The Annals of Probability
1992, Vol. 20, No. 3, 1199–1203

STRING MATCHING: THE ERGODIC CASE[1]

BY PAUL C. SHIELDS

University of Toledo and Eötvös Loránd University

Of interest in DNA analysis is the length $L(x_1^n)$ of the longest sequence that appears twice in a sequence x_1^n of length n. Karlin and Ghandour and Arratia and Waterman have shown that if the sequence is a sample path from an i.i.d. or Markov process, then $L(x_1^n) = O(\log n)$. In this paper, examples of ergodic processes are constructed for which the asymptotic growth rate is infinitely often as large as $\lambda(n)$, where $\lambda(n)$ is subject only to the condition that it be $o(n)$.

Let A denote a finite set and let x_m^n denote the sequence $x_m, x_{m+1}, \ldots, x_n$, where each $x_i \in A$. Define $L(x_1^n)$ to be the length of the longest block which appears at least twice in x_1^n, that is,

$$L(x_1^n) = \max\{k: \exists\, s, t; 0 \le s < t \le n - k, \text{ for which } x_{s+1}^{s+k} = x_{t+1}^{t+k}\}.$$

The asymptotic behavior of $L(x_1^n)$, along sample paths from a stationary, finite-alphabet ergodic process, is of interest in DNA modeling. At first glance it would seem that this asymptotic rate should have the form $(\log n)/H$ where H is the entropy-rate of the process, or at least one would hope that $L(x_1^n) = O(\log n)$. Results of this type have been established by Karlin and Ghandour (1985) and by Arratia and Waterman (1989) for i.i.d. processes and Markov processes. It is not difficult to see, using the ideas in Ornstein and Weiss (1990), that for an arbitrary ergodic process, $(\log n)/H$ is an asymptotic lower bound in the almost sure sense, that is,

$$\liminf_{n\to\infty} \frac{L(x_1^n)}{(\log n)/H} \ge 1, \quad \text{almost surely.}$$

The goal of this paper is to show that, in general, *this entropy bound is not tight for the general class of ergodic processes*. We show that for any ergodic process $\{X_n\}$ and function $\lambda(n)$ which is $o(n)$, there is a stationary coding, $\{Y_n\}$ of $\{X_n\}$, such that

$$\limsup_{n\to\infty} L(Y_1^n)/\lambda(n) \ge 1, \quad \text{almost surely.}$$

Our construction is quite simple, although it uses a technique that may not be well known outside ergodic theory.

For our purposes, a stationary process is a shift-invariant measure μ on the space A^Z of doubly infinite sequences drawn from A; A is called the alphabet.

Received April 1990; revised December 1990.
[1]Partially supported by NSF Grant DMS-87-42630.
AMS 1980 *subject classifications*. Primary 92A10; secondary 60F15.
Key words and phrases. String matching, entropy.

A stationary coding is defined by a shift-invariant measurable function, called the *coding function*, F: $A^Z \mapsto B^Z$, where B is a finite set. Here measurable means measurable with respect to the Borel sets in the two product spaces, and shift-invariant means that $F(T_A x) = T_B F(x)$, for μ-almost all x, where T_A and T_B denote the respective shifts on A^Z and B^Z. The coding function F transports the given stationary measure μ to a stationary measure ν on B^Z which we denote by $\tilde{F}(\mu)$. Thus if μ is the measure corresponding to the stationary sequence $\{X_n\}_{n=-\infty}^{n=\infty}$ of random variables, then ν is the measure corresponding to the sequence $\{Y_n\}_{n=-\infty}^{n=\infty}$ defined by $Y_m = F(\{X_{m+n}\}_{n=-\infty}^{n=\infty})$.

Our basic result is the following.

THEOREM 1. *Let μ be an ergodic process with alphabet A and suppose $\lim_n \lambda(n)/n = 0$. There is a stationary coding function F; $A^Z \mapsto A^Z$ such that if $y = F(x)$, then*

$$\limsup_n L(y_1^n)/\lambda(n) \geq 1, \quad \textit{almost surely}.$$

Our coding function will be obtained by iterating the construction given in Lemma 1 below, a construction that shows how to make $L(y_1^m)$ large in probability for a fixed m by a code that makes only a small percentage change in each sequence x. To quantify changes in infinite sequences, we use the limiting average Hamming distance defined by

$$\bar{d}(x,y) = \limsup_n \frac{1}{n} \sum_{i=1}^{n} d(x_i, y_i), \qquad x, y \in A^Z,$$

where $d(a,b) = 1$ if $a \neq b$, $d(a,b) = 0$ if $a = b$.

