Nonparametric Entropy Estimation for Stationary Processes and Random Fields, with Applications to English Text

I. Kontoyiannis, *Student Member, IEEE* P.H. Algoet Yu.M. Suhov A.J. Wyner, *Member, IEEE*

Abstract – We discuss a family of estimators for the entropy rate of a stationary ergodic process and prove their pointwise and mean consistency under a Doeblin-type mixing condition. The estimators are Cesàro averages of longest match-lengths, and their consistency follows from a generalized ergodic theorem due to Maker. We provide examples of their performance on English text, and we generalize our results to countable alphabet processes and to random fields.

Index Terms – Entropy rate, pattern matching, universal data compression, entropy of English.

I. Introduction

Since the mid 1980's, a lot of work has been done in relating the entropy rate of a stationary ergodic process to the geometry along a single realization. The entropy rate is almost surely an asymptotic lower bound on the per-symbol description length when the process is losslessly

[1]I. Kontoyiannis and P. Algoet are with the Information Systems Laboratory (Durand 141A), Electrical Engineering Department, Stanford University, Stanford CA 94305. Email: `yiannis@isl.stanford.edu`, and `paul@isl.stanford.edu`. Yu. M. Suhov is with the Statistical Laboratory, DPMMS, Cambridge University, and also with the Institute for Problems in Information Transmission, Moscow. Email: `y.m.suhov@statslab.cam.ac.uk`. A.J. Wyner is with the Statistics Department, University of California, Berkeley, CA 94720-3860. Email: `ajw@stat.berkeley.edu`.

[2]I. Kontoyiannis was supported in part by grants NSF #NCR-9205663, JSEP #DAAH04-94-G-0058, ARPA #J-FBI-94-218-2, and by grant EPSRC GR/J31896, during a two-week visit to the Statistical Laboratory, Cambridge University.

[3]A preliminary version of this paper was presented at the 1996 Data Compression Conference, Snowbird, Utah, April 1996.

encoded, and several universal data compression algorithms are known that actually achieve it. In particular, the Lempel-Ziv [31] algorithm attains the entropy lower bound when it is applied to almost every realization of a stationary ergodic source.

A straightforward approach for estimating the entropy rate of an unknown source would be to run a universal coding algorithm on a long segment of the source output. The resulting compression ratio can be used as an upper bound for the entropy. If the data segment is long enough for the algorithm to converge, then the compression ratio is a good estimate for the source entropy. But like for the ergodic theorem, also for data compression there is no universal rate of convergence [21][23]. Moreover, few of the known universal coding algorithms have been shown to achieve the entropy limit in the pointwise sense, and of those, not all are feasible to implement.

In practice, it is often found that universal compression algorithms converge rather slowly, and other approaches, tailored to the specific application at hand, are often employed. After all, estimating the entropy is a simpler task, at least in principle, than compressing an unknown source to the entropy limit.

Wyner and Ziv [29], motivated in part by the problem of providing a pointwise asymptotic analysis of the Lempel-Ziv algorithm, revealed some deep connections between the entropy rate of a stationary ergodic process and the asymptotic behavior of longest match-lengths along a process realization: Let $\{X_i\}$ be a random process with values in a finite alphabet $\mathcal{A}$. A process realization is an element $x = (x_i)_{i \in \mathbf{Z}}$ of the two-sided sequence space $\mathcal{X} = \mathcal{A}^{\mathbf{Z}}$, and $X_i(x) = x_i$ is its ith coordinate. For $i \leq j$, X_i^j denotes the string $(X_i, X_{i+1}, \ldots, X_j)$. For our purposes, the process distribution is a probability measure P on the Borel σ-field on $\mathcal{X}$. We assume that P is invariant under the usual shift transformation $Tx = (x_{i+1})_{i \in \mathbf{Z}}$, so that $\{X_i\}$ is stationary, and we also assume ergodicity. The entropy rate of the process is defined as

$$H \;=\; E\{-\log P(X_0 | X_{-\infty}^{-1})\}.$$

For $n \geq 1$ let L_n denote the minimum length k of the string X_0^{k-1} that starts at time 0 and does not appear as a continuous substring within the past X_{-n}^{-1}. Alternatively, L_n is obtained by adding 1 to the longest match-length:

$$L_n = 1 + \max\{l:\; 0 \leq l \leq n,\; X_0^{l-1} = X_{-j}^{-j+l-1} \text{ for some } l \leq j \leq n\}.$$

Wyner and Ziv [29] showed that, for every ergodic process, L_n grows like $(\log n)/H$ in probability, and Ornstein and Weiss [15] later refined this to pointwise convergence,

$$\frac{L_n}{\log n} \to \frac{1}{H} \quad \text{a.s.} \tag{1}$$

where H is the entropy rate of $\{X_i\}$ (logarithms are to base 2 throughout the paper).

At about the same time, Grassberger [9] suggested an interesting entropy estimator based on average match-lengths. Shields [22] proved the consistency of Grassberger's estimator for independent identically distributed (i.i.d.) processes and mixing Markov chains. Kontoyiannis and Suhov [11] extended this to a wider class of stationary processes, and recently Quas [20] extended it further to certain processes with infinite alphabets and to random fields.

In this paper we introduce three entropy estimators, (a), (b) and (c) below, that are formally similar to the one suggested by Grassberger but which, due to their stationary nature, are much easier to analyze. We establish their pointwise consistency using methods with a familiar information-theoretic flavor (Section III), and we discuss generalizations to random fields (Section IV) and processes with countably infinite alphabets (Section V). We also provide examples of their performance in estimating the entropy of English text, and we compare our results to those obtained using a variety of different techniques (Section II).

Given an instant i and a positive integer n, our main quantity of interest is $\Lambda_i^n(X) = L_n(T^i X)$, the length of the shortest substring X_i^{i+k-1} starting at position i that does not appear as a contiguous substring of the previous n symbols X_{i-n}^{i-1}.

Theorem 1. *Let $\{X_i\}$ be a stationary ergodic process with entropy rate $H > 0$. Then*

$$\begin{aligned}
(a) \quad & \lim_n \frac{1}{n} \sum_{i=1}^{n} \frac{\Lambda_i^n}{\log n} = \frac{1}{H} \quad \text{a.s. and in } L^1, \\
(b) \quad & \lim_n \frac{1}{n} \sum_{i=2}^{n} \frac{\Lambda_i^i}{\log i} = \frac{1}{H} \quad \text{a.s. and in } L^1, \\
(c) \quad & \lim_n \frac{1}{n} \sum_{i=1}^{n} \frac{\Lambda_i^i}{\log n} = \frac{1}{H} \quad \text{a.s. and in } L^1,
\end{aligned}$$

provided the following condition holds:

Doeblin Condition (DC): *There exists an integer $r \geq 1$ and a real number $\beta \in (0,1)$ such that,*

$$\text{for all } x_0 \in \mathcal{A}, \qquad P\{X_0 = x_0 | X_{-\infty}^{-r}\} \leq \beta, \qquad \text{with probability one.}$$

Without (DC), $1/H$ is still an asymptotic lower bound for the estimates in (a) and (b).

Remarks:

1. Applications. Entropy estimators similar to the one in (a) have already appeared in the literature [8][4][5][27][10]. They were applied to experimental data in order to determine the entropy rate of the underlying process, and were demonstrated to be very efficient, even when fed with very limited amounts of data. So part of our motivation is to provide a more general and precise analysis of these practical algorithms.

