $m \geq M$ we begin with $f$ and appending to each block in $\mathcal{G}$ a single 1 concatenate independently until we fill a block of length $m$. This will give us a large supply of $m$-words whose empirical $L$-block distribution is close to optimal. However, this supply will not be large enough since the best we can expect for an upper bound on the number of $m$-blocks in the $\mathcal{X}$-process that we are trying to encode is $2^{m(h+\alpha)}$. To enlarge the supply we do the following. In the last $(\delta/2)m$ places, we no longer insist on using these $(M_0 + 1)$-blocks, but rather allow any letters from the alphabet $A$ (still putting a 1 every $M_0$ places). For fixed $\delta$ if $\alpha$ is sufficiently small and $M$ large compared with $M_0$, this will do what is required, namely give us an adequate supply of $m$-blocks with a good $L$-block distribution.

By the lemma, if $M$ is sufficiently large then in our parsing of $\mathcal{X}_\infty$ by occurrences of $e$ we can identify most (up to a density $\alpha$) of the blocks as belonging to $\mathcal{H}_m$. These blocks will be encoded in a one-to-one fashion by the blocks we have just fashioned with a good $L$-block distribution. The remaining blocks, whose number is estimated by the topological entropy of $\mathcal{X}$ if $M$ is large enough, can be uniquely encoded by using a fixed family of $M_0$-blocks that is disjoint from $\mathcal{G}$.
to replace $G$ in the construction above. This will yield, for $a$ sufficiently small, a process whose topological entropy will be strictly less than $\log |A|$.

To recapitulate, we have described how to associate, in an invertible fashion, to almost every realization of the $X$-process a sequence in the alphabet $A$ with an $L$-block distribution close to that of the $Y$-process. These sequences inherit the distribution of the $X$-process and form a process that is finitarily isomorphic to the $X$-process. This completes the proof of the proposition.

We return now to Lemma 1.

Proof of Lemma 1. The lemma is another variant of the Shannon–McMillan theorem along single orbits which is explained in [W2, Chapter 9]. Begin by choosing an auxiliary small constant $\beta$, and applying the Shannon–McMillan theorem to find an $N$, and a collection $\mathcal{K} \subset A^N$ that satisfies

- $|\mathcal{K}| < 2^{N(h+\beta)}$
- $P[X^N_1 \in \mathcal{K}] > 1 - \beta^2/2$

Next we need to count for large $m$ the number of words $u \in A^m$ such that

$$|[1 \leq i \leq m - N + 1 : u[i, ..., i + N - 1] \in \mathcal{K}]|(m - N + 1) > 1 - \beta.$$

This set will be denoted by $\mathcal{H}_m$ and these are the sets that the lemma calls for. We assume that $m > M$ where $M$ is chosen so that $N/M$ is negligible. To estimate the size of $\mathcal{H}_m$ start with a $u \in \mathcal{H}_m$ and define a sequence $[1 \leq i_1 < i_2 ... < i_k]$ as follows. Let $i_1$ be the minimal index $\geq 1$ such that $u[i_1, ..., i_1 + N - 1] \in \mathcal{K}$. Let $i_2$ be the minimal index $\geq i_1 + N$ such that $u[i_2, ..., i_2 + N - 1] \in \mathcal{K}$, and continue in this fashion as long as $i_k + N - 1 \leq m$. If $N$ is sufficiently large the number of such choices of sequences is exponentially small in $m$ (say at most $2^{m\beta}$) and for a fixed sequence $[1 \leq i_1 < i_2 < ... < i_k]$ the number of ways of filling in with words from $\mathcal{K}$ where appropriate and with arbitrary letters from $A$ in the remaining places (at most $\beta m$ in number) gives the estimate:

$$|\mathcal{H}_m| < 2^{m(h+2\beta+\beta \log |A|)}.$$

By the ergodic theorem for almost every realization of the $X$-process the density of $i$’s such that $x^{i+N-1}_i \in \mathcal{K}$ is at least $1 - \beta^2$. For any parsing into blocks of minimal size $M$, the usual Chebyshev inequality will yield that those blocks where the relative density is at least $1 - \beta$ occupy a set of density at least $1 - \beta$. Choosing $\beta < a/(2 + \log |A|)$ will give us the lemma.

