AN ENTROPY ESTIMATOR FOR A CLASS OF INFINITE ALPHABET PROCESSES

Anthony N. Quas

Department of Pure Mathematics and Mathematical Statistics,
University of Cambridge

June 1997

Abstract. Motivated by recent work by Kontoyiannis and Suhov, and by Shields, we present an entropy estimator which works for a class of ergodic finite entropy infinite symbol processes for which the entropy of the time-zero partition is finite, and which satisfy a 'Doeblin condition'. The results are then extended to random fields indexed by $\mathbb{Z}^d$.

0. Introduction

In this paper, we consider the problem of estimation of the entropy of a sequence of data, by only looking at a finite portion (the reader is referred to [12] for definitions etc). Since the entropy is a measure of the information content of a sequence of data, there are practical reasons for wanting to measure entropy. There are a number of techniques available for estimation of entropy, but the drawbacks in general are that it is hard to prove that the estimators converge to the true value of the entropy, or that one requires an extremely large amount of data to compute an estimate. In general, if we consider the sequence of data to be random variables indexed by $\mathbb{Z}$ taking values in a finite or countable set S, the entropy is much harder to compute if the random variables have long-range dependence. The method presented below is proven to work, is able to deal with systems having long-range dependence (although for any estimator, the longer the range, the more terms will be needed to get a reliable estimate) and can work with a relatively small amount of data.

Let $(X_n)_{n\in\mathbb{Z}}$ be a stationary ergodic stochastic process taking values in a countable set S. Let μ be the induced probability measure on $S^{\mathbb{Z}}$. Throughout this paper, we will need to make the assumption that the time-zero partition $\mathcal{P}$ into sets of the form $[s] = \{x\colon x_0 = s\}$ has finite entropy (i.e. $H(\mathcal{P}) \equiv \sum_{s\in S} -\mu([s]) \log \mu([s]) < \infty$). Under this condition, it follows that the entropy of the process is finite (See [12]). Write h for the entropy of the process.

1991 *Mathematics Subject Classification.* 28D20 28D05 60G10 94A24.

Key words and phrases. Entropy Estimator, Prefixes.

Typeset by $\mathcal{A}\mathcal{M}\mathcal{S}$-TEX

Following the Ziv–Lempel encoding algorithm, Grassberger ([4]) introduced the notion, which we will need, of a prefix. Given a sequence x in $S^{\mathbb{Z}}$, the word consisting of the terms of the sequence $x_m x_{m+1} \dots x_n$ will be denoted by x_m^n. The prefix at scale n of a sequence x starting at i is the shortest word x_i^{i+k-1} which is not equal to x_j^{j+k-1} for $j \neq i$ and $0 \leqslant j < n$ (that is it is the shortest segment of x starting from the ith term which is not repeated starting from some distinct term before the nth). The shortest prefix at scale n starting at i is denoted by $W_i^n(x)$. $L_i^n(x)$ is the length of $W_i^n(x)$. Note that as n increases, the length of the shortest prefix increases as a longer sequence is needed to ensure that it is not repeated. In this paper, we will show that the growth of the prefix length is typically of the order of $\log n$, with the constant related to the entropy of the process.

The other condition which we will need is that there exists an $\alpha < 1$ and an $r \geqslant 1$ such that

$$\mathbb{P}(X_0 = s | X_{-r} = x_{-r},\ X_{-r-1} = x_{-r-1}, \dots) \leqslant \alpha \tag{1}$$

for all s and almost all x. We will refer to this as a Doeblin condition, although Doeblin conditions are usually given for finite state processes and the probability is bounded below rather than above. Since the probabilities sum to 1, the usual Doeblin condition on a finite state space implies the condition (1) which we will use.

We are now ready to state the theorem.

Theorem 1. *Suppose $(X_n)_{n\in\mathbb{Z}}$ is a stationary ergodic process taking values in a countable set S with entropy h and induced measure μ on $S^{\mathbb{Z}}$. Suppose further that the process satisfies the Doeblin condition (1), and that the time-zero partition has finite entropy. Then for μ-almost every $x \in S^{\mathbb{Z}}$, we have*

$$h = \lim_{n\to\infty} \frac{n \log n}{\sum_{i=0}^{n-1} L_i^n(x)}. \tag{2}$$

This theorem has already been shown in the case where S is finite by Kontoyiannis and Suhov ([6]), using a result of Shields ([13]), which in turn was based on a result by Ornstein and Weiss ([9]). The result (lemma 1) of the paper of Ornstein and Weiss on which Shields' paper is based depends crucially on the finiteness of the set S. The main work here is to show how using a direct argument one can avoid the need to use Ornstein and Weiss' lemma. Shields' paper contained a second result which is used in [6], but this result seems to have a small gap in the proof. Since we also need to use Shields' result, we take the opportunity to supply a completion of Shields' proof.

Shields also showed that even for finite state very weak Bernoulli processes, (2) does not hold in general. Kontoyiannis and Suhov ([6]) were the first to show the usefulness of Doeblin conditions in this area, providing a natural condition which gives sufficient regularity of the quantities $L_i^n(x)$. This partly answers a question asked in [14], of what is the right class of processes to consider to get good asymptotic results.

In §2, we describe generalizations of these results to the case of random fields. We apply the theory in §3, leading to a conjecture of a better entropy estimation technique. In §4, we make some preliminary observations about this technique and in §5, we conclude with some problems which seem interesting.

