The Annals of Probability
1992, Vol. 20, No. 1, 403–409

ENTROPY AND PREFIXES[1]

BY PAUL C. SHIELDS

University of Toledo and Eötvös Loránd University

Grassberger suggested an interesting entropy estimator, namely,

$$\frac{n \log n}{\sum_{i=1}^n L_i^n},$$

where L_i^n is the shortest prefix of $x_i, x_{i+1}, \dots,$ which is not a prefix of any other $x_j, x_{j+1}, \dots,$ for $j \le n$. We show that this estimator is not consistent for the general ergodic process, although it is consistent for Markov chains. A weaker trimmed mean type result is proved for the general case, namely, given $\varepsilon > 0$, eventually almost surely all but an ε fraction of the $L_i^n/\log n$ will be within ε of $1/H$. A related Hausdorff dimension conjecture is shown to be false.

1. Introduction. For our purposes a *source*, or stationary ergodic process, is a shift-invariant measure μ on sequences $x = \{x_n\}$ drawn from a finite alphabet A. Let x_m^n denote the finite sequence $x_m, x_{m+1}, \dots, x_n$. If $x = x_1, x_2, \dots$ is an infinite sequence drawn from A, define

$$L_i^n = L_i^n(x) = \min\left\{L \colon x_i^{i+L-1} \ne x_j^{j+L-1}, 1 \le j \le n, j \ne i\right\},$$

that is, $L_i^n(x)$ is the length of the shortest prefix of $x_i, x_{i+1}, \dots,$ which is not a prefix of any other $x_j, x_{j+1}, \dots,$ for $j \le n$. Grassberger (1989) suggests the possibility that the L_i^n can be used to estimate the entropy rate $H = H(\mu)$ of the process and gives some heuristics to suggest that the following holds, almost surely:

$$\lim_n \frac{\sum_{i=1}^n L_i^n(x)}{n \log n} = \frac{1}{H}. \tag{1}$$

The Grassberger prefixes can be viewed as a symmetric form of the celebrated Ziv–Lempel algorithm, which defines prefixes by looking only backwards, that is,

$$\hat{L}^i(x) = \min\left\{L \colon x_i^{i+L-1} \ne x_j^{j+L-1}, 1 \le j < i\right\}.$$

As it is known that $\sum_1^n \hat{L}_i \sim n/(H \log n)$ [Ziv and Lempel (1977)], it is plausible that (1) might be true.

In this paper we show that the conjecture (1) is true for a limited class of processes, that it is not true in general and that a weaker form of the

Received June 1990; revised December 1990.

[1] Partially supported by NSF Grant DMS-87-42630.

AMS 1980 *subject classifications*. Primary 28D20; secondary 60F15.

Key words and phrases. Hausdorff dimension, entropy.

conjecture is true for ergodic processes. These results are summarized in the following theorems.

THEOREM 1. *The limit result* (1) *is true for i.i.d. processes, for mixing Markov processes and for entropy-zero processes.*

THEOREM 2. *There are very weak Bernoulli* (*i.e., finitely determined*) *processes for which the conjecture* (1) *is false, even in probability.*

THEOREM 3. *If* $H > 0$, *then, for every* $\varepsilon > 0$ *and almost every* x, *there is an integer* $N = N(x, \varepsilon)$ *such that, for* $n \geq N$, *all but* εn *of the numbers* $L_i^n(x)/\log n$ *are within* ε *of* $(1/H)$.

The very weak Bernoulli, or finitely determined, processes are precisely the stationary codings of i.i.d. processes [Ornstein (1974)].

As noted by Grassberger, his conjecture is related to a conjecture about Hausdorff dimension. For stationary ergodic processes, this conjecture has the following form: Select n infinite sequences $x(1) = \{x(1)_j\}$, $x(2) = \{x(2)_j\}, \dots,$ $x(n) = \{x(n)_j\}$ independently, using the measure μ that defines the process. Define

$$\tilde{L}_i^n = \min\{L\colon x(i)_1^L \neq x(j)_1^L, 1 \leq j \leq n, j \neq i\}.$$

The Hausdorff dimension conjecture is that, almost surely,

$$\lim_n \frac{\sum_{i=1}^n \tilde{L}_i^n}{n \log n} = \frac{1}{H}. \tag{2}$$

This conjecture is also false; the counterexample we construct in Section 2 also serves as a counterexample for the Hausdorff dimension conjecture. By suitably modifying our methods, Theorem 3 and Theorem 1 can be extended to this case.

