TWO CONSISTENT ENTROPY ESTIMATES
FOR STATIONARY PROCESSES AND RANDOM FIELDS

BY
I. KONTOYIANNIS
AND
P. H. ALGOET
STANFORD UNIVERSITY
AND
YU. M. SUHOV
CAMBRIDGE UNIVERSITY

TECHNICAL REPORT NO. 91
APRIL 1996

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION
GRANT NCR-9205663

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
TWO CONSISTENT ENTROPY ESTIMATES
FOR STATIONARY PROCESSES AND RANDOM FIELDS

BY
I. KONTOYIANNIS
AND
P. H. ALGOET
STANFORD UNIVERSITY
AND
YU. M. SUHOV
CAMBRIDGE UNIVERSITY

TECHNICAL REPORT NO. 91
APRIL 1996

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION
GRANT NCR-9205663

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
Two Consistent Entropy Estimates for Stationary Processes and Random Fields

I. Kontoyiannis  P. H. Algoet  Yu. M. Suhov

Stanford University  Stanford University  Cambridge University

April 25, 1996

Abstract — We propose two new 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 generalize our results to countable alphabet processes and random fields.

Index Terms — Entropy rate, pattern matching, recurrence times, universal data compression.

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 encoded, and several universal data compression algorithms are known for which the per-symbol description length approaches the entropy rate in the pointwise sense. In particular, the Lempel-Ziv [21] algorithm attains the entropy lower bound when it is applied to almost every realization of a stationary ergodic source.

Wyner and Ziv [19], 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 recurrence times. They raised many stimulating questions about the recurrence time, $R_k$, until the initial $k$-block produced by the process is repeated for the first time, and the waiting-time, $W_k$, until a $k$-block from one realization first appears in an independent realization of the same process. Ornstein and Weiss [11] formally proved that $R_k$ grows exponentially, with limiting rate almost surely equal to the entropy rate of the underlying process. Shields [17] argued that similar results hold for the waiting time $W_k$, but only if additional mixing conditions are imposed.

1I. Kontoyiannis and P. Algoet are with the Information Systems Laboratory (Durand 141B), 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, DAMM, Cambridge University, and also with the Institute for Problems in Information Transmission, Moscow. Email: y.m.suhov@statslab.cam.ac.uk.

2I. 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.

3A preliminary version of this paper was presented at the 1996 Data Compression Conference, Snowbird, Utah, April 1996.
These recurrence time results can equivalently be stated in terms of prefix match lengths along a sample path. This is the point of view of Grassberger [7], who suggested an entropy estimator based on average match lengths. Shields [16] proved the consistency of Grassberger’s estimator for independent identically distributed (i.i.d.) processes and mixing Markov processes. Kontoyiannis and Suhov [9] extended this to a wider class of stationary processes, and recently Quas [14] extended it further to certain processes with infinite alphabets and to random fields.

In this paper we introduce two new entropy estimators ((1) and (2) below) that are formally similar to the one suggested by Grassberger but which, due to their stationary nature, are much easier to analyze. We also generalize the results to processes with countably infinite alphabets and to random fields.

We consider a random process \( \{X_i\} \) with values in a finite set \( \mathcal{A} \), called the alphabet. A process realization is an element \( x = (x_i)_{i \in \mathbb{Z}} \) of the two-sided sequence space \( \mathcal{X} = \mathcal{A}^\mathbb{Z} \), and \( X_i(x) = x_i \) is its \( i \)th coordinate. If \( i \leq j \) then \( X^j_\hat{1} \) denotes the string \( (X_i, X_{i+1}, \ldots, X_j) \), otherwise it denotes the empty string. The process distribution is a probability measure \( P \) on the Borel \( \sigma \)-field on \( \mathcal{X} \). We assume that \( P \) is invariant under the usual shift transformation \( T x = (x_{i+1})_{i \in \mathbb{Z}} \), so that \( \{X_i\} \) is a stationary process. We also assume ergodicity.