1. Proof of Theorem 1

Proof of Theorem 1. Fix $\epsilon > 0$. The proof is in three parts. These are as follows:

(1) For almost every $x \in S^{\mathbb{Z}}$, we have

$$\limsup_{n\to\infty} \frac{1}{n} \left| \left\{ i < n : L_i^n(x) \leqslant \frac{\log n}{h}(1-\epsilon) \right\} \right| = 0.$$

(2) For almost every $x \in S^{\mathbb{Z}}$, we have

$$\limsup_{n\to\infty} \frac{1}{n} \left| \left\{ i < n : L_i^n(x) \geqslant \frac{\log n}{h}(1+\epsilon) \right\} \right| = 0.$$

(3) There exists a $c > 0$ such that for almost every x, we have

$$\limsup_{n\to\infty} \max_{0\leqslant i<n} \frac{L_i^n(x)}{\log n} \leqslant c.$$

Given these statements, it follows that for almost all x we have

$$\limsup_{n\to\infty} \left| \frac{1}{n} \sum_{i=0}^{n-1} \frac{L_i^n(x)}{\log n} - \frac{1}{h} \right| \leqslant \frac{\epsilon}{h}.$$

Since ϵ is arbitrary, the theorem follows. We demonstrate (1), (2) and (3) in Lemmas 2, 3 and 4 respectively.

Lemma 2.
If the conditions of Theorem 1 hold, then for almost every $x \in S^{\mathbb{Z}}$, we have

$$\limsup_{n\to\infty} \frac{1}{n} \left| \left\{ i < n : L_i^n(x) \leqslant \frac{\log n}{h}(1-\epsilon) \right\} \right| = 0.$$

Proof. Define

$$f(x) = \limsup_{n\to\infty} \frac{1}{n} \left| \left\{ i < n : L_i^n(x) \leqslant \frac{\log n}{h}(1-\epsilon) \right\} \right|.$$

We may check that this is an invariant function, so by ergodicity of the measure μ, it suffices to check that for each $\delta > 0$, we have that $f(x) \leqslant \delta$ on a set of positive measure. It will then follow that $f(x) \leqslant \delta$ on a set of measure 1. Let $\mathcal{P}_n$ be $\bigvee_{i=0}^{n-1} T^{-i}\mathcal{P}$. This is the partition of $S^{\mathbb{Z}}$ according to the first n symbols. For $x \in S^{\mathbb{Z}}$, let $\mathcal{P}_n(x)$ denote the element of $\mathcal{P}_n$ containing x. By the Shannon–Macmillan–Breiman theorem (see [12]), we have $\frac{1}{n}\log\mu(\mathcal{P}_n(x)) \to -h$ almost everywhere. Set

$$S_N = \left\{ x \colon \left| \frac{1}{n} \log \mu\big(\mathcal{P}_n(x)\big) + h \right| < h\epsilon, \, \forall n > N \right\}.$$

Then we see that S_N is an increasing sequence of sets and $\bigcup_N S_N$ has measure 1. It follows that there exists an N_1 such that $\mu(S_{N_1}) > 1 - \frac{\delta}{2}$. Then if $x \in S_{N_1}$ and $n > N_1$ then we have $\mu(\mathcal{P}_n(x)) \geqslant \exp(-nh(1+\epsilon))$.

Now for almost all x, by Birkhoff's theorem, there exists an $n(x)$ such that

$$n > n(x) \Rightarrow \frac{1}{n}\Big|\{i < n \colon T^i x \in S_{N_1}\}\Big| \geqslant 1 - \frac{\delta}{2}.$$

It follows that there exists an N_2 such that $n(x) \leqslant N_2$ on a set A of positive measure. Suppose now that $x \in A$, $n > N_2$ and $\frac{\log n}{2h} > N_1$. Suppose further that $L_i^n(x) < \frac{\log n}{h}(1-\epsilon)$. Then there are two possibilities:

(1) $T^i(x) \notin S_{N_1}$. This accounts for a proportion less than $\frac{\delta}{2}$ of is with $i < n$. by definition of A.

(2) $T^i(x) \in S_{N_1}$. This implies

$$\mu\left(\mathcal{P}_{\lfloor \frac{\log n}{h}(1-\epsilon)\rfloor}(T^i x)\right) \geqslant \exp\left(\left\lfloor \frac{\log n}{h}(1-\epsilon)\right\rfloor(-h(1+\epsilon))\right) \geqslant n^{-1+\epsilon^2}.$$

Since the $W_i^n(x)$ are distinct, the sets $[W_i^n(x)]$ (i.e. the sets of those points in $S^{\mathbb{Z}}$ which agree with the prefix of x starting at i for the first $L_i^n(x)$ terms) are disjoint. In the case in hand, we have $L_i^n(x) \leqslant \lfloor \frac{\log n}{h}(1-\epsilon)\rfloor$ so $[W_i^n(x)]$ contains $\mathcal{P}_{\lfloor \frac{\log n}{h}(1-\epsilon)\rfloor}(T^i x)$. In particular, we see that $[W_i^n(x)]$ has measure at least $n^{-1+\epsilon^2}$. It follows that the number of i satisfying this condition is at most $n^{1-\epsilon^2}$. Now picking n such that $n^{-\epsilon^2} < \frac{\delta}{2}$, we see that the proportion of i falling into case (2) is at most $\frac{\delta}{2}$.

It follows that the proportion of i with $L_i^n(x) < \frac{\log n}{h}(1-\epsilon)$ is at most δ. This proves that $f(x) \leqslant \delta$ for $x \in A$ as required. □

Lemma 3.