With these propositions it is an easy task to prove the main part of our theorem which asserts that a finitely observable invariant of finitary isomorphism must be a function of the entropy. We will defer the proof of the continuity to the next section. We will formulate this as a theorem:

**Theorem 2.** Fix a finite alphabet $A$, and a value for the entropy $h < \log |A|$. Assume that functions $S_n$ are defined on $A^n$ and take values in some metric space $\Omega$, and that for every ergodic process $X$ with values in $A$, $S_n(x^n)$ converges in probability to a point $J(X)$ in $\Omega$ which depends only on the finitary isomorphism class.
of $\mathcal{X}$. Then there cannot be two $A$-valued processes $\mathcal{X}_i$, $i = 0, 1$ with the same entropy $h$ and topological entropy $h_{\text{top}}(\mathcal{X}) < \log |A|$ such that $f(\mathcal{X}_0) \neq f(\mathcal{X}_1)$.

**Remark.** The hypothesis in the theorem concerning topological entropy doesn’t really affect its applicability to our claim that finitely observable invariants that are isomorphism invariants are functions of the entropy. This is because we are assuming that the function $f$ is defined for all finite state processes and if there were a pair of processes falsifying the claim, then we could arrange for all assumptions of the theorem to be valid by choosing the size of the alphabet $A$ large enough.

**Proof.** As we already explained, the strategy of the proof is to use the hypotheses in order to construct an $A$-valued ergodic process for which the $S_n$’s will not converge in probability. Set $\omega_0 = f(X_0), \omega_1 = f(X_1)$, and $3d = \text{dist}(\omega_0, \omega_1) > 0$. The variables of processes like $\mathcal{X}_0$ where there is a subscript will be written as $0X_i$. Let $L_1$ be sufficiently large such that

$$P(\text{dist}(S_{L_1}(0X_1^{L_0}), \omega_0) < d) > 1 - 10^{-1}$$

Apply now Proposition 2 with $L = L_1, \delta = 10^{-2}$, $\mathcal{Y} = \mathcal{X}_0$, and $\mathcal{X} = \mathcal{X}_1$ to get a process, which we will denote by $\mathcal{Z}_1$, that is finitarily isomorphic to $\mathcal{X}_1$ but has its $L_1$-block distributions within $10^{-2}$ of those of $\mathcal{X}_0$. As a consequence we get

$$P(\text{dist}(S_{L_1}(1z_1^{L_1}), \omega_0) < d) > 1 - 10^{-1} - 10^{-2}$$

Next we use the fact that $\mathcal{Z}_1$ is finitarily isomorphic to $\mathcal{X}_1$ in order to find an $L_2 > L_1$ large enough such that

$$P(\text{dist}(S_{L_2}(1z_1^{L_2}), \omega_1) < d) > 1 - 10^{-2}$$

Apply now Proposition 2 with $L = L_2, \delta = 10^{-3}$, $\mathcal{Y} = \mathcal{Z}_1$ and $\mathcal{X} = \mathcal{X}_0$ to get a process, which we shall denote by $\mathcal{Z}_2$, that is finitarily isomorphic to $\mathcal{X}_0$ but has its $L_2$-block distributions within $10^{-3}$ of those of $\mathcal{Z}_1$. As a consequence we get

$$P(\text{dist}(S_{L_2}(2z_1^{L_2}), \omega_1) < d) > 1 - 10^{-2} - 10^{-3}$$

This ping-pong between processes, alternately isomorphic to $\mathcal{X}_1$ and $\mathcal{X}_0$, is continued indefinitely and since the $L_i$ tend to infinity there is a limiting process $\mathcal{Z}$ for which