For \( n \geq 1 \) let \( L_n \) denote the minimum length \( k \) of the string \( X^k_0 \) that starts at time 0 and does not appear as a continuous substring within the past \( X^{-n}_{-n} \). Alternatively, \( L_n \) is obtained by adding 1 to the longest match length:

\[
L_n = 1 + \max\{l : 0 \leq l \leq n, \ X^l_0 = X^{-l}_{j-1} \ \text{for some} \ l \leq j \leq n\}.
\]

From the work of Wyner and Ziv [19] and Ornstein and Weiss [11], we know that 1

\[
\frac{L_n}{\log n} \to \frac{1}{H} \quad \text{a.s.},
\]

where \( H \) is the entropy rate. Given an instant \( i \) and a positive integer \( n \), our main quantity of interest is \( \Lambda^n_i(x) = L_n(T^i x) \), the length of the shortest substring \( X^{i+k}_{i+k-1} \) starting at position \( i \) that does not appear as a contiguous substring of the previous \( n \) symbols \( X^{i-1}_{i-n} \). One may interpret \( \Lambda^n_i \) as the length of the next phrase to be encoded by the sliding-window Lempel-Ziv algorithm [20], when the window size is \( n \). Similarly, \( \Lambda^+_i \) is the length of the phrase that would be encoded next by the Lempel-Ziv algorithm [21] with knowledge of the past \( X^{i-1}_0 \). Notice that \( \Lambda^+_i \) is determined by the one-sided sequence \( \{X_j : j \geq 0\} \), whereas \( \Lambda^n_i \) generally depends on observations \( X_j \) with both positive and negative indices \( j \).

Since \( \Lambda^n_i(x) = L_n(T^i x) \), it follows by stationarity that, for any fixed index \( i \), \( \Lambda^n_i / \log n \) will converge to \( 1/H \) with probability one as \( n \to \infty \). But the rate of convergence seems slow in practice. In order to make more efficient use of the data, we propose to estimate \( 1/H \) by the Cesàro averages of the normalized match lengths \( \Lambda^n_i / \log n \) or \( \Lambda^+_i / \log i \), for \( i \) between 1 and \( n \). The main result of the paper is that if we impose an additional mixing condition, then we can invoke a generalized ergodic theorem due to Maker [10] and conclude that the Cesàro averages are pointwise consistent estimates for \( 1/H \).

---

1Logarithms are to base 2 throughout the paper.
Theorem. Let \( \{X_i\} \) be a stationary ergodic process with entropy rate \( H \). Then

\[
\frac{1}{n} \sum_{1 \leq i \leq n} \frac{\Lambda_i^n}{\log n} \to \frac{1}{H}, \quad \text{a.s. and in } L^1,
\]

\[
\frac{1}{n} \sum_{2 \leq i \leq n} \frac{\Lambda_i^n}{\log i} \to \frac{1}{H} \quad \text{a.s. and in } L^1,
\]

provided the following mixing condition holds:

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

\[
P\{X_0 = x_0 | X_{-\infty} \} \leq \beta, \quad \text{for all } x_0 \in \mathcal{A}.
\]

(DC) ensures that \( \sup_n L_n / \log n \) has finite expectation. Without this integrability condition, \( 1/H \) is still an asymptotic lower bound for the estimates in (1) and (2).

By stationarity, the quantities \( \Lambda_i^n / \log n \) always converge to \( 1/H \) in probability, and if (DC) holds then our Theorem says that their Cesàro means also converge to \( 1/H \), with probability one. However, Pittel [15] and Szpankowski [18] have shown that the quantities \( \Lambda_i^n / \log n \) themselves keep fluctuating. They interpret \( \Lambda_i^n \) as the length of a feasible path in a suffix tree and under certain mixing conditions they identify two natural constants \( H_1 \) and \( H_2 \) such that \( H_1 > H > H_2 \) and

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

The Doeblin condition was originally introduced in the analysis of Markov chains, where it is typically used to prove the existence and uniqueness of an invariant distribution and to show that convergence to the invariant distribution occurs with exponential rate. In the context of this paper, (DC) was first introduced by Kontoyiannis and Suhov [9], where its properties are discussed in greater detail. Here we note that (DC) holds for i.i.d. processes, for irreducible aperiodic Markov chains, and also for certain non-Markov processes. Our present formulation of (DC), first introduced by Quas [14], is equivalent to that in [9] when the alphabet is finite, but it has the advantage of being applicable also to processes with countably infinite alphabets.

An interesting property of (DC) is that it is satisfied by any stationary ergodic process that is observed through a discrete memoryless channel which transforms letters of the alphabet to other letters with positive but arbitrarily small probability. In particular, 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 \( \beta = \max\{p, 1-p\} \).