For almost every $x \in S^{\mathbb{Z}}$, we have

$$\limsup_{n\to\infty} \frac{1}{n}\left|\left\{i < n : L_i^n(x) \geqslant \frac{\log n}{h}(1+\epsilon)\right\}\right| = 0. \tag{3}$$

Proof. It was claimed in [13] that (3) holds if S is a finite set, but the proof seems to be incomplete. We present here a completion of Shields' proof and then show how to apply it in the situation of this paper, where there are infinitely many symbols. To describe the problem, we need to introduce two quantities, $S_k(x)$ and $R_k(x)$. These are defined as follows:

$$R_k(x) = \inf\{m \geqslant k \colon x_m^{m+k-1} = x_0^{k-1}\}$$
$$S_k(x) = \inf\{m \geqslant 1 \colon x_m^{m+k-1} = x_0^{k-1}\}$$

In lemma 3 of [13], part of [10] is quoted, showing that for almost every x, $\log R_k(x)$ $/k \to h$ as $k \to \infty$. This concerns the time before the first non-overlapping recurrence of the block of k symbols. Shields' lemma makes no distinction between overlapping recurrence and non-overlapping recurrence, so in fact the result needed for lemma 3 of [13] to hold is that for almost every x,

$$\lim_{k\to\infty} \frac{\log S_k(x)}{k} = h.$$

We now prove this, given that $\log R_k(x)/k \to h$ for almost all x.

To prove this, it is clearly necessary and sufficient to show that for almost every x, the number of overlapping recurrences is finite. Let

$$A = \{x\colon x_0^{k-1} = x_k^{2k-1} \text{ for infinitely many } k\}.$$

We start off by showing that $\mu(A) = 0$. Pick ϵ satisfying $0 < \epsilon < \frac{1}{3}$. As before set

$$S_N = \left\{x\colon \left|\frac{1}{n}\log\mu(\mathcal{P}_n(x)) + h\right| < h\epsilon,\ \forall n \geqslant N\right\}.$$

Then, as before, by the Shannon–Macmillan–Breiman theorem, we have $\mu(S_N) \to 1$ as $N \to \infty$. Next, let $B_m = \{x : x_0^{m-1} = x_m^{2m-1}\}$. Let $m \geqslant N$ and suppose $x \in B_m \cap S_N$. Then we see that $\mu(\mathcal{P}_m(x)) > \exp(-hm(1+\epsilon))$ and $\mu(\mathcal{P}_{2m}(x)) < \exp(-2hm(1-\epsilon))$. Let W be any word of length m. Write $[W]$ for those points whose first m terms are those of W and $[WW]$ for those points whose first m terms and second block of m terms are both W. Then we see from the above argument that if $[WW] \cap S_N$ is non-empty, then $\mu([WW]) < \mu([W])\exp(-hm(1-3\epsilon))$. In particular, for any word W, we have $\mu([WW] \cap S_N) < \mu([W])\exp(-hm(1-3\epsilon))$. Now summing over words W of length m, we see that $\mu(B_m \cap S_N) \leqslant \exp(-hm(1-3\epsilon))$. We now apply a Borel–Cantelli argument. Set $C_l = \bigcup_{m \geqslant l} B_m$. Then the above shows that $\mu(C_l \cap S_l) < \exp(-hl(1-3\epsilon))/(1-\exp(-h(1-3\epsilon)))$. It then follows that $\mu(C_l) \leqslant \exp(-hl(1-3\epsilon))/(1-\exp(-h(1-3\epsilon))) + \mu({S_l}^c)$. Since this expression converges to 0 and A is $\bigcap_{l=1}^{\infty} C_l$, this implies A has measure 0 as required.

To complete the proof, note that any periodic points lie in A. Now take $x \in A^c$. Then there are at most finitely many k such that $x_0^{k-1} = x_k^{2k-1}$. Let these k be $k_1, \ldots, k_r$. Since $x \in A^c$, x is aperiodic. It follows that there exist $i_1, \ldots, i_r$ such that $x_{i_j} \neq x_{k_j+i_j}$ for $j = 1, \ldots, r$. Next let $t = \max\{i_j : 1 \leqslant j \leqslant r\}$. Then suppose $s > t$ and $x_0^{s-1} = x_k^{k+s-1}$ for some $k < s$. Then $x_0^{k-1} = x_k^{2k-1}$, so it follows that $k = k_j$ for some $j \leqslant r$. But then $x_{i_j} \neq x_{k+i_j}$ so we see that $x_0^{s-1} \neq x_k^{k+s-1}$ which contradicts our assumption.

In particular, it follows that for $x \in A^c$, there are at most finitely many n for which there exists $k < n$ with $x_0^{n-1} = x_k^{k+n-1}$. This completes the proof of Shields' result. We now adapt Shields' result to our purposes.

We will assume without loss of generality that the countable set S is $\mathbb{Z}^+$, the positive integers. Let $({X^{(m)}}_n)_{n\in\mathbb{Z}}$ denote the stationary process defined by ${X^{(m)}}_n = \min(X_n, m)$. Then $({X^{(m)}}_n)$ gives rise to a shift-invariant measure $\mu^{(m)}$ on $\{0, \ldots, m\}^{\mathbb{Z}}$. This measure is in fact the projection of the measure μ under the map $\pi^{(m)} : {\mathbb{Z}^+}^{\mathbb{Z}} \to \{0, \ldots, m\}^{\mathbb{Z}}$; $\pi^{(m)}(x)_n = \min(x_n, m)$. Write $h^{(m)}$ for the entropy of the measure $\mu^{(m)}$. It is a standard property of entropy that $h^{(m)} \to h$ as $m \to \infty$ (See [12] proposition 5.2.11). Pick m such that $\frac{h^{(m)}}{h} > \frac{1+\frac{\epsilon}{2}}{1+\epsilon}$. Then by Shields' result, for almost every $y \in \{0, \ldots, m\}^{\mathbb{Z}}$ with respect to $\mu^{(m)}$, we have

$$\limsup_{n\to\infty} \frac{1}{n}\left|\left\{i < n\colon L_i^n(y) \geqslant \frac{\log n}{h^{(m)}}\left(1+\frac{\epsilon}{2}\right)\right\}\right| = 0.$$

From this it follows that for almost every $x \in S^{\mathbb{Z}}$ with respect to μ, we have

$$\limsup_{n\to\infty} \frac{1}{n}\left|\left\{i < n\colon L_i^n\left(\pi^{(m)}(x)\right) \geqslant \frac{\log n}{h^{(m)}}\left(1+\frac{\epsilon}{2}\right)\right\}\right| = 0.$$