2. The Doeblin condition. The Doeblin condition was originally introduced in the analysis of Markov chains [7]. In the context of this paper (DC) was first introduced by Kontoyiannis and Suhov [11], where its properties are discussed in greater detail. Here we note that (DC) holds for i.i.d. processes, for ergodic Markov chains of any order, and also for certain non-Markov processes. Our present formulation of (DC), introduced by Quas [20], is equivalent to that in [11] when the alphabet is finite, and has the additional advantage of being applicable to processes with countably infinite alphabets.

In practice, (DC) is not a fierce restriction. What (DC) requires is that, after some number r of time steps, everything is possible again with positive probability, independently of whatever may have occurred in the past. This is certainly satisfied by natural languages, and it is highly plausible for most sources of data encountered in practice. Observe also that (DC) is satisfied by any stationary ergodic process observed through a discrete memoryless channel which transforms any letter of the alphabet to any other letter, with non-zero (but arbitrarily small) probability. For example, if $\{\xi_n\}$ is a stationary ergodic binary process and $\{\epsilon_n\}$ is an independent identically distributed noise sequence with $P\{\epsilon_n = 1\} = p$, then the dithered process $X_i = \xi_i + \epsilon_i$ satisfies (DC) with $r = 1$ and $\beta = \max\{p, 1-p\}$.

3. Without the Doeblin Condition. Without assuming (DC), the estimates in (a) and (b) can still be used to provide lower bound estimates for the entropy. Notice that these bounds are in the opposite direction from the ones provided by universal coding algorithms, so that, even if (DC) does not hold, we can use either (a) or (b) in conjunction with a data compression algorithm to estimate upper and lower bounds for the process entropy.

4. Interpretation. The match-length Λ_i^n can be thought of as the length of the next phrase to be encoded by the sliding-window Lempel-Ziv algorithm [30] when the window size is n. In fact, the entropy estimator in (a) above is a special case of the sliding window estimator [8] defined by

$$\hat{H}_{k,n} = \frac{1}{k} \sum_{i=1}^{k} \frac{\Lambda_i^n}{\log n}, \tag{2}$$

where k (as well as n) are freely chosen positive integers. From (a) we learn that SWE is almost surely consistent for $k = n$ tending to infinity. Also, it is a consequence of the ergodic theorem that $\hat{H}_{k,n}$ is almost surely consistent if we first let $k \to \infty$ and then $n \to \infty$, provided that $E[L_n/(\log n)] = 1/H + o(1)$. This is true for stationary ergodic Markov sources [28] and, by (6) of Theorem $1'$ below combined with (1) it is also true for all stationary ergodic processes satisfying (DC).

Similarly, Λ_i^i is the length of the phrase that would be encoded next by the Lempel-Ziv algorithm [31] with knowledge of the past X_0^{i-1}. From (1) and stationarity it follows that, for any fixed index i, $\Lambda_i^n/\log n \to 1/H$ with probability one, and also that $\Lambda_n^n/\log n \to 1/H$ in probability. Theorem 1 says that, when (DC) holds, the Cesàro means of these quantities also converge to $1/H$, with probability one.

5. Maker's generalized ergodic theorem. The proof of Theorem 1 is based on the fact that, under (DC), we can invoke a generalized ergodic theorem due to Maker [14] and conclude that the Cesàro averages in Theorem 1 are pointwise consistent estimates for $1/H$. Maker's theorem includes, as a special case, Breiman's ergodic theorem, which was used in [3] to prove the Shannon-McMillan-Breiman theorem. In the Appendix we present a simplified proof of Maker's generalized ergodic theorem, and some extensions that are used in Sections III and IV.

6. A word of caution. Theorem 1 says that the Cesàro averages of the quantities $\Lambda_i^n/\log n$ and of the quantities $\Lambda_i^i/\log i$ converge with probability one, but Pittel [19] and Szpankowski [24] have shown that the quantities $\Lambda_n^n/\log n$ themselves keep fluctuating. Interpreting Λ_n^n as the length of a feasible path in a suffix tree they identify two natural constants H_1 and H_2 with $H_1 > H > H_2$, and they show that, under certain mixing conditions,

$$\frac{1}{H_1} = \liminf_n \frac{\Lambda_n^n}{\log n} < \limsup_n \frac{\Lambda_n^n}{\log n} = \frac{1}{H_2} \quad \text{a.s.}$$

II. Applications to English Text

In this section we present numerical results for the performance of our estimators when applied to English text data. We also provide a heuristic discussion that motivates the results and offers an alternative derivation for the form of these estimators.

Our experiments were done on Jane Austen's four novels *Mansfield Park*, *Northanger Abbey*, *Persuasion* and *Sense and Sensibility*, a total of 2,364,200 characters, converted to 27-character text (26 letters plus space). We use this text to demonstrate the convergence properties of our methods. We show that our estimators are very efficient even for small sample sizes, and we claim that they offer a significant improvement over the universal estimators that are based on the corresponding versions of Lempel-Ziv coding algorithms (cf. remark 4 in Section I).

For example, using estimator (a) from Theorem 1, we obtain an estimate of 1.777 bits-per-character (bpc) based on a sample of about 75,000 words from the novel *Mansfield Park* by Jane Austen (the first 23 chapters, about 400,000 characters), a rather modest sample size.

A. Experimental Results

A naive approach to the problem of estimating the entropy of a given text would be to use the Wyner-Ziv-Ornstein-Weiss result (1), which says that $L_n/\log n$ converges to $1/H$ for any stationary ergodic process. But it is intuitively clear that such estimates would depend heavily on our choice of the exact position in the text where we look for a match, and hence they would fluctuate a lot (cf. remark 6 in Section I). Indeed, this behavior is demonstrated by the results shown in Table I.

TABLE I

ENTROPY ESTIMATES BASED ON A SINGLE MATCH. THE WINDOW-LENGTH n IS EQUAL TO 100,000 CHARACTERS, THE TEXT IS JANE AUSTEN'S NOVEL *Mansfield Park*, AND THE SEGMENT OF THE TEXT THAT IS BEING MATCHED IS `"...tea when you and your mamma went out of the room..."`.

Matching position i	Match Length $L_n(T^i x)$	Entropy Estimate	Matching Phrase
100,000	4	**4.152**	`"tea "`
100,003	11	**1.510**	`" when you a"`
100,006	8	**2.076**	`"en you a"`
100,009	7	**2.373**	`"you and"`
100,012	12	**1.384**	`" and your ma"`

In accordance with standard statistical methodology, what we would want to do in order to decrease the fluctuations of these estimates is to calculate match-lengths $\Lambda_i^n(x) = L_n(T^i x)$ at several positions i along the "text" x and take averages. This should reduce the variance of the estimates, but also introduce more systematic error for finite values of n (increase the bias). Since the bias eventually decreases with increasing n, it is natural to calculate match lengths into the largest available cache of past observations. The entropy estimators of Theorem 1, parts (b) and (c) both form averages of match lengths for values of i ranging from 1 to n, partly explaining our choice of the estimator implicit in (a) instead of (b) or (c).

Theorem 1 part (a) says that the corresponding entropy estimates

$$\hat{H}_n = \left[\frac{1}{n} \sum_{i=1}^{n} \frac{L_n(T^i x)}{\log n} \right]^{-1} \tag{3}$$

are consistent, provided the underlying process satisfies (DC). (A similar case can be made for the estimators in parts (b) and (c) of Theorem 1.) Using (3) we obtained the results shown in Table II.

TABLE II

ENTROPY ESTIMATES BASED ON AVERAGE MATCH-LENGTHS, USING ESTIMATOR (3) WITH VARYING WINDOW SIZES. THE TEXT IS THE FOUR JANE AUSTEN NOVELS *Mansfield Park*, *Northanger Abbey*, *Persuasion* AND *Sense and Sensibility*.