Entropy estimators similar to the one in (1) have appeared in the literature [4], [6], where they were applied to experimental data in order to determine the entropy rate of the underlying process. So apart from the theoretical interest of (1) and (2), part of our motivation was to provide a more precise analysis for the use of these methods in practice.

In Section II we state and prove our main result for finite-alphabet processes. We generalize this result to processes with an infinite alphabet in Section III, and to random fields in Section IV. In the Appendix we present a simplified proof of Maker’s generalized ergodic theorem and some extensions that are used in Sections II and IV.
II. Main Results

We first prove the main result for a stationary ergodic process with values in a finite alphabet \(\mathcal{A}\). We invoke Maker's generalized ergodic theorem, which requires a certain integrability condition. We then verify that the integrability condition follows from (DC).

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

(a) \[ \liminf_n \frac{\sum_{1 \leq i \leq n} \Lambda_i^n}{n \log n} \geq \frac{1}{H} \quad \text{a.s.}, \]

(b) \[ \liminf_n \frac{1}{n} \sum_{2 \leq i \leq n} \frac{\Lambda_i^i}{\log i} \geq \frac{1}{H} \quad \text{a.s.} \]

If the random variables \(L_n/\log n\) are \(L^1\)-dominated (in particular, if (DC) holds), then

(c) \[ \lim_n \frac{\sum_{1 \leq i \leq n} \Lambda_i^n}{n \log n} = \frac{1}{H} \quad \text{a.s. and in } L^1, \]

(d) \[ \lim_n \frac{1}{n} \sum_{2 \leq i \leq n} \frac{\Lambda_i^i}{\log i} = \frac{1}{H} \quad \text{a.s. and in } L^1, \]

(e) \[ \lim_n \frac{\sum_{1 \leq i \leq n} \Lambda_i^i}{n \log n} = \frac{1}{H} \quad \text{a.s. and in } L^1. \]

**Proof:** Recall that \(L_n/\log n \rightarrow 1/H\) with probability one. In Lemma 1 below, we prove that (DC) implies integrability of \(\sup_n L_n/\log n\). We now invoke Maker's generalized ergodic theorem. Assertions (a) and (c) follow from Theorem 4 in the appendix by setting \(g_{n,i} = L_n/\log n\), whereas (b) and (d) follow by setting \(g_{n,i} = L_i/\log i\).

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

\[ \frac{1}{n} \left( l_2 \log 2 + \ldots + l_n \log n \right) \rightarrow \frac{1}{H} \quad \text{a.s. and in } L^1. \]  

(3)

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

\[ 0 \leq \frac{l_i}{\log i} - \frac{l_i}{\log n} \leq \delta_n \frac{l_i}{\log i}, \]

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

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

\[ \leq \epsilon \frac{1}{en} \left( l_2 \log 2 + \ldots + l_en \log(en) \right) + \frac{\delta_n}{n} \left( \frac{l_2 \log 2 + \ldots + l_n \log n}{\log n} \right). \]

Letting \(n \rightarrow \infty\) and using (3), we may conclude that for any \(\epsilon > 0\),

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

The limit must vanish, and this fact in combination with (3) yields (e), since \(l_i = \Lambda_i^i\). \(\blacksquare\)

It remains to verify that (DC) implies integrability of \(\sup_n L_n/\log n\).
Lemma 1. Let \( \{X_i\} \) be a stationary process. If the Doeblin condition holds, then
\[
P\{L_n > k\} \leq n\beta^{[k/r]}, \quad k \geq 1, \tag{4}
\]
and the random variables \( L_n / \log n \) are \( L^1 \)-dominated:
\[
E \left\{ \sup_n \frac{L_n}{\log n} \right\} < \infty.
\]