Now using the facts that $L_i^n\left(\pi^{(m)}(x)\right) \geqslant L_i^n(x)$ and $\frac{\log n}{h^{(m)}}\left(1+\frac{\epsilon}{2}\right) \leqslant \frac{\log n}{h}(1+\epsilon)$, we see that for almost every $x \in S^{\mathbb{Z}}$,

$$\limsup_{n\to\infty} \frac{1}{n}\left|\left\{i<n\colon L_i^n(x) > \frac{\log n}{h}(1+\epsilon)\right\}\right| = 0$$

as required. □

Lemma 4. *There exists a $c>0$ such that for almost every x, we have*

$$\limsup_{n\to\infty} \max_{0\leqslant i<n} \frac{L_i^n(x)}{\log n} \leqslant c.$$

Proof. The proof of this Lemma is identical to the proof given in [6]. It is reproduced here for the convenience of the reader. We start by estimating $\mathbb{P}(L_i^n(x) > c\log n)$. Fix n and let $l=\lfloor c\log n\rfloor$. Then by definition of L_i^n, $L_i^n(x)>l$ if and only if there is a $j<n$ distinct from i such that $x_i^{i+l-1}=x_j^{j+l-1}$. For a fixed j, $\mathbb{P}(X_i^{i+l-1}=X_j^{j+l-1})$ may be estimated as follows:

$$\begin{aligned}\mathbb{P}(X_i^{i+l-1}=X_j^{j+l-1}) \leqslant &\mathbb{P}(X_j=X_i)\cdot\mathbb{P}(X_{j+r}=X_{i+r}|X_j=X_i)\cdot\ldots\cdot\\ &\mathbb{P}(X_{j+sr}=X_{i+sr}|X_{j+(s-1)r}=X_{i+(s-1)r};\ldots;X_j=X_i),\end{aligned}$$

where $s=\lceil l/r\rceil-1$. But by (1), each of the probabilities in the product is bounded above by α. It follows that $\mathbb{P}(X_i^{i+l-1}=X_j^{j+l-1})\leqslant\alpha^{\lceil l/r\rceil}\leqslant\alpha^{l/r}$. Summing over j, it follows that $\mathbb{P}(L_i^n(X)>c\log n)\leqslant n\alpha^{(c\log n-1)/r}$. From this, we see that $\mathbb{P}(\max_{0\leqslant i<n}L_i^n(X)>c\log n)\leqslant n^2\alpha^{c\log n/r-1}=n^{2+(c\log\alpha)/r}/\alpha$. Now choosing $c>-3r/\log\alpha$, we see that the sequence is summable, and hence by the Borel–Cantelli Lemma, for almost every sequence x, we have

$$\limsup_{n\to\infty} \max_{0\leqslant i<n} \frac{L_i^n(x)}{\log n} \leqslant c.$$

□

This completes the proof of the theorem. □

2. Generalizations to Higher Dimensions

The theorem has an analogue which works in higher dimensional situations: Suppose $(X_v)_{v\in\mathbb{Z}^d}\colon\Omega\to S$ is a stationary random field taking values in a finite set S (note that in this section the finiteness of S is needed as the version of the Shannon–Macmillan–Breiman theorem which we will use is slightly more restrictive. It may be possible to remove this restriction - see the section on possible extensions.) There is a naturally associated $\mathbb{Z}^d$ action on $S^{\mathbb{Z}^d}$ defined by $(T_u(x))_v=x_{u+v}$. Define also a map $\pi:\Omega\to S^{\mathbb{Z}^d}$; $(\pi(\omega))_v=X_v(\omega)$. Then π pushes forward $\mathbb{P}$, the distribution on Ω to a measure μ invariant under the $\mathbb{Z}^d$ action. There is a corresponding definition of ergodicity for $\mathbb{Z}^d$ actions and we will need to assume that the measure μ is ergodic.

For $v \in \mathbb{Z}^d$, write $v \geqslant 0$ if $v_i \geqslant 0$ for $i = 1, \ldots, d$. Define $|v|$ to be $\max |v_i|$. Write $A_k = \{v\colon v \geqslant 0; |v| < k\}$. For $x \in S^{\mathbb{Z}^d}$, let $\mathcal{P}_k(x) = \{y \in S^{\mathbb{Z}^d}\colon y_v = x_v,\ \forall v \in A_k\}$ and $L_v^n(x) = \inf\{k\colon \mathcal{P}_k(T_v(x)) \neq \mathcal{P}_k(T_u(x)),\ \forall u \in A_n \setminus \{v\}\}$.

The entropy h of a $\mathbb{Z}^d$ action is defined similarly to the entropy of a single transformation (see Krengel [7] for a relatively complicated definition). Ornstein and Weiss have proved a version of the Shannon–Macmillan–Breiman theorem (see [8]) which is applicable in this situation. It should be noted that the proof given by Ornstein and Weiss is exclusively for finite partitions, and it is for this reason that the results of this section are given in the case where S is finite. The result states (in this context) that for almost all $x \in S^{\mathbb{Z}^d}$,

$$\lim_{n\to\infty} \frac{1}{n^d} \log \mu(\mathcal{P}_n(x)) = -h. \tag{4}$$

Heavy use is made of this result and the constructions used in it in the proof of the higher dimensional version of Theorem 1.