Window Size n	$\hat{H}_n$	Total Data length
100	**2.356**	300
500	**2.047**	1,500
1,000	**2.009**	2,500
5,000	**1.997**	10,500
10,000	**1.937**	20,500
50,000	**1.886**	100,500
100,000	**1.841**	200,500
300,000	**1.774**	600,500
500,000	**1.794**	1,000,500
700,000	**1.769**	1,400,500
1,000,000	**1.761**	2,000,500
1,181,850	**1.749**	2,364,200

B. Comparison with Universal Coding Algorithms.

Here we compare the performance of our estimators with that of the corresponding Lempel-Ziv coding schemes (cf. remark 4 in Section I). In particular, we argue that we get a significant improvement in performance, at the price of assuming the extra condition (DC).

For the purposes of comparison, we begin by recalling the sliding-window Lempel-Ziv (SWLZ) coding algorithm (or, rather, a slightly idealized version of it). Suppose that the encoder and the decoder both have available to them the previous n characters of the text, x_{-n+1}^{0}, which we call the "window" of length n, and the encoder's task is to describe the next n characters, x_1^n, to the decoder. SWLZ operates as follows.

The encoder calculates the length $\Lambda = \Lambda_1^n$, and then describes the phrase x_1^Λ in two stages: First, Λ is described; this takes $\log(\Lambda) + C\log\log(\Lambda)$ bits. Then the encoder either describes the position in the window where the match occurs plus the last character x_Λ, or the actual phrase x_1^Λ in binary, whichever of the two is shorter (plus a one-bit flag to say which of the two was used). So the description length of x_1^Λ (in bits) is

$$\log(\Lambda_1^n) + C\log\log(\Lambda_1^n) + \min\{\log n + \lceil \log 27 \rceil,\ \lceil \Lambda_1^n \log 27 \rceil\} + 1.$$

Having described x_1^Λ, the encoder shifts the window by $i = \Lambda_1^n$ places to the right, and then repeats the above process to encode the next phrase,

$$x_{i+1}^{i+\Lambda_{i+1}^n}.$$

Then the window is shifted again, and the same process is repeated until the entire string x_1^n has been encoded. This encoding scheme produces a sequence of positions $1 = i_1 < i_2 < \cdots < i_J$ where the successive phrases begin, and the overall description length for x_1^n can be written as

$$\sum_{j=1}^{J} \left[\log(\Lambda_{i_j}^n) + C\log\log(\Lambda_{i_j}^n) + \min\{\log n + \lceil \log 27 \rceil,\ \lceil \Lambda_1^n \log 27 \rceil\} + 1\right],$$

where J in the total number of phrases. Since

$$n = \sum_{j=1}^{J} \Lambda_{i_j}^n,$$

the number of bits-per-character used to describe x_1^n can be written as

$$\frac{\sum_{j=1}^{J} \left[\log(\Lambda_{i_j}^n) + C\log\log(\Lambda_{i_j}^n) + \min\{\log n + \lceil \log 27 \rceil,\ \lceil \Lambda_1^n \log 27 \rceil\} + 1\right]}{\sum_{j=1}^{J} \Lambda_{i_j}^n}. \tag{4}$$

We could use (4) to obtain an upper bound, but in practice such bounds tend to overestimate the entropy (see Table III), especially for small values of the window length n. In order to get lower (and hence better) estimates, we drop all of the terms in the numerator of (4) except the leading $\log n$ term, and set

$$\begin{aligned} \tilde{H}_n &= \frac{J \log n}{\sum_{j=1}^J \Lambda_{i_j}^n} \\ &= \left[\frac{1}{J} \sum_{j=1}^J \frac{L_n(T^{i_j} x)}{\log n} \right]^{-1}. \end{aligned} \tag{5}$$

We no longer know that $\tilde{H}_n$ converges to H, but we do know that the estimates it produces will be significantly lower than those produced by (4). Now if instead of looking at matches along the subsequence $\{i_j\}$, we look at all positions $i = 1, 2, \ldots, n$ along x_1^n, then, setting $J = n$, reduces $\tilde{H}_n$ in (5) to $\hat{H}_n$ in (3), and from Theorem 1 we know that, under (DC), $\hat{H}_n$ indeed converges to the entropy rate.

In Table III we compare the performance of our estimator $\hat{H}_n$ with that of the expression $\tilde{H}_n$ in (5), and with the compression-ratio achieved by SWLZ.

TABLE III

ENTROPY ESTIMATES BASED ON OUR ESTIMATOR $\hat{H}_n$ FROM THEOREM 1 PART (a), ON THE EXPRESSION $\tilde{H}_n$ IN (5), AND ON THE COMPRESSION-RATIO H_n^{SWLZ} ACHIEVED BY SWLZ.

Window Size n	H_n^{SWLZ}	$\tilde{H}_n$	$\hat{H}_n$
100	5.449	2.458	**2.356**
500	3.838	2.295	**2.047**
1,000	3.570	2.212	**2.009**
5,000	3.153	2.145	**1.997**
10,000	2.974	2.046	**1.937**
50,000	2.747	1.962	**1.886**
100,000	2.693	1.891	**1.841**
300,000	2.661	1.823	**1.774**
500,000	2.666	1.846	**1.794**

We point out that there are many other, perhaps more suitable, choices of the the number of phrases k in (2). In applications, one must consider restrictions on the available data as well as particular source characteristics, like changing source statistics. This point is certainly true for English text, and it is a possible explanation for the increase in the entropy estimate observed for the largest value of the window size given in the final entry of table III (it is likely that, as the window grows, differences in vocabulary across chapters have the effect of slightly elevating the entropy).

The estimator we use here is (2) with $k = n$; since we expect its variance to decrease like $O(1/k)$ and the bias like $O(1/\log n)$, it is likely that the number of phrases k does not need to be as large as n, and may in fact be quite a bit smaller. Indeed, the proof of Theorem 1 can easily be extended to prove the consistency of $\hat{H}_{k,n}$ for k being linear in n. It is an open problem to determine the most general way in which k and n can tend to infinity while $H_{k,n}$ remains consistent.

C. Other Methods.

The problem of estimating the entropy of English text has a long history. The most successful methods almost always involve a training stage or some sort of preprocessing of the data. (There is extensive literature that deals with language- and text-modeling, and several special-purpose algorithms that provide very good estimates; see, for example, Teahan and Cleary [25] and the references therein.) The best results to date are those reported by Teahan and Cleary [26], using PPM-related methods. They obtain an estimate of 1.603 bpc for the complete works of Jane Austen (over 4,000,000 characters), and then use training in conjunction with an alphabet enlargement technique ("bigram encoding") to improve this to 1.48 bpc. Although our results may not be as accurate as those obtained by the best known techniques, they do have several advantages which make them applicable to a wide range of problems rather than just to English text: they do not require prior training or preprocessing of the data, they do not assume that the underlying source belongs to some presupposed parametric class, they have relatively good performance for small data sets, and they are very easy to implement in practice.

III. Proof of Theorem 1

In this section we prove Theorem 1 by invoking Maker's generalized ergodic theorem (discussed and proved in the appendix). In fact, we prove a slightly stronger result:

Theorem 1′. *Let $\{X_i\}$ be a stationary ergodic process with entropy rate $H > 0$. Then*

$$
\begin{aligned}
(a)\quad & \lim_n \frac{1}{n}\sum_{i=1}^{n} \frac{\Lambda_i^n}{\log n} = \frac{1}{H} \quad \text{a.s. and in } L^1,\\
(b)\quad & \lim_n \frac{1}{n}\sum_{i=2}^{n} \frac{\Lambda_i^i}{\log i} = \frac{1}{H} \quad \text{a.s. and in } L^1,\\
(c)\quad & \lim_n \frac{1}{n}\sum_{i=1}^{n} \frac{\Lambda_i^i}{\log n} = \frac{1}{H} \quad \text{a.s. and in } L^1,
\end{aligned}
$$

provided the random variables $L_n/\log n$ are L^1-dominated:

$$
E\left\{\sup_n \frac{L_n}{\log n}\right\} < \infty. \tag{6}
$$

Without (6), $1/H$ is still an asymptotic lower bound for the estimates in (a) and (b).