Proof: 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 \( X_{-n} \), for some \( k \leq j \leq n \). Therefore,
\[
P\{L_n > k\} \quad \leq \quad \sum_{k \leq j \leq n} P\{X_0^{k-1} = X_{-j}^{-j+k-1}\}
\]
\[
\leq \quad (n-k+1) \max_{k \leq j \leq n} P\{X_0^{k-1} = X_{-j}^{-j+k-1}\}. \tag{5}
\]
Write \( x^k \) for \( (x_0, \ldots, x_{k-1}) \) and observe that
\[
P\{X_0^{k-1} = X_{-j}^{-j+k-1}\} = \sum_{x^k \in \mathbb{A}^k} P\{X_0^{k-1} = x^k | X_{-j}^{-j+k-1} = x^k\} P\{X_{-j}^{-j+k-1} = x^k\}.
\]
Using (DC),
\[
P\{X_0^{k-1} = x^k | X_{-j}^{-j+k-1} = x^k\} \leq P\{X_{st} = x_{st}, 0 \leq t < [k/r] | X_{-j}^{-j+k-1} = x^k\}
\]
\[
\leq \prod_{t=0}^{[k/r]} P\{X_{st} = x_{st} | X_{-j}^{-j+k-1} = x^k, X_{rs} = x_{rs}, 0 \leq s < t\}
\]
\[
\leq \beta^{[k/r]}.
\]
It follows that \( P\{X_0^{k-1} = X_{-j}^{-j+k-1}\} \leq \beta^{[k/r]} \). Substituting this bound in (5) above, we obtain (4). 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 \( [k \log n]/r \geq (k \log n)/r - 1 - 1/r \geq (k \log n)/(2r) \). Consequently, for \( K \geq 4r \),
\[
E \left\{ \sup_{n \geq 2} \frac{L_n}{\log n} \right\} = \int_0^\infty P \left\{ \sup_{n \geq 2} \frac{L_n}{\log n} > k \right\} \, dk
\]
\[
\leq K + \int_K^\infty \sum_{n \geq 2} P\{L_n > k \log n\} \, dk
\]
\[
\leq K + \sum_{n \geq 2} \int_K^\infty n \beta^{[k \log n]/r} \, dk
\]
\[
\leq K + \sum_{n \geq 2} \int_K^\infty n \beta^{(k \log n)/(2r)} \, dk
\]
\[
= K + \sum_{n \geq 2} \int_K^\infty n^{1-\gamma k} \, dk
\]
\[
= K + \sum_{n \geq 2} \frac{n^{1-\gamma K}}{\gamma \ln n}.
\]
The sum is finite if the constant \( K \) is chosen so that \( K > 2/\gamma \). \( \blacksquare \)
We do not know if (DC) is necessary for the $L^1$-domination result of Lemma 1, but we believe that some mixing condition is required. Our proof requires that the tails of $L_n$ vanish exponentially so that the double sum $\sum_k \sum_n P\{L_n / \log n > k\}$ is finite.

III. Infinite Alphabets

The proofs of Theorem 1 and Lemma 1 carry over verbatim when $A$ is countable. It only remains to show that for a stationary ergodic process $\{X_i\}$ with finite entropy rate $H$, 

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

This result was conjectured in [19] and formally established in [11] for the finite-alphabet case. Here we extend the proof of [11] to the countable-alphabet case. We assume $H < \infty$.

**Proof:** The recurrence time $R_k$ is defined as the minimum length $n$ such that the string $X_0^{k-1}$ can be found within the database $X_{-n}^{-1}$:

$$R_k = \inf\{n \geq k : X_{-n}^{n+k-1} = X_0^{k-1}\}.$$

As observed in [19], $L_n > k$ iff $R_k \leq n$ iff $X_0^{k-1}$ recurs within $X_{-n}^{-1}$. In view of this duality between $L_n$ and $R_k$, it suffices to prove that

$$\frac{1}{k} \log R_k \to H \quad \text{a.s.}$$

Assume, without loss of generality, that $A$ 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

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

The process $\{X_i^{(m)}\}$ is also stationary ergodic, and its entropy rate $H^{(m)}$ increases to the entropy rate $H$ of the process $\{X_i\}$ as $m \to \infty$ (see Chapter 7 of Pinsker [13] for a general proof). Let $R_k^{(m)}$ be defined in terms of the process $\{X_i^{(m)}\}$ in the same way as $R_k$ was defined in terms of $\{X_i\}$. Then $R_k \geq R_k^{(m)}$, so