The analogue of the Doeblin condition is there exists an $\alpha < 1$ and an $r \geqslant 1$ such that for all s and almost all x,

$$\mathbb{P}(X_0 = s | X_v = x_v, \forall |v| \geqslant r) \leqslant \alpha. \tag{5}$$

We then state the generalization of Theorem 1:

Theorem 5. *Suppose S is a finite set and μ is an ergodic invariant measure of a $\mathbb{Z}^d$ action on $S^{\mathbb{Z}^d}$ satisfying (5). Then we have for almost all $x \in S^{\mathbb{Z}^d}$,*

$$\lim_{n\to\infty} \frac{\frac{1}{n^d}\sum_{v\in A_n} L_v^n(x)}{(\log n)^{\frac{1}{d}}} = (d/h)^{\frac{1}{d}}\,.$$

Since the proof of this theorem is very similar to that of Theorem 1, we will not present a full proof here, but rather indicate the significant alterations from the proof given. The proof is divided into three lemmas, similar to the ones proving Theorem 1. Fix $\epsilon > 0$.

Lemma 6. *For almost every $x \in S^{\mathbb{Z}^d}$,*

$$\limsup_{n\to\infty} \frac{1}{n^d} \left| \left\{ v \in A_n\colon L_v^n(x) \leqslant \left(\frac{(1-\epsilon) d \log n}{h} \right)^{\frac{1}{d}} \right\} \right| = 0.$$

Lemma 7. *For almost every $x \in S^{\mathbb{Z}^d}$,*

$$\limsup_{n\to\infty} \frac{1}{n^d} \left| \left\{ v \in A_n\colon L_v^n(x) \geqslant \left(\frac{d \log n}{h-\epsilon} \right)^{\frac{1}{d}} \right\} \right| = 0.$$

Lemma 8. *There exists a $c > 0$ such that for almost every $x \in S^{\mathbb{Z}^d}$,*

$$\limsup_{n\to\infty} \max_{v\in A_n} \frac{L_v^n(x)}{(\log n)^{\frac{1}{d}}} \leqslant c.$$

The proofs of Lemmas 6 and 8 are so similar to the proofs of Lemmas 2 and 4 as to require nothing but obvious changes. Lemma 7 requires more attention. The main reason for this is that the theorem in [10] upon which Lemma 3 is based is only proved in the one-dimensional case. However, Ornstein and Weiss have recently produced a version of this result which is valid in higher dimensions (see [11]). We include a statement of the lemma in this paper which we need.

Define

$$R_k(x) = \inf\{|u| \colon \mathcal{P}_k(T_u(x)) = \mathcal{P}_k(x)\}.$$

Note that in the definitions, we do not require that the u in the infimum satisfy $u \geqslant 0$. This definition differs slightly from the one in [11], but the following result follows from [11].

Lemma 9. *For a stationary ergodic random field indexed by $\mathbb{Z}^d$ we have for almost every x,*

$$\lim_{n\to\infty} \frac{\log R_n(x)}{n^d} = \frac{h}{d}.$$

Assuming this, to prove Lemma 7, we proceed as in lemma 3 of Shields' paper, using Lemma 9 in place of the result from [10].

3. Application to Test Data

As a test of the methods described above, I tried to apply the technique to some test data for which the entropy was known. The results, while showing that the method is comparatively inaccurate suggest another entropy estimator, which is much more accurate. For this section, we restrict our attention to the one-dimensional case.

There were 4 classes of data to which I applied the estimator from Theorem 1:

(i) Bernoulli data (randomly generated by computer) with M equally probable symbols;

(ii) Markov data generated by computer with a simple 3×3 transition matrix;

(iii) Continued fraction data generated by calculating accurately a large number of digits of the continued fraction expansion of π.

(iv) Continued fraction data generated by simulating the continued fraction mapping $x \mapsto 1/x - \lfloor 1/x \rfloor$ on a computer with random initial conditions and with the inherent rounding error.

To apply the theory to the continued fraction map, T, I first took the standard generating partition $\mathcal{P} = \{\{0\}, (\frac{1}{2}, 1], (\frac{1}{3}, \frac{1}{2}], \dots\}$ and invariant measure m given by $dm(x) = 1/(\log 2(1+x))d\lambda(x)$. Then by standard symbolic dynamics techniques, each point x of [0,1] gives rise to a sequence of non-negative integers, the labels of the elements of the partition containing the successive iterates of x under T. These non-negative integers are also the terms of the continued fraction expansion. The invariant measure m for T gives rise to a shift-invariant measure μ on $(\mathbb{Z}^+)^{\mathbb{Z}^+}$. It is straightforward to check that this shift-invariant measure satisfies the conditions of Theorem 1. Its entropy can in fact be computed analytically. It is $h_{CF} = \pi^2/(6\log 2)$ (see [2]).

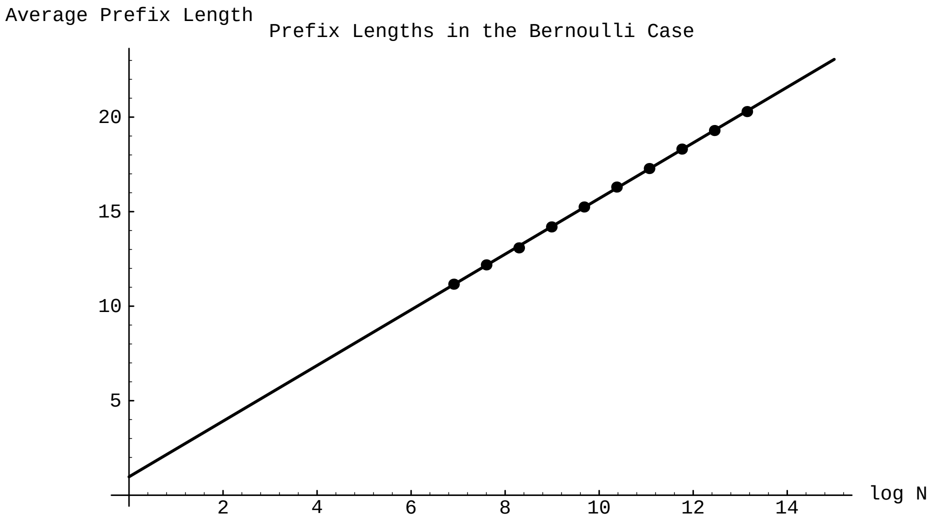
To apply Theorem 1, I took a large number N, computed a string of $N + n_0$ symbols of data (where I took n_0 to be 10^4 - this extra data is needed to be able to calculate the prefixes of the right-hand points of the sequence), and computed the prefix lengths L_i^N. Taking the average as in (2) gave experimental estimates of h for the various systems. My findings were that