And then we also check that (DC) implies (6).

Lemma 1. *Let $\{X_i\}$ be a stationary process. If (DC) holds, then*

$$
P\{L_n > k\} \le n\beta^{\lfloor k/r \rfloor}, \quad k \ge 1, \tag{7}
$$

and the random variables $L_n/\log n$ are L^1-dominated.

Proof of Theorem 1′: Recall that $L_n/\log n \to 1/H$ with probability one and invoke Maker's generalized ergodic theorem. Assertion (a) follows from Theorem 4 in the appendix by setting $g_{n,i} = L_n/\log n$, whereas (b) follows by setting $g_{n,i} = L_i/\log i$. (Similarly for the one-sided counterparts to (a) and (b).)

We can deduce (c) from (b) as follows. Let $l_i = \Lambda_i^i$ and observe that by (b),

$$
\frac{1}{n}\left(\frac{l_2}{\log 2} + \ldots + \frac{l_n}{\log n}\right) \to \frac{1}{H} \quad \text{a.s. and in } L^1. \tag{8}
$$

If $0 < \epsilon < 1$ and $\epsilon n \le i \le n$, then $0 \le \log n - \log i \le \log(1/\epsilon)$ and hence

$$
0 \le \frac{l_i}{\log i} - \frac{l_i}{\log n} \le \delta_n \frac{l_i}{\log i},
$$

where $\delta_n = -\log\epsilon/\log n$. It follows that

$$
\begin{aligned}
0 \;&\le\; \frac{1}{n}\left(\frac{l_2}{\log 2} + \ldots + \frac{l_n}{\log n}\right) - \frac{l_2 + \ldots + l_n}{n\log n}\\
&\le\; \epsilon\,\frac{1}{\epsilon n}\left(\frac{l_2}{\log 2} + \ldots + \frac{l_{\epsilon n}}{\log(\epsilon n)}\right) + \frac{\delta_n}{n}\left(\frac{l_2}{\log 2} + \ldots + \frac{l_n}{\log n}\right).
\end{aligned}
$$

Letting $n \to \infty$ and using (8), we may conclude that for any $\epsilon > 0$,

$$0 \leq \lim_n \left[\frac{1}{n} \left(\frac{l_2}{\log 2} + \ldots + \frac{l_n}{\log n} \right) - \frac{l_2 + \ldots + l_n}{n \log n} \right] \leq \epsilon.$$

The limit must vanish, and this fact in combination with (8) yields (e), since $l_i = \Lambda_i^i$. ∎

Proof of Lemma 1: This follows along the lines of the proof of Lemma 3.1 in [11]. First we prove that L_n has exponentially vanishing tails. If $L_n > k$ then X_0^{k-1} appears as a substring X_{-j}^{-j+k-1} of X_{-n}^{-1}, for some $k \leq j \leq n$. Therefore,

$$\begin{aligned} P\{L_n > k\} &\leq \sum_{k \leq j \leq n} P\{X_0^{k-1} = X_{-j}^{-j+k-1}\} \\ &\leq (n-k+1) \max_{k \leq j \leq n} P\{X_0^{k-1} = X_{-j}^{-j+k-1}\}. \end{aligned} \tag{9}$$

Write x^k for $(x_0, \ldots, x_{k-1})$ and observe that

$$P\{X_0^{k-1} = X_{-j}^{-j+k-1}\} = \sum_{x^k \in \mathcal{A}^k} P\{X_0^{k-1} = x^k | X_{-j}^{-j+k-1} = x^k\} P\{X_{-j}^{-j+k-1} = x^k\}.$$

Using (DC),

$$\begin{aligned} P\{X_0^{k-1} = x^k | X_{-j}^{-j+k-1} = x^k\} &\leq P\{X_{rt} = x_{rt},\ 0 \leq t < \lfloor k/r \rfloor \,|\, X_{-j}^{-j+k-1} = x^k\} \\ &\leq \prod_{t=0}^{\lfloor k/r \rfloor} P\{X_{rt} = x_{rt} | X_{-j}^{-j+k-1} = x^k, X_{rs} = x_{rs},\ 0 \leq s < t\} \\ &\leq \beta^{\lfloor k/r \rfloor}. \end{aligned}$$

It follows that $P\{X_0^{k-1} = X_{-j}^{-j+k-1}\} \leq \beta^{\lfloor k/r \rfloor}$. Substituting this bound in (9) above, we obtain (7). It is now a routine calculation to verify the L^1-domination of the random variables $L_n/\log n$. Indeed, let $\gamma = (-\log \beta)/(2r)$ and observe that for $k \geq 4r$ we have $\lfloor \lfloor k \log n \rfloor / r \rfloor \geq (k \log n)/r - 1 - 1/r \geq (k \log n)/(2r)$. Consequently, for $K \geq 4r$,

$$\begin{aligned} E\left\{ \sup_{n \geq 2} \frac{L_n}{\log n} \right\} &= \int_0^\infty P\left\{ \sup_{n \geq 2} \frac{L_n}{\log n} > k \right\} dk \\ &\leq K + \int_K^\infty \sum_{n \geq 2} P\{L_n > k \log n\}\, dk \\ &\leq K + \sum_{n \geq 2} \int_K^\infty n \beta^{\lfloor \lfloor k \log n \rfloor / r \rfloor}\, dk \\ &\leq K + \sum_{n \geq 2} \int_K^\infty n\, \beta^{(k \log n)/(2r)}\, dk \end{aligned}$$

$$
\begin{aligned}
&= K + \sum_{n\geq 2} \int_K^\infty n^{1-\gamma k}\, dk \\
&= K + \sum_{n\geq 2} \frac{n^{1-\gamma K}}{\gamma \ln n}.
\end{aligned}
$$

The sum is finite if the constant K is chosen so that $K > 2/\gamma$. ∎

IV. Generalization to Random Fields

We generalize our results to random fields on the integer lattice $\mathbf{Z}^d$. Such a random field is a family of random variables $\{X_u : u \in \mathbf{Z}^d\}$ indexed by d-dimensional integer vectors $u = (u_1, \ldots, u_d)$. We assume that all X_u take values in a finite set $\mathcal{A}$. The process distribution is a stationary ergodic probability measure P on the product space $\mathcal{X} = \prod \{\mathcal{A}_u : u \in \mathbf{Z}^d\}$, where each $\mathcal{A}_u$ is a copy of $\mathcal{A}$. If $x = \{x_u : u \in \mathbf{Z}^d\}$ is a realization in $\mathcal{X}$, then $X_u(x) = x_u$ is the coordinate at position u. For a subset $U \subseteq \mathbf{Z}^d$ let $X_U(x) = (x_u)_{u\in U}$. For any vector $v \in \mathbf{Z}^d$, let $T_v x$ denote the realization with coordinates $X_u(T_v x) = X_{u+v}(x) = x_{u+v}$. We say that X_U occurs at position v if $X_U = X_{v+U}$.