$$\lim_{i \to \infty} P(\text{dist}(S_{L_i}(iZ_1^{L_i}), \omega_1) \leq d) = 1$$

on the one hand, while on the other hand

$$\lim_{i \to \infty} P(\text{dist}(S_{L_{2i+1}}(iZ_1^{L_{2i+1}}), \omega_0) \leq d) = 1.$$
this proposition, when choosing \( L_2 \) for example, we also choose it large enough so that we could find a subset \( G_{L_2} \subset A^{L_2} \) that satisfies \( P(x_1^{L_2} \in G_{L_2}) > 1 - 10^{-2} \) and

\[
|D(b | u) - D(b | v)| < 10^{-2}, \text{for all } b \in A^2 \text{ and for all } u, v \in G_{L_2}
\]

This condition will be preserved in the limit with \((10/9)^{10} \) replacing \( 10^{-2} \), and doing this at all stages will give us the required ergodicity. \( \Box \)

3. Proof of the Main Theorem

After we have Theorem 2 in hand what remains to be done is to show that if \( F \) is a function from the nonnegative reals to a metric space such that \( J(\mathcal{X}) = F(h(\mathcal{X})) \) is finitely observable, then \( F \) is necessarily a continuous function. Once again our strategy is to argue by contradiction. Namely, we assume that for some finite \( t \geq 0 \) there is a sequence \( t_i \) converging to \( t \) such that

\[
\text{dist}(F(t_i), F(t)) > d > 0
\]

for all \( i \) and proceed to construct an ergodic process \( \mathcal{X} \) with entropy equal to \( t \) for which \( S_n(x_i^t) \) will not converge in probability to \( F(t) \). We may assume that the sequence \( t_i \) is monotonic. The argument is simpler in the case where \( t_i \) is decreasing and we begin with that case.

Case I - Discontinuity from above. The class of processes that we will use for this construction is the class of block-independent processes. These are processes consisting of independent concatenations of blocks of a fixed length \( m \), say, on an alphabet \( A \) with a distribution \( \mu \) on \( A^m \). The starting point of these blocks is a uniform random variable on \( \{1, 2, ..., m\} \). This is to ensure that the resulting process is ergodic. As is well known, the entropy of any such process is given by \( 1/mH(\mu) \) where \( H \) is the standard entropy function for finite distributions. To begin with we can take \( m = 1 \) and choose a distribution \( \mu_1 \) such that \( H(\mu_1) = t_1 \). Let \( \mathcal{X}_1 \) be the independent process with this distribution. Since \( J \) is assumed to be finitely observable, there is an \( L_1 \) such that

\[
P(\text{dist}(S_{L_1}(x_1^{t_1}), F(t)) > d) > 1 - 10^{-1}.
\]

The idea is now to build a process \( \mathcal{X}_2 \) whose \( L_1 \)-block distribution is close to that of \( \mathcal{X}_1 \) but having entropy equal to \( t_2 \). This is easy to achieve if \( t_2 < t_1 \). Indeed, by the ergodic theorem, if \( M \) is sufficiently large, with high probability the empirical distribution of blocks of length \( L_1 \) in words of length \( M_1 \) is within \( 10^{-2} \) of the distribution induced by \( \mu_1 \). Furthermore, if we were to use only one of these \( M_1 \)-strings and repeat periodically we would get zero entropy, while using all of them with the distribution coming from \( \mathcal{X}_1 \) would give us average entropy \( t_1 \), so that by continuity it is clear that we can use some of these good words of length \( M_1 \) and find a distribution \( \mu_2 \) on \( M_1 \)-blocks with average entropy equal to \( t_2 \) exactly. Concatenating \( M_1 \)-blocks independently with this distribution gives us an ergodic process \( \mathcal{X}_2 \) with entropy equal to \( t_2 \) but still satisfying

\[
P(\text{dist}(S_{L_1}(x_1^{t_2}), F(t)) > d) > 1 - 10^{-1} - 10^{-2}.
\]
Since \( J = F(h) \) is finitely observable by assumption (8) there is an \( L_2 \) so that

\[
P(\text{dist}(S_{L_2}(2x_1^{L_2}), F(t)) > d) > 1 - 10^{-2}
\]