LEMMA 1. *Let μ be an ergodic process with alphabet A, let m and k be positive integers such that $k < m/4$ and let ε be a positive number. There is a coding function $F = F(\mu, k, m, \varepsilon)$ such that the encoded process $\nu = \tilde{F}(\mu)$ satisfies*:

(a) $\nu(\{y_1^m: L(y_1^m) \geq k\}) \geq 1 - \varepsilon$.

(b) $\mu(\{x: F(x)_0 \neq x_0\}) \leq 2k/m$.

PROOF. Let g be the greatest integer in $m/2$. The function $F(x)$ will be defined by partitioning x into nonoverlapping blocks, each of length no more than g, such that most of the blocks will have length exactly equal to g. Then k changes will be made in each g-block to guarantee that each such block will contain matching strings of length k. We take for our g-block coder the function that simply replaces the last k symbols in a g-block by the first k symbols, namely the function $\phi_{k,m}$: $A^m \mapsto A^m$ defined by

$$\phi_{k,m}(a_1^g) = b_1^g, \qquad b_i = a_i, \qquad 1 \leq i \leq g - k, \qquad b_{g-k+i} = a_i, \qquad 1 \leq i \leq k.$$

Note that our definition insures that if $b_1^g = \phi_{k,m}(a_1^g)$, then $L(b_1^g) \geq k$ and $\sum_1^g d(a_i, b_i) \leq gk$.

The partitioning of a given sequence x is done 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 δ be a positive number to be specified later and choose a cylinder set C such that $0 < \mu(C) < \delta$. Fix $x \in A^Z$ and define the increasing sequence $n_i = n_i(x)$ by the condition $T^{n_i}x \in C$. (Note that the set $\tilde{X}$ of sequences x for which $\inf n_i = -\infty$ and $\sup n_i = \infty$ has measure 1 with respect to every ergodic measure, so the coding function can be defined arbitrarily outside $\tilde{X}$.) The g-block coder is applied to successive g-blocks starting with n_i, until we get within g of n_{i+1}. To make this precise, for each i determine nonnegative integers q_i, r_i such that $n_{i+1} - n_i = q_i g + r_i$, $0 \le r_i < g$. Define $F(x) = y$, where

$$y_{n_i+jg}^{n_i+(j+1)g-1} = \phi_{k,m}\left(x_{n_i+jg}^{n_i+(j+1)g-1}\right), \qquad 0 \le j < q_i,$$

$$y_{n_i+q_ig+j} = x_{n_i+q_ig+j}, \qquad 0 \le j < r_i.$$

This defines the coding function on a set of full measure. At most k changes are made in any block of length $g \le m/2$, so that condition (b) certainly holds. The proof will be complete if we can show that δ can be chosen so that condition (a) holds. If a block of length m in x contains one of the blocks $x_{n_i+jg}^{n_i+(j+1)g-1}$, $0 \le j < q_i$, then the definition of ϕ guarantees matching strings of length at least k. If a block of length m contains no such block, then it must contain at least one of the $x_{n_i+q_ig+j}$, $0 \le j < r_i$; by the ergodic theorem the probability of this happening can be made arbitrarily small by choosing δ small enough. This completes the proof of the lemma. □

PROOF OF THEOREM 1. Let $\{\varepsilon_i\}$ be a sequence of positive numbers such that $\Sigma_i \varepsilon_i < \infty$ and let $\{n_i\}$ be an increasing sequence of positive integers such that $\lambda(n_i) < n_i$. Put $\mu_1 = \mu$ and use the lemma to define inductively

$$F_i = F_i(\mu_i, \lambda(n_i), n_i, \varepsilon_i); \ \mu_{i+1} = \tilde{F}_i(\mu_i).$$

Condition (b) of the lemma guarantees that

$$\mu_{i+1}\left(\left\{x: F_i\left(F_{i-1}\left(\cdots\left(F_1(x)\right)\cdots\right)\right)_0 \ne x_0\right\}\right) < \sum_{j=1}^{i} 2\lambda(n_j)/n_j.$$

Thus if we assume

$$\sum_i 2\lambda(n_i)/n_i < \infty,$$

then we are guaranteed that almost surely each coordinate will get changed only a finite number of times and hence the sequence of codes will converge to a limit code F.