For $u, w \in \mathbf{Z}^d$ let $[u, w) = \{v \in \mathbf{Z}^d : u_j \leq v_j < w_j \text{ for all } j\}$. The d-dimensional cube with side k is defined, for any integer $k \geq 1$, as the cartesian product $[0, k)^d$:

$$
C(k) = \{u \in \mathbf{Z}^d : 0 \leq u_j < k \text{ for all } j\}.
$$

We define L_n here as the analogous quantity to the match-length L_n in one dimension: L_n is the minimum value of k such that $X_{-C(k)}$ does not occur anywhere in $-C(n)$ except at position $\mathbf{0}$:

$$
L_n = \inf\{k \geq 1 : X_{-C(k)} \neq X_{-u-C(k)} \text{ for some } u \in C(n),\ u \neq \mathbf{0}\}.
$$

Figure 1 shows an example of L_6 for a binary random field in two dimensions. We also define a dual quantity, the recurrence time R_k for an integer $k \geq 1$ as the minimum value of n such that the block $X_{-C(k)}$ occurs at some position other than position $\mathbf{0}$ inside $-C(n)$:

$$
R_k = \inf\{n \geq 1 : X_{-C(k)} = X_{-u-C(k)} \text{ for some } u \in C(n),\ u \neq \mathbf{0}\}.
$$

Notice that R_k and L_n are related by the following relationship:

$$
L_n \leq k \quad \text{iff } R_k > n. \tag{10}
$$

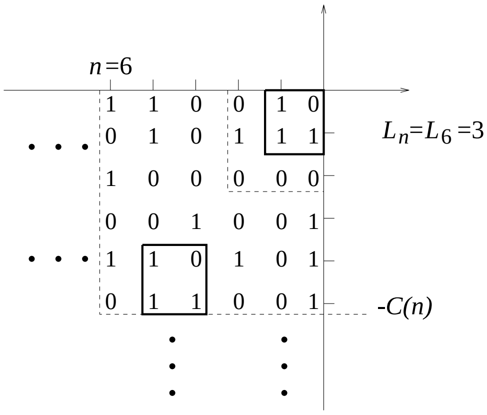

Figure 1: Example of L_n in two dimensions.

Applying a result of Ornstein and Weiss [16] to the reflected field $\{X_{-u}\}$, we see that

$$\frac{\log R_k^d}{k^d} \to H \quad \text{a.s.} \tag{11}$$

and from the duality relationship (10) it follows immediately that

$$\frac{L_n^d}{\log n^d} \to \frac{1}{H} \quad \text{a.s.} \tag{12}$$

Note that k^d, n^d, R_k^d and L_n^d are the volumes of cubes with sides k, n, R_k and L_n, respectively.

We now introduce a Doeblin-type condition for random fields $\{X_u : u \in \mathbf{Z}^d\}$:

d-dimensional Doeblin Condition (dDC): *There exists an integer $r \geq 1$ and a real number $\beta \in (0,1)$ such that, for all $x_0 \in \mathcal{A}$,*

$$P\{X_0 = x_0 \,|\, X_{-\infty}^{-r}\} \leq \beta \quad \text{a.s.}$$

In the d-dimensional case, the "past" $X_{-\infty}^{-r}$ is defined as the family of random variables X_u with index vectors $u = (u_1, \ldots, u_d)$ such that $(\lceil u/r \rceil, \ldots, \lceil u/r \rceil)$ lexicographically precedes $\mathbf{0} = (0, \ldots, 0)$ in $\mathbf{Z}^d$. In particular, if $r = 1$ then $X_{-\infty}^{-r} = \{X_u : u \in \mathbf{Z}^d, u \prec \mathbf{0}\}$ is the part of the random field that lexicographically precedes X_0 in a generalized raster scan. Figure 2 shows the two-dimensional region $X_{-\infty}^{-r}$ when $d = 2$ and $r = 5$.

We can now study the analog of Theorem 1 for random fields. For $n \geq 1$ and any vector $u = (u_1, \ldots, u_d)$ in $\mathbf{Z}^d$ let $\Lambda_u^n(x) = L_n(T_u x)$ denote the smallest integer k such that the block $X_{u-C(k)}$ does not occur within the translated cube $u - C(n)$ except at position u. Figure 3 shows an example of Λ_u^n for a two-dimensional binary random field.

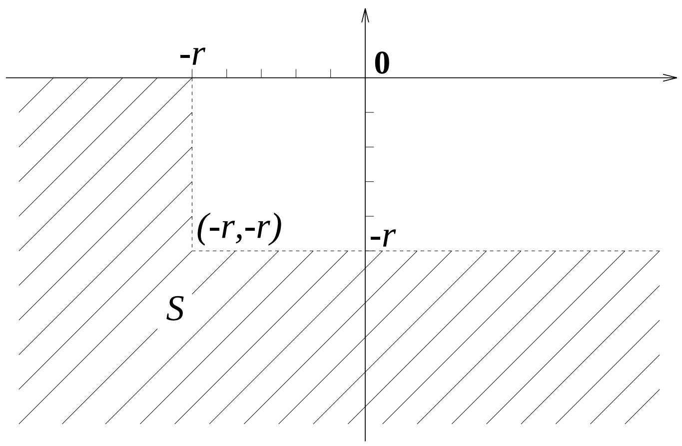

Figure 2: Example of the conditioning region for (dDC) in two dimensions.

Theorem 2. *Let $\{X_u : u \in \mathbf{Z}^d\}$ be a stationary ergodic random field with entropy rate $H > 0$. Then*

$$\liminf_{n\to\infty} \frac{\sum_{u\in C(n)} (\Lambda_u^n)^d}{n^d \log n^d} \geq \frac{1}{H} \quad \text{a.s.}$$

If the sequence $\{L_n^d/\log n^d\}$ is L^1-dominated, in particular, if (dDC) holds, then

$$\frac{\sum_{u\in C(n)} (\Lambda_u^n)^d}{n^d \log n^d} \to \frac{1}{H} \quad \text{a.s. and in } L^1.$$

Proof: Recall that $L_n^d/\log n^d \to 1/H$ almost surely as $n \to \infty$. If (dDC) holds, then $\sup_n L_n^d/\log n^d$ is integrable by Lemma 2. The stated results follow by application of Theorem 5 in the appendix to $g_{n,u} = L_n^d/\log n^d$. ∎

Lemma 2. *Let $\{X_u : u \in \mathbf{Z}^d\}$ be a stationary random field. If (dDC) holds, then*

$$P\{L_n > k\} \leq n^d \beta^{\lfloor k/r \rfloor^d}, \quad k \geq 1,$$

and the sequence of random variables $\{L_n^d/\log n^d\}$ is L^1-dominated:

$$E\left\{\sup_n \frac{L_n^d}{\log n^d}\right\} < \infty.$$

Proof: The proof of Lemma 2 parallels that of Lemma 1. To get the desired bound on the tail probability of L_n, recall that the event $\{L_n > k\}$ occurs if there is at least one match for a cube

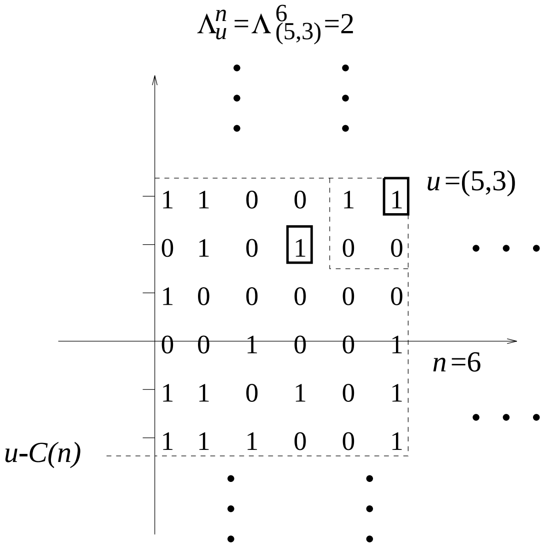

Figure 3: Example of Λ_u^n in two dimensions.

with volume k^d within a larger cube with volume n^d. The number of positions where a match can occur is less than n^d, and the probability of a match of volume k^d at any position may be bounded by the product of conditional probabilities at $\lfloor k/r \rfloor^d$ lattice points that are regularly spaced a distance r apart in each dimension. By (dDC), each term in the product is bounded by β since it is a weighted average of conditional probabilities given patterns that appeared earlier in the chain rule expansion, at least a distance r away in each dimension.