Having this \( L_2 \) in hand we can find an \( M_2 \) sufficiently large so that with high probability \( M_2 \)-blocks are good in the sense that the empirical distribution of the \( L_2 \)-words in them is within \( 10^{-3} \) of the distribution coming from the \( \mathcal{X}_2 \)-process. As before since \( t_3 < t_2 \) we can use some of them to get a distribution \( \mu_3 \) on good blocks of length \( M_2 \) having an average entropy equal to \( t_3 \). Concatenating \( M_2 \)-blocks independently with this distribution gives us an ergodic process \( \mathcal{X}_3 \) with entropy equal to \( t_3 \) but still satisfying equations (10) and (11) with an additional \( -10^{-3} \) on the right-hand side. It should be pretty clear now how to continue this procedure indefinitely, yielding a sequence of processes \( \mathcal{X}_i \) that converge in finite distribution to a limiting process \( \mathcal{X} \). This limiting process can be guaranteed to be ergodic by the use of Proposition 1 of the preceding section. Furthermore, its entropy is equal to the limit of the \( t_i \), namely \( t \). In the limit equations like (9) will continue to hold \( \mathcal{X} \) with \( 1 - (10/9)10^{-1} \) on the right-hand side, and these will contradict the assumptions made on \( J = F(h) \).

**Case II - Discontinuity from below.** Our assumption now is that equation (8) holds for some increasing sequence \( t_i \) converging to \( t \). The argument is a bit trickier now, since the semi-continuity of the entropy does not allow us to preserve arbitrary finite distributions while increasing the entropy of a process and maintaining at the same time a fixed finite alphabet. The way around this problem is to work with an alphabet \( A \) with \( \log |A| > t \) and use the fact that an increase in the entropy can be arranged that is proportional to the change in finite distributions. Following is the lemma that we shall use.

**Lemma 2.** Let \( s < t < \log |A| \) be given together with an ergodic process \( \mathcal{X}_s \) on the alphabet \( A \) with entropy equal to \( s \). Given any \( M \) and \( \delta > 0 \) there is an ergodic process \( \mathcal{Y} \) with entropy equal to \( s + \delta(\log |A| - t) \) and with \( M \) block distribution within \( 2\delta \) of that of \( \mathcal{X}_s \).

**Proof.** By applying the ergodic theorem and the Shannon–McMillan theorem we can find an \( n \) large enough so that with very high probability the empirical distribution of \( M \)-blocks in strings of length \( n \) are within \( \delta \) of their distribution in the \( \mathcal{X}_s \)-process. We can arrange things so that the average entropy of those good \( n \)-blocks is extremely close to \( s \). If we define a distribution on blocks of length \( N/(1 + \delta) \) by concatenating independently this distribution on good \( n \)-blocks with the uniform distribution on all \( A \)-valued blocks of length \( N \delta/(1 + \delta) \) we get an average entropy at least equal to \( s + \delta(\log |A| - t) \) and modifying the uniform distribution a little bit would enable us to get precisely that value for the entropy. Concatenating this distribution on blocks of length \( N \) yields a process \( \mathcal{Y} \) with the same entropy and it is clear that we achieve the desired degree of approximation for the distributions of length \( M \).