2. Positive results. We first prove Theorem 3, then use it to establish Theorem 1. Our proof of Theorem 3 will be based on two lemmas, Lemmas 1 and 3, stated later. The proof of the lower bound result, Lemma 1, uses the covering idea contained in Ornstein and Weiss (1990a); see also the recent results of Ornstein and Weiss (1990b) about partitions of a sequence into distinct blocks. The proof of the upper bound result, Lemma 3, uses a return-time result from Ornstein and Weiss (1990b).

For $i \leq n$, define $W_i^n = W_i^n(x) = x_i^{i+L_i^n(x)-1}$, so that W_i^n is the shortest prefix of $x_i, x_{i+1}, \dots$ which is not a prefix of $x_j, x_{j+1}, \dots,$ for $j \leq n$, $j \neq i$. Of course, W_i^n has length $L(W_i^n) = L_i^n = L_i^n(x)$. We use the following two properties of the W_i^n to prove Theorem 3.

(P1) For fixed n, the W_i^n are distinct.

(P2) Given $i \le n$ and $1 \le L < L_i^n$, there is a $j \le n$, $j \ne i$, such that $x_i^{i+L-1} = x_j^{j+L-1}$.

Our first lemma makes use of property (P1). For $H, \varepsilon > 0$, define

$$\mathscr{B}(x, n, \varepsilon) = \{i \colon L_i^n(x) < (1-\varepsilon)(\log n)/H, 1 \le i \le n\}.$$

Our goal is to show that this collection of "too short" blocks cannot be too large, relative to n.

Lemma 1. *If $\varepsilon > 0$, then for almost every x there is an integer $N = N(x, \varepsilon)$ such that if $n \ge N$ then $|\mathscr{B}(x, n, \varepsilon)| \le \varepsilon n$.*

Proof. To exploit property (P1) we need to gain control over the number of words of a given length. Such control is provided by entropy through the Shannon–McMillan theorem; we need to know a uniform version of a variable-length form of this theorem. Let δ be a positive number to be specified later. The Shannon–McMillan theorem gives an integer m and a collection $\mathscr{C}_m \subset A^m$ such that the following hold:

(a) $\mu(\mathscr{C}_m) \ge 1 - \delta/2$;
(b) $|\mathscr{C}_m| \le 2^{m(H+\delta)}$.

We think of $\mathscr{C}_m$ as the "building blocks" of the process because, by the ergodic theorem, most sufficiently long blocks are mostly made up of these blocks. This fact, together with property (P1), will enable us to count the number of too-short blocks. To do this, we first define what we mean by "mostly made up of." For each $k > m$, define $\mathscr{G}_k$ to be the set

$$\left\{x_1^k \colon \left|\left\{x_i^{i+m-1} \in \mathscr{C}_m, 1 \le i \le k-m+1\right\}\right| > (1-\delta)(k-m+1)\right\},$$

that is, $\mathscr{G}_k$ is the collection of k-blocks all but a δ-fraction of whose m-subblocks come from the collection $\mathscr{C}_m$. The size of the collections $\mathscr{G}_k$ can be controlled by suitable choice of δ, for [as noted in Lemma 1 of Ornstein and Weiss (1990a)] it is easy to see that δ can be chosen, independently of k, so that

$$|\mathscr{G}_k| \le 2^{k(H+\varepsilon^2)}. \tag{3}$$

Fix m and δ so that (a), (b) and (3) hold, for $k > m$. The ergodic theorem and $\mu(\mathscr{C}_m) \ge 1 - \delta/2$ imply that, for almost every sequence x, the initial segment x_1^k will belong to $\mathscr{G}_k$ for all sufficiently large k; hence we can choose K such that the set

$$\tilde{\mathscr{G}}_K = \left\{x \colon x_1^k \in \mathscr{G}_k, k \ge K\right\}$$

has measure greater than $1 - \delta/2$. The ergodic theorem then tells us that, almost surely, $T^i x \in \tilde{\mathscr{G}}_K$ for all but fewer than a limiting δ-fraction of the indices i, where T denotes the shift operator. This fact, in the "almost uniform" version we need, is stated as follows.