To prove L^1-domination, we follow the same steps as in the proof of Lemma 1, with the obvious modifications. If $\gamma = (-\log \beta)/(2r)^d$ then for $K \geq (4r)^d/d$, we obtain

$$\begin{aligned} E\left\{\sup_{n\geq 2} \frac{L_n^d}{\log n^d}\right\} &= \int_0^\infty P\left\{\sup_{n\geq 2} \frac{L_n^d}{\log n^d} > k\right\} dk \\ &\leq K + \sum_{n\geq 2} \int_K^\infty n^{d(1-\gamma k)}\, dk \\ &= K + \sum_{n\geq 2} \frac{n^{d(1-\gamma K)}}{\gamma \ln n^d}. \end{aligned}$$

This is finite if $K > (1 + d^{-1})/\gamma$. ∎

What is said in Theorem 1 about the Cesàro averages of $\Lambda_i^i/\log i$ can also be generalized to the random field case. For any nonnegative integer vector $u \in \mathbf{Z}_+^d$ let

$$L_u = \inf\{k \geq 1 : X_{-C(k)} \text{ does not occur in } (-u, \mathbf{0}] \text{ except at } \mathbf{0}\}.$$

Pick $0 < \epsilon < 1$, and observe that $\log(\epsilon n)^d \sim \log n^d$. Let $\pi(u) = \prod_j u_j$ denote the volume of the rectangle $[\mathbf{0}, u)$. If $u \in [\epsilon n, n)^d$ then $\log \pi(u) \sim \log n^d$ and $L_{\epsilon n} \le L_u \le L_n$. By (12),

$$\frac{L_u^d}{\log \pi(u)} \to \frac{1}{H} \quad \text{a.s.}$$

as $u \to \infty$ in the sector $\{u \in \mathbf{Z}_+^d : \min_j u_j \ge \epsilon \max_j u_j\}$. The shifted random variable $\Lambda_u^u(x) = L_u(T_u x)$ is equal to the minimum value of k such that the cube $X_{u-C(k)}$ fails to occur in the rectangle $(\mathbf{0}, u]$ except at position u.

Theorem 3. *Let $\{X_u : u \in \mathbf{Z}^d\}$ be a stationary ergodic random field with entropy rate $H > 0$. Then*

$$\liminf_{n\to\infty} \frac{1}{n^d} \sum_{\substack{u \in C(n) \\ \min_j u_j \ge 2}} \frac{(\Lambda_u^u)^d}{\log \pi(u)} \ge \frac{1}{H} \quad \text{a.s.} \tag{13}$$

If the integrability condition (16) is satisfied (in particular, if (dDC) holds), then

$$\frac{1}{n^d} \sum_{\substack{u \in C(n) \\ \min_j u_j \ge 2}} \frac{(\Lambda_u^u)^d}{\log \pi(u)} \to \frac{1}{H} \quad \text{a.s. and in } L^1, \tag{14}$$

$$\frac{\sum_{u \in C(n)} (\Lambda_u^u)^d}{n^d \log n^d} \to \frac{1}{H} \quad \text{a.s. and in } L^1. \tag{15}$$

Proof: Recall that $L_u^d / \log \pi(u) \to 1/H$ as $\inf(u_1, \ldots, u_d) \to \infty$ while u remains in the sector S^ϵ. Assertions (13) and (14) follow by setting $g_{n,u} = L_u^d / \log \pi(u)$ and invoking Theorem 5 in the appendix. Finally, (15) follows from (14) by a multivariate generalization of the technique in Theorem 1. ∎

Lemma 3. *Suppose $\{X_u\}$ is a stationary random field. If (dDC) holds, then*

$$P\{L_u < k\} \le \pi(u)\, \beta^{\lfloor k/r \rfloor^d}, \quad k \ge 1,$$

$$E\left\{ \sup_{u:\min_j u_j \ge 2} \frac{L_u^d}{\log \pi(u)} \right\} < \infty. \tag{16}$$

Proof: We mimic the proof of Lemma 2, but replace the volume n^d of the cube $[0, n)^d$ by the volume $\pi(u)$ of the rectangle $[\mathbf{0}, u)$. To prove L^1-domination, observe that

$$E\left\{ \sup_{u:\min_j u_j \ge K} \frac{L_u^d}{\log \pi(u)} \right\} = \int_0^\infty P\left\{ \sup_{u:\min_j u_j \ge K} \frac{L_u^d}{\log \pi(u)} > k \right\}$$

$$\leq \quad K + \sum_{u:\min_j u_j \geq K} \frac{\pi(u)^{1-\gamma K}}{\gamma \ln \pi(u)}$$

$$\leq \quad K + \frac{1}{\gamma} \prod_{1\leq i\leq d} \left(\sum_{u_j \geq K} u_j^{1-\gamma K} \right)$$

is finite when K is large, and that $L_u < K$ when $\min_j u_j < K$. ∎

V. Infinite Alphabets

In this section we generalize our results from Sections III and IV to processes and random fields with countably infinite alphabets. The proofs of Theorems 1, 2 and 3, as well as those of the corresponding Lemmas, carry over *verbatim* when $\mathcal{A}$ is countable. We only need to show that in the case of countably infinite alphabets, (12) remains valid. This, in turn, will follow from (11). In the following proposition we show that this is indeed the case.

Proposition. *Let $\{X_u\}$ be a stationary ergodic random field with entropy rate H, and assume that $E\{-\log P(X_0)\}$ is finite. Then,*

$$\frac{\log R_k^d}{k^d} \to H \quad \text{a.s.}$$

Proof: Assume, without loss of generality, that $\mathcal{A}$ is the set of nonnegative integers. For any fixed $m \geq 2$ we may lump the symbols $m, m+1, \ldots$ into a single super-symbol and define

$$X_u^{(m)} = \begin{cases} X_u & \text{if } 0 \leq X_u < m, \\ m & \text{if } X_u \geq m. \end{cases}$$

The random field $\{X_u^{(m)}\}$ is also stationary ergodic, and its entropy rate $H^{(m)}$ increases to the entropy rate H of the random field $\{X_u\}$ as $m \to \infty$ (see Chapter 7 of Pinsker [18] for a general discussion). Let $R_k^{(m)}$ be defined in terms of $\{X_u^{(m)}\}$ in the same way as R_k was defined in terms of $\{X_u\}$. Then $R_k \geq R_k^{(m)}$, so

$$\liminf_k \frac{1}{k^d} \log R_k^d \geq \lim_k \frac{1}{k^d} \log R_k^{(m)^d} = H^{(m)} \quad \text{a.s.}$$

Since $H^{(m)}$ increases to H as $m \to \infty$, we may conclude that

$$\liminf_k \frac{1}{k^d} \log R_k^d \geq H \quad \text{a.s.} \tag{17}$$

On the other hand, Ornstein and Weiss [16] have shown that in the finite alphabet case,

$$\limsup_k \frac{1}{k^d} \log R_k^d \leq H \quad \text{a.s.} \tag{18}$$

Their argument [16] is also valid in the infinite alphabet case, provided the Shannon-McMillan-Breiman theorem holds for the random field $\{X_u\}$. According to Ornstein and Weiss [17], this is indeed true if $E\{-\log P(X_0)\}$ is finite.