The main point of the lemma is that the entropy increase is independent of the length \( M \). The proof that was carried out in case I above can now be repeated.
The role of the fixed changes that were allowed there, namely $10^{-i}$, is taken over by the increments $t_{i+1} - t_i$ that play the role of the $10^{-i}$. At the $k$-th stage in the construction we want to increase the entropy from $t_k$ to $t_{k+1}$. Since $t_{k+1} < t$, by applying the lemma, we can do this with a change in finite distribution (of any required length) less than $2(t_{k+1} - t_k)/(\log|A| - t)$. Both the ergodicity of the limiting distribution and the fact that its entropy is really equal to $t$ come about by having the changes take place in sufficiently long finite distributions.

Combining the two cases we get that only continuous functions of the entropy can be finitely observable. Together with the preceding section this completes the proof of the theorem:

**Theorem 3.** If $J$ is a finitely observable function, defined on all stationary ergodic finite-valued processes, that is an invariant of finitary isomorphism, then $J$ is a continuous function of the entropy.

4. Process Functions with Restricted Domain

The results of the preceding sections were based on the assumption that the process function under discussion is defined on the entire class of finite-valued stationary ergodic processes. It is also quite natural to suppose that the function $J$ has a restricted domain such as the class of mixing processes, zero-entropy processes, processes that are $K$ (i.e., with a trivial tail or, in probabilistic language, satisfying Kolomogorov’s 0-1 law, whence the label $K$), pure point spectrum processes, etc. The main theorem that we established in the preceding section will continue to be valid in any situation where the construction that was carried out there concludes with an ergodic process $\mathcal{Z}$ that is in the domain of the function $J$. We will give a couple of examples where we know how how to carry out this program and conclude with several open questions.

The simplest class is that of weakly mixing processes. One of the characterizations of weak mixing of a process $X$ is that the independent joining is ergodic. This means that if we take an independent copy of $X$, say $X'$, then the pair process defined by $y_n = (x_n, x'_n)$ is ergodic. The same technique that we used to guarantee that the limit process in the construction was ergodic can also be used to guarantee weak mixing - assuming of course that the processes that are used in the construction are known to be weak mixing. This enables us to prove the following:

**Proposition 3.** If $J$ is a finitely observable function, defined on the class $\mathcal{W}$ of weakly mixing finite-valued stationary processes, that is an invariant of finitary isomorphism, then $J$ is a continuous function of the entropy.

Recall that the Bernoulli processes (B-processes) are those processes isomorphic to an independent process. We can also prove the following:

**Theorem 4.** If $J$ is a finitely observable function, defined on some class $\mathcal{C}$ of finite-valued stationary processes, closed under isomorphism, and containing the B-processes, that is an invariant of finitary isomorphism, then $J$ is a continuous function of the entropy.
The proof involves a portion of the basic theory of B-processes, exposed for example in [O], which we will summarize briefly. There is a metric on stationary processes denoted by \( \bar{d} \) that takes into account the long-range behavior of a process in a way that the weak* topology doesn’t. With this metric the space of finite-valued processes with a fixed alphabet becomes a complete metric space but it is far from separable. For example, if one partitions the unit circle into two half circles and considers two rotations \( \alpha, \beta \) that are independent over the rationals then the \( \bar{d} \) distance between them is always equal to 1/2. One of the basic properties of B-processes is that they are finitely determined (FD). A process \( \mathcal{X} \) is finitely determined if for any \( \epsilon > 0 \) there is a \( \delta > 0 \) and an \( L \) such that for any process \( \mathcal{Y} \) whose \( L \)-block distributions are within \( \delta \) of those of \( \mathcal{X} \) and whose entropy is within \( \delta \) of the entropy of \( \mathcal{X} \) one has that \( \bar{d}(\mathcal{X}, \mathcal{Y}) < \epsilon \). In fact this property characterizes the B-processes. Another basic fact is that the class of B-processes is closed under the \( \bar{d} \) metric.