LEMMA 2. *For almost every x there is an integer $N(x)$ such that if $n \geq N(x)$, then, for all but δn indices $i < n$, the i-fold shift $T^i x$ belongs to $\tilde{\mathscr{G}}_K$.*

Now we proceed to complete the proof of Lemma 1. For a given sequence x and integer n, an index $i \leq n$ will be called good if $L_i^n \geq K$ and $W_i^n \in \mathscr{G}_{L_i^n}$. Since the prefixes W_i^n are all distinct, there are at most $|A|^K$ indices $i \leq n$ such that $L_i^n \leq K$. Combining this with Lemma 2, we see that if n is large enough and δ is small enough, the set of nongood indices will have cardinality at most $\varepsilon n/2$.

Fix $k \geq K$ and consider the set of good indices i for which $L_i^n = k$. There are at most $2^{k(H+\varepsilon^2)}$ such indices because the W_i^n are distinct members of the collection $\mathscr{G}_k$ which has cardinality at most $2^{k(H+\varepsilon^2)}$. Hence there are at most (constant)$2^{J(H+\varepsilon^2)}$ good indices i for which $K \leq L_i^n < J$. Thus, if n is sufficiently large and ε is sufficiently small, there will be at most $\varepsilon n/2$ good indices i for which $L_i^n \leq (1-\varepsilon)(\log n)/H$. This, combined with the bound on the number of nongood indices of the preceding paragraph, completes the proof of Lemma 1. □

Next we prove an upper bound result by using the fact that, except for the final letter, each W_i^n appears at least twice, which is property (P2). If a too-long block appears twice, then it has too short return-time; recent return-time results [Ornstein and Weiss (1990b)] will show that this can only happen rarely. Fix $0 < \varepsilon < H$ and let

$$\mathscr{U}(x,n,\varepsilon) = \left\{ i \colon L_i^n(x) > 1 + \frac{\log n}{H-\varepsilon}, 1 \leq i \leq n \right\}.$$

LEMMA 3. *For almost every x there is an integer $N = N(x,\varepsilon)$ such that if $n \geq N$, then $|\mathscr{U}(x,n,\varepsilon)| \leq \varepsilon n$.*

PROOF. Define the return-time function

$$R_k(x) = \inf\{m \geq k \colon x_{m+1}^{m+k} = x_1^k\}. \tag{4}$$

Then $\lim_k (\log R_k(x))/k = H$, a.s., [Ornstein and Weiss (1990b)], so there is a K such that if $k \geq K$, then $(\log R_k(x))/k \geq H - \varepsilon$, except for a set B of sequences of measure less than $\varepsilon/3$. Therefore, for almost every x there is an integer $N(x)$ such that if $n \geq N(x)$, then $T^{i-1}x \in B$ for at most $\varepsilon n/2$ indices $i \leq n$.

If $i \in \mathscr{U}(x,n,\varepsilon)$ and $k = L_i^n - 1$, then there is a $j \leq n$, $j \neq i$, such that $x_i^{i+k-1} = x_j^{j+k-1}$, by property (P2). Thus either $R_k(T^{i-1}x) \leq n$

or $R_k(T^{j-1}x) \le n$. Therefore, if $L_i^n > 1 + (\log n)/(H - \varepsilon)$, then either $T^{i-1}x \in B$ or $T^{j-1}x \in B$. This shows that if $n \ge N(x)$, then $|\mathcal{U}(x, n, \varepsilon)| \le 2\varepsilon n/2$, which establishes the lemma.

Theorem 3 follows easily from Lemmas 1 and 3. To establish the stronger result, Theorem 1, for the Markov case, it is enough to show that the return-time $R_k(x)$ can be too short only on a set of exponentially small probability; such results are already known, for example, Arratia and Waterman (1989). Our proof of Lemma 1 can be extended easily to the zero entropy case, hence Theorem 1 holds in that case also. □

3. The counterexample. Let $\Omega = \{0, 1\}^Z$ be the set of doubly infinite binary sequences and let λ denote the measure on Ω defined by coin-tossing, that is, $\lambda([a_1^n]) = 2^{-n}$, where $[a_1^n]$ denotes the cylinder set determined by a_1^n, $a_i \in \{0, 1\}$. Given $\varepsilon > 0$, we construct a binary, ergodic process μ with the following properties.