Combining (17) and (18) completes the proof. ∎

Appendix

Breiman [3] developed a generalized ergodic theorem and used it to prove pointwise convergence in what is now called the Shannon-McMillan-Breiman theorem. See also Barron [2] for a one-sided version and Algoet [1] for other applications. It turns out that Breiman's generalization is a special case of an older and more general ergodic theorem due to Maker [14]. We prove the one-sided version and then generalize to random fields. See also Theorem 7.5 on p. 66 of Krengel [13].

Theorem 4. (Maker) *Let T be a measure preserving transformation of a probability space $(\mathcal{X}, \mathcal{B}, P)$ and let $\mathcal{I}$ denote the σ-field of invariant events. Let $\{g_{n,i}\}_{n,i\geq 1}$ be a two-dimensional array of real-valued random variables.*

(a) If $E\{\inf_{n,i} g_{n,i}\} > -\infty$ and $g = \liminf_{n,i\to\infty} g_{n,i}$, then

$$\liminf_{n,i\to\infty} \frac{1}{n} \sum_{1\leq i\leq n} g_{n,i}(T^i x) \geq E\{g|\mathcal{I}\} \quad \text{a.s.}$$

(b) If $\sup_{n,i} |g_{n,i}|$ is integrable and $g_{n,i} \to g$ almost surely as $n, i \to \infty$, then

$$\frac{1}{n} \sum_{1\leq i\leq n} g_{n,i}(T^i x) \to E\{g|\mathcal{I}\} \quad \text{a.s. and in } L^1.$$

Proof: To prove (a), pick some integer $k \geq 0$ and consider the random variable

$$g_k = \inf_{n,i\geq k} g_{n,i}.$$

If $n \geq k$, then $g_{n,i} \geq g_1$ for $i \geq 1$ and $g_{n,i} \geq g_k$ for $i \geq k$, hence

$$\sum_{1\leq i\leq n} g_{n,i}(T^i x) \geq \sum_{1\leq i<k} g_1(T^i x) + \sum_{k\leq i\leq n} g_k(T^i x). \tag{19}$$

Dividing both sides by n and taking the $\liminf$ as $n \to \infty$, we see that

$$\liminf_n \frac{1}{n} \sum_{1 \le i \le n} g_{n,i}(T^i x) \ge E\{g_k | \mathcal{I}\} \quad \text{a.s.},$$

by the pointwise ergodic theorem. Now $g_1 = \inf_{n,i} g_{n,i}$ has expectation $E\{g_1\} > -\infty$ and g_k increases to g, so $E\{g_k|\mathcal{I}\}$ increases to $E\{g|\mathcal{I}\}$ by the monotone convergence theorem. Since $k \ge$ was arbitrary this completes the proof of (a).

The pointwise convergence in (b) follows by application of (a) to both $g_{n,i}$ and $-g_{n,i}$. To prove convergence in L^1, observe that by the mean ergodic theorem,

$$\frac{1}{n} \sum_{1 \le i \le n} g(T^i x) \to E\{g|\mathcal{I}\} \quad \text{in } L^1. \tag{20}$$

By assumption, $g_{n,i} \to g$ and $|g_{n,i} - g|$ is L^1-dominated, so $E|g_{n,i} - g| \to 0$ as $n, i \to \infty$ by the dominated convergence theorem. It follows by stationarity that

$$\begin{aligned} E\left|\frac{1}{n} \sum_{1 \le i \le n} g_{n,i}(T^i x) - g(T^i x)\right| &\le \frac{1}{n} \sum_{1 \le i \le n} E|g_{n,i}(T^i x) - g(T^i x)| \\ &= \frac{1}{n} \sum_{1 \le i \le n} E|g_{n,i} - g| \to 0 \quad \text{as } n \to \infty. \end{aligned} \tag{21}$$

The L^1 convergence follows from (20) and (21) and the triangle inequality. ∎

Maker's theorem can be generalized to random fields. A comprehensive treatment of ergodic theorems for random fields is provided in Chapter 6 of Krengel [13].

Let $\{T_u : u \in \mathbf{Z}_+^d\}$ be an abelian semigroup of measure preserving transformations of the probability space $(\mathcal{X}, \mathcal{B}, P)$. Given a random variable X, we consider the random field $\{X_u : u \in \mathbf{Z}_+^d\}$ where $X_u(x) = X(T_u x)$. If g is an integrable random variable then

$$\frac{1}{n^d} \sum_{u \in C(n)} g(T_u x) \to E\{g|\mathcal{I}\} \quad \text{a.s. and in } L^1,$$

where $\mathcal{I}$ is the σ-field of invariant events and $C(n) = [0, n)^d$ is the cube with side length n. Given $0 < \epsilon < 1$, the cube $C(n)$ can be partitioned into the cube $C^\epsilon(n) = [\epsilon n, n)^d$ and its complement $D^\epsilon(n) = C(n) \backslash C^\epsilon(n)$. Note that $C^\epsilon(n)$ is contained in the sector

$$S^\epsilon = \{u \in \mathbf{Z}_+^d : \min_j u_j \ge \epsilon \max_j u_j\}.$$

For any integer $n \geq 0$ and any nonnegative integer vector $u \in \mathbf{Z}_+^d$ let $g_{n,u}$ be a real valued random variable defined on $(\mathcal{X}, \mathcal{B}, P)$. For $0 < \epsilon < 1$ and $k \geq 1$ let

$$g_k^\epsilon = \inf_{\substack{\inf(n,u_1,\ldots,u_d)\geq k \\ u\in S^\epsilon}} g_{n,u}.$$

As k increases, the infimum is taken over smaller sets and increases to $g^\epsilon = \lim_k g_k^\epsilon$. If now $\epsilon \downarrow 0$ then g^ϵ decreases to a limit $g = \lim_{\epsilon\downarrow 0} g^\epsilon$.

Theorem 5. *(a) Suppose the family $\{g_{n,u} : n \in \mathbf{Z}_+,\, u \in \mathbf{Z}_+^d\}$ is bounded below by an integrable random variable g_0, and let $g = \lim_{\epsilon\downarrow 0} \lim_k g_k^\epsilon$ as above. Then*

$$\liminf_{n\to\infty} \frac{1}{n^d} \sum_{u\in C(n)} g_{n,u}(T_u x) \geq E\{g|\mathcal{I}\} \quad \text{a.s.}$$

(b) Suppose that for any $0 < \epsilon < 1$, $g_{n,u} \to g$ almost surely as $n, u_1, \ldots, u_d \to \infty$ while $u = (u_1, \ldots, u_d)$ stays in the sector S^ϵ. If the family $\{g_{n,u}\}$ is L^1-dominated, then

$$\frac{1}{n^d} \sum_{u\in C(n)} g_{n,u}(T_u x) \to E\{g|\mathcal{I}\} \quad \text{a.s. and in } L^1.$$

Proof: Pick some $0 < \epsilon < 1$ and $k \geq 1$ and observe that for large n, $n\epsilon \geq k$ and

$$\sum_{u\in C(n)} g_{n,u}(T_u x) \geq \sum_{u\in D^\epsilon(n)} g_0(T_u x) + \sum_{u\in C^\epsilon(n)} g_k^\epsilon(T_u x).$$

Dividing by n^d and taking the $\liminf$ as $n \to \infty$, we obtain

$$\liminf_n \frac{1}{n^d} \sum_{u\in C(n)} g_{n,u}(T_u x) \geq (1-(1-\epsilon)^d) E\{g_0|\mathcal{I}\} + (1-\epsilon)^d E\{g_k^\epsilon|\mathcal{I}\} \quad \text{a.s.}$$

The right hand side increases to $(1-(1-\epsilon)^d)E\{g_0|\mathcal{I}\} + (1-\epsilon)^d E\{g^\epsilon|\mathcal{I}\}$ as $k \to \infty$. Letting $\epsilon \searrow 0$ yields (a). Part (b) can also be proved as in the one-dimensional case. ∎

Acknowledgements

The Jane Austen novels *Mansfield Park*, *Northanger Abbey* and *Persuasion* we got from Project Gutenebrg at `http://192.76.144.75/books/gutenberg`. The last novel, *Sense and Sensibility*, we got from the Educational Resources of the University of Maryland at College Park, at `http://www.inform.umd.edu:8080/EdRes/Topic/WomenStudies/ReadingRoom`.