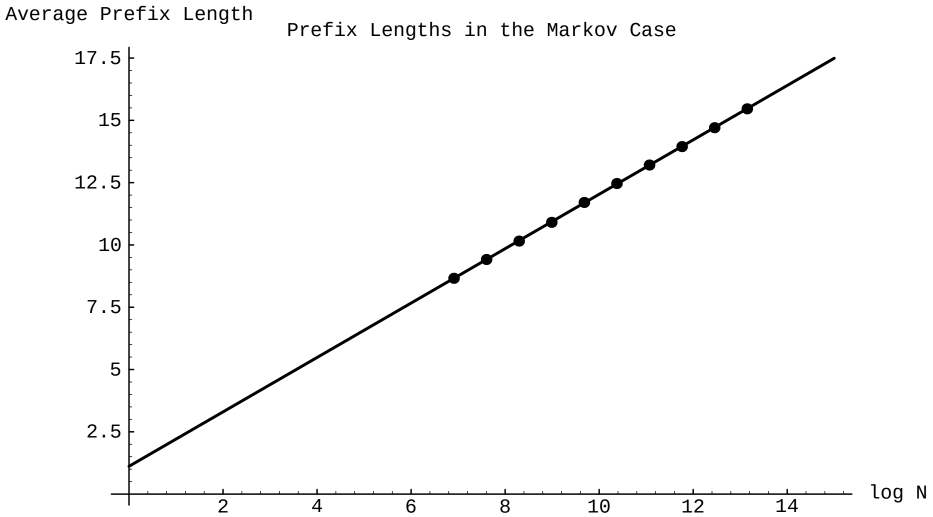
(i) The true continued fraction data for π and the simulated continued fraction map gave almost identical estimates for h_{CF}.

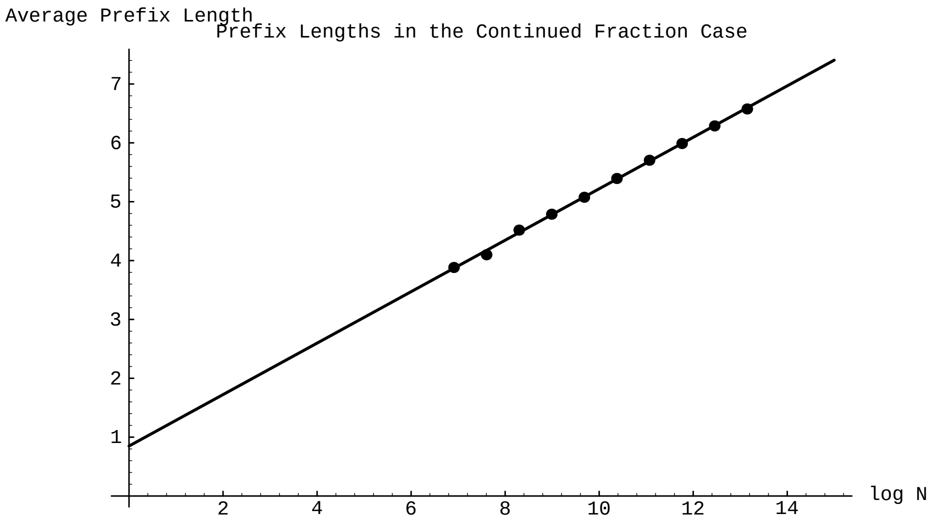
(ii) There were often very significant errors in the estimates of h (they were often 10–20% too low).

(iii) The estimated value of h for a given system and a given value of N was very robust to variations of the initial setting of the random number generator. For the Bernoulli system with $N = 10^6$, the estimates of h differed from one another by less than 0.1% on average. They were however 6% below the true figure. The other systems showed similar robustness (with the average variations decreasing in N as expected). For the continued fraction map, the error in the estimate for h_{CF} with $N = 10^6$ was about 15%. The average difference between estimates of h_{CF} was about 0.1% for $N = 10^6$.

These results were initially disappointing suggesting that the method is of limited applicability, but upon closer examination, it became clear that the errors had a systematic quality. For a given system, the average prefix length $\widehat{L^N} = \frac{1}{N}\sum_{i=0}^{N-1} L_i^N$ appears to differ from what one would like to see, $(\log N)/h$, by a constant. If this is so, one would have $\widehat{L^N} = (\log N)/h + C$. This would allow h to be estimated by linear regression. The following graphs show for a single sequence of data $\widehat{L^N}$ plotted against $\log N$.


Average Prefix Length
Prefix Lengths in the Bernoulli Case
20
15
10
5
2
4
6
8
10
12
14
log N


Average Prefix Length
Prefix Lengths in the Markov Case
17.5
15
12.5
10
7.5
5
2.5
2
4
6
8
10
12
14
log N


Average Prefix Length
Prefix Lengths in the Continued Fraction Case
7
6
5
4
3
2
1
2
4
6
8
10
12
14
log N

These clearly suggest strongly that there is an approximate linear relationship between $\widehat{L^N}$ and $\log N$. Performing the linear regression for this data gave estimates of h as follows:

System	Bernoulli	Markov	Continued Fraction
h_{actual}	0.693147	0.912696	2.373138
$h_{\text{estimated}}$	0.688037	0.915675	2.42215

If instead of using the data from a single sequence of symbols, the average data from a large number of sequences is used, the estimates are even better. The following data used linear regression on the average values of $\widehat{L^N}$ from 20 trials with $N = 10^4$, 10^5 and 10^6 giving estimates as follows:

System	Bernoulli	Markov	Continued Fraction
h_{actual}	0.693147	0.912696	2.373138
$h_{\text{estimated}}$	0.691644	0.911943	2.37403

The estimates are all within 0.3% of the true values, making this quite an accurate method.