(A1) There is a measurable, shift-invariant function F: $\Omega \mapsto \Omega$ such that $\mu(A) = \lambda(F^{-1}A)$ for all cylinder sets A.

(A2) There an increasing sequence n_k of positive integers such that if $y = F(x)$, then, with probability greater than $1 - 2^{-k}$, there is an i, $0 < i < n_k - n_k^{3/4}$, such that $y_{i+j} = 0$, $0 \le j \le n_k^{3/4}$.

(A3) $\text{prob}(x_0 \ne F(x)_0) \le \varepsilon$.

The process μ provides a counterexample to (1). Property (A1) says that μ is a stationary coding of an i.i.d. process, hence μ must be very weak Bernoulli. Property (A1) also guarantees positive entropy. Property (A2) guarantees that eventually almost surely there will be infinitely many n_k such that $L_i^{n_k} \ge n_k^{3/4}/2$ for at least $n_k^{3/4}/2$ many indices $i < n_k$, so that $\sum_{i \le n} L_i^n \ge n_k^{3/2}/4$, and hence, (1) fails. Property (A3) implies that we can produce such examples by making arbitrarily small changes in the sample paths of the i.i.d. process λ.

Our construction will be similar to our construction of ergodic processes for which sample paths of length n tend to have matching blocks longer than that predicted by entropy [Shields (1991)]. We shall first show how to construct a stationary coding function F_n so that property (A2) holds for a given large n.

Lemma 4. *Given $\delta > 0$, there is an N_0 such that if $n \ge N_0$, then there is a measurable, shift-invariant function F_n: $\Omega \mapsto \Omega$ such that if $y = F_n(x)$, then the following hold*:

(i) *With probability greater than $1 - \delta$, there is an i, $0 < i < n - n^{3/4}$, such that $y_{i+j} = 0$, $0 \le j \le n^{3/4}$.*

(ii) $\text{prob}(x_0 \ne F_n(x)_0) \le \delta$.

PROOF. The basic coding idea is quite simple: Just partition a sample path x into blocks $x_{jn+1}^{(j+1)n}$ of length n and define $y_{jn+t} = 0$, $1 \leq t \leq n^{3/4}$, $y_{jn+t} = x_{jn+t}$, otherwise. This code is, of course, not stationary. Such a block code can be modified to obtain a stationary code by using the "punctuation" scheme described in Shields and Neuhoff (1977), which is essentially the same as the Rochlin tower coding technique used in Ornstein (1974). Let γ be a positive number to be specified later and choose a cylinder set C such that $0 < \mu(C) < \gamma$. Fix $x \in \Omega$ and define the increasing sequence $m_i = m_i(x)$ by the condition $T^{m_i}x \in C$. The n-block coder is applied to successive n-blocks starting with m_i, until we get within n of m_{i+1}. To make this precise, for each i, determine nonnegative integers q_i, r_i such that $m_{i+1} - m_i = q_i n + r_i$, $0 \leq r_i < n$. Define $F_n(x) = y$, where

$$y_{m_i+jn+t} = \begin{cases} 0, & \text{if } 0 \leq j < q_i,\ 1 \leq t \leq n^{3/4}, \\ x_{m_i+jn+t}, & \text{otherwise.} \end{cases}$$

The average distance between m_{i+1} and m_i is greater than $1/\gamma$, so if γ is small enough (to guarantee that most of the time is spent in the blocks x_{m_i+jn+t}, $0 \leq j < q_i$, $t \leq n$) and n large enough (so that $n^{3/4}/n$ is small), then, with probability greater than $1 - \delta$, there is an $i < n - n^{3/4}$ such that $y_{i+j} = 0$, $0 \leq j \leq n^{3/4}$, and, furthermore, $\text{prob}(x_0 \neq F_n(x)_0) \leq \delta$. This completes the proof of Lemma 4. □

Note that the fact that we started with the i.i.d. measure λ is unimportant. Given any ergodic measure ν, the numbers n and γ can be chosen so that Lemma 4 holds. This is because the set of sequences x for which $\inf m_i = -\infty$ and $\sup m_i = \infty$ has measure 1 with respect to every ergodic measure.