We wish to thank the anonymous referees for several interesting comments that greatly improved the presentation of our paper and for bringing to our attention reference [17], and the Associate Editor, M. Feder, for suggesting that we include a section with numerical examples of the performance of our algorithms. Also we would like to thank Prof. J. Ziv for pointing out to us that (DC) holds for any process observed through a memoryless channel with arbitrarily small noise.

References

[1] P.H. Algoet, "The strong law of large numbers for sequential decisions under uncertainty," *IEEE Trans. Inform. Theory,* vol. 40, pp. 609–634, May 1994.

[2] A.R. Barron, "The strong ergodic theorem for densities: generalized Shannon-McMillan-Breiman theorem," *Ann. Probab.,* vol. 13, pp. 1292–1303, Nov. 1985.

[3] L. Breiman, "The individual ergodic theorem of information theory," *Ann. Math. Statist.,* vol. 28, pp. 809–811, 1957. Correction: *ibid.,* vol. 31, pp. 809–810, 1960.

[4] S. Chen and J.H. Reif, "Using difficulty of prediction to decrease computation: Fast sort, priority queue and convex hull on entropy bounded inputs," in *Proceedings of the 34th Symposium on Foundations of Computer Science*, Los Alamitos, California: IEEE Computer Society Press, 1993, pp. 104–112.

[5] S. Chen and J.H. Reif, "Fast pattern matching for entropy bounded text," in *Proceedings of DCC'95 Data Compression Conference*, Snowbird, UT, USA: IEEE Computer Society Press, 1995, pp. 282–291.

[6] T.M. Cover and J. Thomas, *Elements of Information Theory.* New York: Wiley, 1991.

[7] W. Doeblin, "Sur le propriétés asymptotiques de mouvement régis par certains types de chaines simples," *Bull. Math. Soc. Roum. Sci.*, vol. 39, nr. 1 pp. 57–115, and vol. 39, nr. 2 pp. 3–61, 1937.

[8] M. Farach, M. Noordewier, S. Savari, L. Shepp, A.J. Wyner and J. Ziv. "On the entropy of DNA: Algorithms and measurements based on memory and rapid convergence," in *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms,* Philadelphia: Society for Industrial and Applied Mathematics, 1995, pp. 48-57.

[9] P. Grassberger, "Estimating the information content of symbol sequences and efficient codes," *IEEE Trans. Inform. Theory,* vol. 35, pp. 669–675, May 1989.

[10] P. Juola, "What can we do with small corpora: Document categorization via cross-entropy," *Comparative Linguistics,* to appear, SimCat '97.

[11] I. Kontoyiannis and Yu.M. Suhov, "Prefixes and the entropy rate for long-range sources," in *Probability Statistics and Optimization*, F. P. Kelly, ed. Chichester, England: Wiley, 1994, pp. 89–98.

[12] I. Kontoyiannis and Yu.M. Suhov, "Stationary entropy estimation via string matching," In *Proceedings of the Data Compression Conference DCC 96,* Snowbird, UT, April 1996.

[13] U. Krengel, *Ergodic Theorems*. Berlin: de Gruyter, 1985.

[14] P.T. Maker, "The ergodic theorem for a sequence of functions," *Duke Math. J.*, vol. 6, pp. 27–30, 1940.

[15] D.S. Ornstein and B. Weiss, "Entropy and data compression schemes," *IEEE Trans. Inform. Theory,* vol. 39, pp. 78–83, Jan. 1993.

[16] D.S. Ornstein and B. Weiss, "Entropy and recurrence rates for stationary random fields," preprint; also presented at *IEEE Int. Symp. Inform. Theory,* Trondheim, Norway, Jun. 27–Jul. 1, 1994.

[17] D.S. Ornstein and B. Weiss, "The Shannon-McMillan-Breiman theorem for countable partitions," unpublished manuscript.

[18] M.S. Pinsker, *Information and Information Stability of Random Variables and Processes.* Izd. Akad. Nauk SSSR, Moscou, 1960. Translated and edited by A. Feinstein. San Francisco: Holden-Day, 1964.

[19] B. Pittel, "Asymptotic growth of a class of random trees," *Ann. Probab.*, vol. 13, nr. 2, pp. 414–427, 1985.

[20] A.N. Quas, "An entropy estimator for a class of infinite alphabet processes." Technical Report 95–3, Statistical Laboratory, University of Cambridge. Cambridge, England, May 1995.

[21] P.C. Shields, "Universal redundancy rates do not exist," *IEEE Trans. Inform. Theory,* vol. 39, nr. 2, pp. 520–524, 1993.

[22] P.C. Shields, "Entropy and prefixes," *Ann. Probab.*, vol. 20, pp. 403–409, 1992.

[23] P.C. Shields and B. Weiss, "Universal redundancy rates for the class of B-processes do not exist," *IEEE Trans. Inform. Theory,* vol. 41, nr. 2, pp. 508–512, 1995.

[24] W. Szpankowski, "Asymptotic properties of data compression and suffix trees," *IEEE Trans. Inform. Theory,* vol. 39, pp. 1647–1659, Sept. 1993.

[25] W.J. Teahan and J.G. Cleary, "Models of English text," in *Proceedings of DCC'97 Data Compression Conference*, Snowbird, UT, USA: IEEE Computer Society Press, 1997, pp. 12–21.

[26] W.J. Teahan and J.G. Cleary, "The entropy of English using PPM-based models," in *Proceedings of DCC'96 Data Compression Conference*, Snowbird, UT, USA: IEEE Computer Society Press, 1996, pp. 53–62.

[27] A.J. Wyner, "Entropy and patterns," In *Proceedings IEEE Inform. Theory Workshop,* Haifa, Israel, June 1996.

[28] A.J. Wyner, "The redundancy and distribution of the phrase lengths of the fixed-database Lempel-Ziv algorithm," *IEEE Trans. Inform. Theory,* vol. 43, pp. 1452–1464, Sept. 1997.

[29] A.D. Wyner and J. Ziv, "Some asymptotic properties of entropy of a stationary ergodic data source with applications to data compression," *IEEE Trans. Inform. Theory,* vol. 35, pp. 1250–1258, Nov. 1989.

[30] J. Ziv and A. Lempel, "A universal algorithm for sequential data compression," *IEEE Trans. Inform. Theory,* vol. 23, pp.337–343, May 1977.

[31] J. Ziv and A. Lempel, "Compression of individual sequences via variable rate coding," *IEEE Trans. Inform. Theory,* vol. 24, pp. 530–536, Sept. 1978.

Captions of tables and figures

List of tables

List of figures