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

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

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

It remains to show that also

$$\limsup_k \frac{1}{k} \log R_k \leq H \quad \text{a.s.} \quad (6)$$

To prove this, assume that $H < \infty$ and pick constants $D$ and $C$ such that $H < D < C$. For each $k \geq 1$ consider the event

$$\Delta_k = \{P(X_{-k}^{-1}) > 2^{-Dk}, R_k > 2^{Ck}\}.$$
Let $A_k$ denote the set of sequences $x_{-k}^{-1}$ such that $P(x_{-k}^{-1}) > 2^{-Dk}$, and for each $x_{-k}^{-1} \in A_k$ define the cylinder set

$$\Gamma(x_{-k}^{-1}) = \{X_{-k}^{-1} = x_{-k}^{-1}, R_k > 2^{Ck}\}.$$ 

If $\Gamma = \Gamma(x_{-k}^{-1})$ for some $x_{-k}^{-1} \in A_k$, then $\{T_i^k \Gamma : 0 \leq i < 2^{Ck}/k\}$ is a collection of disjoint events. (Indeed, if $x \in T_i^k \Gamma \cap T_j^k \Gamma$ for some $0 \leq i < j < 2^{Ck}/k$, then we would have $R_k(T^{-j^k}x) \leq (j - i)k \leq jk < 2^{Ck}/k$, contradicting the definition of $\Gamma$.) These $2^{Ck}/k$ disjoint events are obtained from $\Gamma$ by repeated application of the measure preserving transformation $T$, so they all have the same probability, bounded by $k2^{-Ck}$. Observe that $A_k$ has cardinality $\|A_k\| \leq 2^{Dk}$ since $P\{X_{-k}^{-1} = x_{-k}^{-1}\} \geq 2^{-Dk}$ for all $x_{-k}^{-1} \in A_k$. Since

$$\Delta_k = \bigcup_{x_{-k}^{-1} \in A_k} \Gamma(x_{-k}^{-1}),$$

and $P(\Gamma(x_{-k}^{-1})) \leq k2^{-Ck}$ for all $x_{-k}^{-1} \in A_k$, we may conclude that

$$P(\Delta_k) \leq \|A_k\| \max_{x_{-k}^{-1} \in A_k} P(\Gamma(x_{-k}^{-1})) \leq k2^{(D-C)k}$$

and consequently

$$\sum_{k \geq 1} P(\Delta_k) \leq \sum_{k \geq 1} k2^{(D-C)k} < \infty.$$ 

It follows by the Borel-Cantelli lemma that eventually for large $k$, either $P(X_{-k}^{-1}) \leq 2^{-Dk}$ or $R_k \leq 2^{Ck}$. Since $P(X_{-k}^{-1}) > 2^{-Dk}$ eventually by the Shannon-McMillan-Breiman theorem for the time-reversed process $\{X_{-i}\}$, we may conclude that $R_k \leq 2^{Ck}$ for large $k$. This conclusion holds for arbitrary $C > H$, so (6) holds as claimed. \hfill \blacksquare

**IV. Generalization to random fields**

In this section we generalize our results to random fields on the integer lattice $\mathbb{Z}^d$. Such a random field is a family of random variables $\{X_u : u \in \mathbb{Z}^d\}$ 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 $\mathcal{X} = \prod \{A_u : u \in \mathbb{Z}^d\}$, where each $A_u$ is a copy of $\mathcal{A}$. If $x = \{x_u : u \in \mathbb{Z}^d\}$ is a realization in $\mathcal{X}$, then $X_u(x) = x_u$ is the coordinate at position $u$. For a subset $U \subseteq \mathbb{Z}^d$ let $X_U(x) = (x_u)_{u \in U}$. For any vector $v \in \mathbb{Z}^d$, 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 $u, w \in \mathbb{Z}^d$ let $\{u, w\} = \{v \in \mathbb{Z}^d : u_j \leq v_j < w_j \text{ for all } j\}$. The $d$-dimensional cube with side $k$ is defined, for any integer $k \geq 1$, as the cartesian product $[0, k)^d$:

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

We define the recurrence time $R_k$ for any integer $k \geq 1$ as the minimum value of $n$ such that the block $X_{-C(k)}$ occurs at some position other than position $0$ inside $-C(n)$:

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

The dual quantity here is the match length $L_n$, defined as the minimum value of $k$ such that $X_{-C(k)}$ does not occur anywhere in $-C(n)$ except at position $0$:

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

Figure 1 shows an example of $L_n$ for a binary random field in two dimensions.

Applying a result of Ornstein and Weiss [12] to the reflected field $\{X_{-u}\}$, we see that

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

Since $L_n \leq k$ iff $R_k > n$, it follows immediately that

$$\frac{L_n^d}{\log n^d} \to \frac{1}{H} \quad \text{a.s.} \quad (7)$$

If the Shannon-McMillan-Breiman theorem were known to hold for random fields with values in countable set $A$, then assertion (7) and the conclusions that we derive from it in Theorem 2 and Theorem 3 below would also hold. 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 $\{X_u : u \in Z^d\}$. It is expressed exactly like (DC), but the past $X_{-\infty}$ is defined differently in the $d$-dimensional case.