4. Bounds on $\widehat{L^N} - \log N/h$

In this section, we derive some bounds for $|\widehat{L^N} - \log N/h|$ where the data is taken to be Bernoulli data. What is required is to show that $\widehat{L^N} - \log N/h = C + o(1)$ for some constant C. So far, I have been unable to show this, and comparison with the problem in [1] suggests that this may be quite hard. As a preliminary result, we give estimates of $\mathbb{P}(L_i^n(X) > \log n/h + k)$ and $\mathbb{P}(L_i^n(X) < \log n/h - k)$.

Lemma 10. *If $(X_i)_{i\in\mathbb{Z}}$ is a sequence of independent identically distributed equally weighted random variables taking values in $\{1, \dots, M\}$, then $\mathbb{P}(L_i^n(X) > l) \leqslant n/M^l$.*

Proof. If $L_i^n(X) > l$ then it follows that the word X_i^{i+l-1} is equal to a word X_j^{j+l-1} for some $j < n$ which is distinct from i. For fixed j, the probability of this happening is $1/M^l$, so we see straightforwardly that $\mathbb{P}(L_i^n(X) > l) \leqslant n/M^l$ as required. □

Lemma 11. *If $(X_i)_{i\in\mathbb{Z}}$ is a sequence of independent identically distributed equally weighted random variables taking values in $\{1, \dots, M\}$, then $\mathbb{P}(L_i^n(X) < l) \leqslant 3\exp(\frac{-n}{3M^l})$.*

Proof. If $L_i^n(X) < l$ then it follows that the block X_i^{i+l-1} does not recur for any $j < N$. To analyse the probability of this, observe that considering the blocks of length l gives a Markov chain, where the transition probability between two blocks $x_0 \dots x_{l-1}$ and $y_0 \dots y_{l-1}$ is $1/M$ if $y_0 = x_1; \dots; y_{l-2} = x_{l-1}$ and 0 otherwise. Given that X_j^{j+l-1} is a fixed block $W = w_0 \dots w_{l-1}$, this Markov chain may be further reduced as follows: Define $Y_j = \max\{k \leqslant l\colon X_{j+l-k} = w_0, X_{j+l-k+1} = w_1, \dots, X_{j+l-1} = w_{k-1}\}$. It is straightforward to check that Y_j is a Markov chain taking values in $\{0, \dots, l\}$. For each $m < l$, the probability of transition to $m + 1$ is given by $1/M$ and all the other states which can be reached in one step are less than or equal to m. As a simplifying assumption, we will assume for now that

$i = 0$. This helps as we then have $Y_0 = l$ and have to work out the probability that $Y_j \neq l$ for each $0 < j < n$. If i is non-zero, then we have to deal separately with the possibilities that $j < i$ and $j > i$. Let $p_0(W)$ be the probability that the word W never recurs given that it occurs in the first position. Then we have $p_0(W) \leqslant \mathbb{P}(Y_2 \neq l; \ldots ; Y_{n-1} \neq l |\, Y_1 = 0)$. One can then check that that the W for which this quantity is maximized is $W_0 = 00\ldots 0$, as then the transition matrix is the $(l+1) \times (l+1)$ matrix P given by

$$\begin{pmatrix} \frac{M-1}{M} & \frac{1}{M} & 0 & \cdots & & \\ \frac{M-1}{M} & 0 & \frac{1}{M} & 0 & \cdots & \\ \frac{M-1}{M} & 0 & 0 & \ddots & 0 & \cdots \\ \vdots & \vdots & & & \ddots & \\ \frac{M-1}{M} & 0 & \cdots & & 0 & \frac{1}{M} \\ \frac{M-1}{M} & 0 & \cdots & & 0 & \frac{1}{M} \end{pmatrix}$$

This transition matrix has the property that any added digit apart from a zero 'returns one to the bottom of the ladder', so has the least probability of hitting the l state from the 0 state. We then calculate the probability of not hitting l from the 0 state in $n-2$ steps for this Markov chain. This quantity is seen to be equal to $(1\ 0\ 0\ \ldots\ 0)Q^{n-2}(1\ 1\ 1\ \ldots\ 1)^T$ where Q is the $l \times l$ matrix obtained by truncating the bottom row and rightmost column of P:

$$Q = \begin{pmatrix} \frac{M-1}{M} & \frac{1}{M} & 0 & \cdots & & \\ \frac{M-1}{M} & 0 & \frac{1}{M} & 0 & \cdots & \\ \frac{M-1}{M} & 0 & 0 & \ddots & 0 & \cdots \\ \vdots & \vdots & & & \ddots & \\ \frac{M-1}{M} & 0 & \cdots & & 0 & \frac{1}{M} \\ \frac{M-1}{M} & 0 & \cdots & & 0 & 0 \end{pmatrix}$$

We proceed by estimating this. It may be verified that $(1\ \gamma\ \gamma^2\ \ldots\ \gamma^{l-1})$ is a left eigenvector of Q if γ satisfies $\gamma = \frac{1}{M} + \frac{M-1}{M}\gamma^{l+1}$. The eigenvalue is $1/(M\gamma)$. By the theory of Perron-Frobenius matrices (see [3]), Q has a unique left eigenvector with positive entries, and the corresponding eigenvalue is the eigenvalue of maximum modulus of Q.