Now we want to guarantee that for the limit process $\nu = \tilde{F}(\mu)$, the probability

$$\nu\left(\left\{x_1^{n_i}: L\left(x_1^{n_i}\right) < \lambda(n_i)\right\}\right)$$

is summable in i, so that almost surely $L(x_1^{n_i})$ can be less than $\lambda(n_i)$ only finitely often. We are already guaranteed by condition (a) of the lemma that at the ith stage,

$$\mu_i(\{x_1^{n_i}: L(x_1^{n_i}) < \lambda(n_i)\}) < \varepsilon_i$$

and need only make sure that the subsequent codes do not have large effects on this probability. This can be accomplished by selecting the n_i a bit more carefully, namely, we choose n_{i+1} so much larger than n_i that the density of changes, $2\lambda(n_{i+1})/n_{i+1}$, is so small that the following holds:

$$\mu_i(\{x_1^{n_j}: L(x_1^{n_j}) < \lambda(n_j)\}) \le \mu_{i-1}(\{x_1^{n_j}: L(x_1^{n_j}) < \lambda(n_j)\}) + \frac{\varepsilon_j}{2^{i-j}}, \qquad j < i.$$

This will guarantee that for the limit measure ν the following holds for all j:

$$\nu(\{x_1^{n_j}: L(x_1^{n_j}) < \lambda(n_j)\}) \le 2\varepsilon_j.$$

Thus, since we assumed that $\sum \varepsilon_i < \infty$, the theorem is established. □

Note that the Ornstein isomorphism theory [see Ornstein (1974)] guarantees if the original process μ is i.i.d., then the limit process will be a Bernoulli process, that is, it will be isomorphic to some i.i.d. process; in particular, the limit process will have very strong mixing properties. Note also that our proof actually shows that codes can be obtained that satisfy the theorem and that make as small a density of changes in the original process as we please.

REMARK. The referees raised the following two questions.

QUESTION 1. "What is going on here?" That is, what conditions on a stationary process are necessary and sufficient for $\log n$ type laws to hold for string matching?

QUESTION 2. Does every ergodic process have an isomorphic image with arbitrary string matching asymptotics, that is, the property given in Theorem 1?

The construction used in this paper does not shed much light on Question 1 other than the negative result that even very strong mixing properties, such as the very weak Bernoulli (VWB) property, are not enough. The author has recently constructed examples showing that conditions stronger than VWB are needed to get positive results about several other entropy related properties, such as prefix generation, waiting times, exponential rates of type II error probabilities and redundancy rates. [The prefix result is in Shields (1992); the other results are more recent.] These and our string matching results raise a whole host of questions about what is going on, including the question of necessary and sufficient conditions in each case and the more general question of whether any of these results are related.

We suspect that there are no nice necessary and sufficient conditions for any of these problems. The more interesting question is to determine a class of processes which is big enough to include the processes of interest in a large class of applications, such as DNA modeling, universal coding, hypothesis testing and engineering design, but small enough that nice entropy properties will hold.

An isomorphism is a stationary coding for which the coding function F is invertible. The answer to Question 2 is unknown in general, but a positive answer can be given for the isomorphism class of an i.i.d. process. This is because there is a VWB process of any given entropy with arbitrary string matching asymptotics, for it is easy to modify the proof of Theorem 1 to obtain a homomorphic image of any given smaller entropy. Since VWB processes of the same entropy are isomorphic, this indeed shows that the isomorphism class of an i.i.d. process contains processes with arbitrary string matching asymptotics.

Acknowledgment. The author wishes to thank Gábor Tusnády of the Mathematics Institutes of the Hungarian Academy of Sciences for bringing this problem to his attention.

REFERENCES

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

KARLIN, S. and GHANDOUR, G. (1985). Comparative statistics for DNA and protein sequences—single sequence analysis. *Proc. Nat. Acad. Sci. U.S.A.* **82** 5800–5804.

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

ORNSTEIN, D. S. and WEISS, B. (1990). Entropy and data compression schemes. *IEEE Trans. Inform. Theory*. To appear.

SHIELDS, P. (1992). Entropy and prefixes. *Ann. Probab.* **20** 403–409.

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

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF TOLEDO
TOLEDO, OHIO 43606