Now we are ready to iterate the preceding construction. To facilitate the discussion, let us define a sequence y_1^n to be *n-good* if there is an $i < n - n^{3/4}$ such that $y_{i+j} = 0$, $0 \leq j \leq n^{3/4}$; otherwise y_1^n will be called *n-bad*. Let $\mathscr{B}_n$ denote the set of n-bad sequences. Let us suppose we have determined $n_1 < n_2 < \cdots < n_k$ such that for $\phi_j = F_{n_j}$, $j \leq k$, and $\mu_j = \mu_{j-1}\phi_j^{-1}$, $y^{(j)} = \phi_j(y^{(j-1)})$, $j \leq k$, where $\mu_0 = \lambda$, $y^{(0)} = x$, the following hold:

$$\text{(C1)} \qquad \lambda\{x : y_0^{(k)} \neq y_0^{(k-1)}\} \leq \frac{\varepsilon}{2^{k+1}}.$$

$$\text{(C2)} \quad \mu_j(\mathscr{B}_{n_i}) \leq \frac{1}{2^{i+1}} + \frac{1}{2^{i+2}} + \cdots + \frac{1}{2^{i+(j-i)+1}}, \qquad i \leq j \leq k.$$

We then apply Lemma 4 to choose $n_{k+1} > n_k$ and $\phi_{k+1} = F_{n_{k+1}}$ to obtain:

1. $\lambda\{x: y_0^{(k+1)} \neq y_0^{(k)}\} < \delta$,
2. $\mu_{k+1}(\mathscr{B}_{n_{k+1}}) < \delta$,

where $\mu_{k+1} = \mu_k\phi_{k+1}^{-1}$, $y^{(k+1)} = \phi_{k+1}(y^{(k)})$. If δ is small enough then so few changes are made by ϕ_{k+1} that we can guarantee that (C1) and (C2) hold with k replaced by $k + 1$.

Now we pass to the limit, defining $F(x) = \lim_k(\phi_k(\cdots(\phi_1(x))\cdots))$. This definition makes sense for condition (C1) guarantees that, almost surely, each coordinate is changed only a finite number of times. Conditions (C1) and (C2) guarantee that the limit process has the desired properties (A1)–(A3). This completes the construction of our counterexample to the conjecture (1).

REMARK. The same example provides a counterexample to the Hausdorff dimension conjecture, (2). If $n = n_k$, then most of a sample path will be covered by nonoverlapping blocks of length n such that the initial $n^{3/4}$ of each such block are 0's. If n independent paths are selected and n is large enough, then it can be shown that, with probability bounded away from 0, at least $n^{3/4}/4$ will have to start in the first half of such an initial block of 0's. This is because it is (approximately) like the problem of throwing n balls into $2n^{1/4}$ boxes. The expected number of balls in the first box is $n^{3/4}/2$ and the variance is $\sim$ (constant)$n^{3/4}$.

Acknowledgments. I wish to thank Imre Csiszár, Kati Marton and Don Ornstein for useful conversations about the ideas in this paper.

REFERENCES

ARRATIA, R. and WATERMAN, M. S. (1989). The Erdös–Rényi strong law for pattern matching with a given proportion of mismatches. *Ann. Probab.* **17** 1152–1169.

GRASSBERGER, P. (1989). Estimating the information content of symbol sequences and efficient codes. *IEEE Trans. Inform. Theory* **IT-35** 669–675.

ORNSTEIN, D. (1974). *Ergodic Theory, Randomness, and Dynamical Systems*. Yale Univ. Press.

ORNSTEIN, D. and WEISS, B. (1990a). How sampling reveals a process. *Ann. Probab.* **18** 905–930.

ORNSTEIN, D. and WEISS, B. (1990b). Entropy and data compression schemes. Unpublished manuscript.

SHIELDS, P. (1992). String matching: The ergodic case. *Ann. Probab.* **20**. To appear.

SHIELDS, P. and NEUHOFF, D. (1977). Block and sliding-block source coding. *IEEE Trans. Inform. Theory* **IT-23** 211–215.

ZIV, J. and LEMPEL, A. (1977). A universal algorithm for sequential data compression. *IEEE Trans. Inform. Theory* **IT-23** 337–343.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF TOLEDO
TOLEDO, OHIO 43606