It follows that

$$p_0(W) \leqslant (M\gamma)^{-(n-2)} \frac{1-\gamma^l}{1-\gamma} \leqslant \left(\frac{1}{M\gamma}\right)^{n-2} \frac{1}{1-\gamma}.$$

Estimating γ, we find $M\gamma \geqslant 1 + \frac{M-1}{M}\frac{1}{M^l}$ and can then show that $p_0(W_0) \leqslant 3\exp\left(\frac{-n}{3M^l}\right)$. It follows that $p_0(W) \leqslant 3\exp\left(\frac{-n}{3M^l}\right)$ for all W and so $\mathbb{P}(L_0^n(X) < l) \leqslant 3\exp\left(\frac{-n}{3M^l}\right)$. Dealing separately with the terms left of i and right of i gives the same estimate for $L_i^n(X)$ for $i \neq 0$. □

Noting that for the Bernoulli system, $h = \log M$, we see that setting $l_0 = \log n/h$, we have $\mathbb{P}(L_i^n(X) > l_0 + k) \leqslant 1/M^k$ and $\mathbb{P}(L_i^n(X) < l_0 - k) \leqslant 3\exp(-M^k/3)$. From this, it follows that $\mathbb{E}|L_i^n(x) - \log n/h|$ is bounded for all n and i, so in particular,

$\mathbb{E}|\widehat{L^n} - \log n/h|$ is bounded for all n. This is of course much weaker than what is required to prove that linear regression should give good estimates of h.

To get better bounds, the near independence of the random variables $L_i^n(X)$ would have to be taken into account. Some possible techniques work for identically distributed random variables. Defining $\widetilde{L_i^n}(X)$ to be $\inf\{k\colon X_i^{i+k-1} \neq X_j^{j+k-1},\ \forall |j-i| < n\}$ gives a sequence of random variables which is stationary. It is relatively straightforward to check that Theorem 1 holds with $\widetilde{L_i^n}$ replacing L_i^n throughout. This might give a sequence of random variables which would be more amenable to theoretical analysis.

5. Conjectures and Problems

We conclude with a number of questions and conjectures which have arisen in the course of this work.

The main conjecture is the one described in §3, that for a class of stochastic processes including the ones considered in this paper, with probability one, $\widehat{L^N} - \log N/h = C + o(1)$ for some constant C as $N \to \infty$.

It would be interesting to know how large the relevant class of stochastic processes is. In all the examples in this paper, if one (following Keane [5]) defines a function g by

$$g(x_0x_1\dots) = \mathbb{P}(X_0 = x_0 | X_{-1} = x_1,\ X_{-2} = x_2,\ \dots),$$

one finds that g is a Hölder continuous function. Stochastic processes with a Hölder continuous g-function have been widely studied and are known to have very good properties (see [16]). Maybe this is the right class of processes in which to try to prove the above conjecture.

If this conjecture can be proven, then one can get an entropy estimator by the linear regression technique of §3. This should be more accurate than the estimator derived in Theorem 1, but it would be interesting to see how it works with other kinds of data.

It would also be interesting to know if Ornstein and Weiss' result ([8]) holds in the context of countable partitions satisfying $\sum_{B\in\mathcal{P}} -\mu(B)\log(\mu(B)) < \infty$. This would allow the result for random fields to be generalized to the case of countably many symbols.

I should like to thank Yurii Suhov for introducing me to this subject area and for useful discussions in the course of the preparation of this paper. Thanks also to Robert Harley and Sheldon Parnes for providing me with a large number of digits in the continued fraction expansion of π.

References

1. R. Arratia and M. S. Waterman, *An Erdős-Rényi law with shifts*, Adv. Math. **55** (1985), 13–23.
2. P. R. Baldwin, *A multidimensional continued-fraction and some of its statistical properties*, J. Stat. Phys. **66** (1992), 1463–1505.
3. F. R. Gantmacher, *The theory of matrices*, New York, 1960.
4. P. Grassberger, *Estimating the information content of symbol sequences and efficient codes*, IEEE Trans. Inform. Theory **35**, 669–675.
5. M. Keane, *Strongly mixing g-measures*, Invent. Math **16** (1972), 309–324.

6. I. Kontoyiannis and Yu. Suhov, *Prefixes and the entropy rate for long-range sources*, Probability, statistics and optimization (F. P. Kelly, ed.), Wiley, New York, 1993.
7. U. Krengel, *Ergodic theorems*, de Gruyter, Berlin, 1985.
8. D. S. Ornstein and B. Weiss, *The Shannon–Macmillan–Breiman Theorem for a class of amenable groups*, Israel J. Math. **44** (1983), 53–60.
9. D. S. Ornstein and B. Weiss, *How sampling reveals a process*, Ann. Prob. **18** (1990), 905–930.
10. D. S. Ornstein and B. Weiss, *Entropy and data compression schemes*, IEEE Trans. Inform. Theory **39** (1993), 78–83.
11. D. S. Ornstein and B. Weiss, *Preprint: Entropy and recurrence rates for stationary random fields* (1995).
12. K. Petersen, *Ergodic Theory*, C.U.P., Cambridge, England, 1983.
13. P. C. Shields, *Entropy and Prefixes*, Ann. Prob. **20** (1992), 403–409.
14. P. C. Shields, *String matching: the ergodic case*, Ann. Prob. **20** (1992), 1199–1203.
15. P. C. Shields, *String matching bounds via coding*, Ann. Prob. **25** (1997), 329–336.
16. P. Walters, *Ruelle's operator theorem and g-measures*, Trans. Amer. Math. Soc. **214** (1975), 375–387.

Statistical Laboratory, Department of Pure Mathematics and Mathematical Statistics, 16 Mill Lane, Cambridge, CB2 1SB, ENGLAND

Current address: University of Memphis, Department of Mathematical Sciences, Campus Box 526429, Memphis, TN 38152-6429, U.S.A.

E-mail address: quasa@msci.memphis.edu