With this background it is easy to see how to prove Theorem 4. We begin with the analogue of Theorem 2. We may assume that one of the two processes on which the function \( J \) takes different values is a B-process. At stage \( k \), where we are constructing a process isomorphic to the non-B-process we can make sure that its finite distributions are close enough to those of the B-process so that its \( \bar{d} \) distance to the B-process is less than \( 1/2^k \). In the same way, for the following step where a process is constructed isomorphic to the B-process we ensure a small error in \( \bar{d} \) by controlling the finite distributions. In this way one sees that the limiting process, \( \mathcal{Z} \), is in the \( \bar{d} \)-closure of the B-processes and hence it too is a B-process. In a similar way one guarantees that the processes constructed to show the continuity of the function that is applied to the entropy are B-processes. There the argument is even easier since there is more freedom in the choice of the process. However, one still needs the result on \( \bar{d} \)-limits of B-processes.

Recently, Y. Gutman and M. Hochman [GH] have proved a theorem to the effect that if \( \mathcal{X} \) is a zero-entropy extension of \( \mathcal{Y} \) and \( \mathcal{C} \) is the class of processes isomorphic to either \( \mathcal{X} \) or \( \mathcal{Y} \), then any finitely observable invariant defined on \( \mathcal{C} \) is a constant. This easily implies that any finitely observable invariant defined on the Kronecker systems or the zero-entropy mixing systems is a constant. Many questions of this nature still remain. Probably foremost among them is whether there can be a nonconstant finitely observable invariant on the \( K \) systems with a fixed entropy that are not Bernoulli.

REFERENCES

[B] D. H. Bailey, Sequential schemes for classifying and predicting Ergodic processes. Ph. D. thesis, Stanford University, 1976.

[DGS] M. Denker, C. Grillenberger and K. Sigmund, Ergodic Theory on Compact Spaces, LNM 527, Springer, 1976.

[GH] Y. Gutman and M. Hochman, On some classes of systems for which entropy is the only finitely observable invariant (preprint).
Entropy per unit time as a metric invariant of automorphisms (Russian), Dokl. Akad. Nauk SSSR 124 (1959) 754–755.

I. Kontoyiannis, P. Algoet, Y. M. Suhov and A. J. Wyner, Nonparametric entropy estimation for stationary processes and random fields, with application to English text, IEEE Transactions on Information Theory 44 (1998) 1319–1327.

M. Keane and M. Smorodinsky, Bernoulli schemes of the same entropy are finitarily isomorphic, Ann. of Math. 109 (1979) 397–406

G. Monroy and L. Russo, A family of codes between some Markov and Bernoulli schemes, Comm. Math. Phys. 43 (1975) 155–159.

D. S. Ornstein, Ergodic Theory, Randomness, and Dynamical Systems, Yale University Press, New Haven, 1974.

D. S. Ornstein and B. Weiss, How sampling reveals a process, The Annals of Probability 18 (1990) 905–930.

D. S. Ornstein and B. Weiss, Entropy and data compression schemes, IEEE Transactions on Information Theory 39 (1993) 78–83.

D. J. Rudolph, Fundamentals of Measurable Dynamics: Ergodic Theory on Lebesgue Spaces, Oxford University Press, New York, 1990.

C. E. Shannon, A mathematical theory of communication, Bell System Tech. J. 27 (1948) 379–423, 623–656.

P. Shields, The Theory of Bernoulli Shifts, The University of Chicago Press, Chicago, 1973.

B. Weiss, Some remarks on filtering and prediction of stationary processes, Israel J. of Math. 149 (2005) 345–360.

B. Weiss, Single Orbit Dynamics, AMS–CBMS No. 95, 2000.

J. Ziv, Coding theorems for individual sequences, IEEE Transactions on Information Theory 24 (1978) 405–412.

DONALD ORNSTEIN <ornstein@math.stanford.edu>: Department of Mathematics, Stanford University, Stanford, CA 94305

BENJAMIN WEISS <weiss@math.huji.ac.il>: Institute of Mathematics, Hebrew University of Jerusalem, Jerusalem, Israel 91904