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

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

In the $d$-dimensional case, we define $X_{-\infty}^{-r}$ as the family of random variables $X_u$ with index vectors $u = (u_1, \ldots, u_d)$ such that $([u/r], \ldots, [u/r])$ lexicographically precedes $0 = (0, \ldots, 0)$ in $Z^d$. In particular, if $r = 1$ then $X_{-\infty}^{-r} = \{X_u : u \in Z^d, u < 0\}$ 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 $r = 5$.

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

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

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 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 $[k/r]^d$ lattice points that are regularly spaced a distance $r$ apart in each dimension. By (dDC), each term in the product is bounded by $\beta$ 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

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

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

We now study the analog of Theorem 1 for random fields. For $n \geq 1$ and any vector $u = (u_1, \ldots, u_d)$ in $\mathbb{Z}^d$ 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 $\Lambda_u^n$ for a two-dimensional binary random field.

**Theorem 2.** Let $\{X_u : u \in \mathbb{Z}^d\}$ 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

\[
\frac{\sum_{u \in C(n)} (\Lambda_u^n)^d}{n^d \log n^d} \to \frac{1}{H} \quad \text{a.s. and in } L^1.
\]
Figure 3: Example of $\Lambda_u^n$ in two dimensions.

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 $g_{n,u} = L_n^d/\log n^d$.

Theorem 2 is about averages of the volumes $(\Lambda_u^n)^d$. A similar theorem holds for averages of the side lengths $\Lambda_u^n$. In fact, for any finite $\delta > 0$ one can deduce results analogous to Theorem 2 regarding the Cesàro mean convergence of the quantities $(\Lambda_u^n)^\delta$. Indeed, $L_n^\delta/(\log n^d)^{\delta/d} \to 1/H^{\delta/d}$ almost surely and by the one-sided version of Maker’s theorem,

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

If (dDC) holds, then $\sup_n L_n^\delta/(\log n^d)^{\delta/d}$ is integrable by a straightforward generalization of the argument of Lemma 2. By Maker’s theorem,

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

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

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

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 $[0, u)$. If $u \in \epsilon n, n]^d$ then $\log \pi(u) \sim \log n^d$ and $\epsilon n \leq L_u \leq n$. By (7),

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

as $u \to \infty$ in the sector $\{u \in \mathbb{Z}_+^d : \min_j u_j \geq \epsilon \max_j u_j\}$. 

10
Lemma 3. Suppose \( \{X_u\} \) is a stationary random field. If \( (dDC) \) holds, then
\[
P\{L_u < k\} \leq \pi(u) \beta^{[k/r]^d}, \quad k \geq 1,
\]
\[
E \left\{ \sup_{u: \min_j u_j \geq 2} \frac{L_u^d}{\log \pi(u)} \right\} < \infty. \tag{8}
\]

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 \([0, u]\). To prove \( L^1 \)-domination, observe that
\[
E \left\{ \sup_{u: \min_j u_j \geq K} \frac{L_u^d}{\log \pi(u)} \right\} = \int_0^\infty P \left\{ \sup_{u: \min_j u_j \geq K} \frac{L_u^d}{\log \pi(u)} > k \right\}
\leq K + \sum_{u: \min_j u_j \geq K} \frac{\pi(u)^{1-\gamma K}}{\gamma \ln \pi(u)}
\leq K + \frac{1}{\gamma} \prod_{1 \leq i \leq d} \left( \sum_{u_j \geq K} u_j^{1-\gamma K} \right)
\]
is finite when \( K \) is large, and that \( L_u < K \) when \( \min_j u_j < K \). \qed

The shifted random variable \( \Lambda_u^w(x) = L_u(T_u x) \) is equal to the minimum value of \( k \) such that the cube \( X_{u-C(k)} \) fails to occur in the rectangle \((0, u)\) except at position \( u \).

Theorem 3. Let \( \{X_u : u \in \mathbb{Z}^d\} \) be a stationary ergodic random field with entropy rate \( H > 0 \). Then
\[
\liminf_{n \to \infty} \frac{1}{n^d} \sum_{u \in C(n) \atop \min_j u_j \geq 2} \frac{(\Lambda_u^w)^d}{\log \pi(u)} \geq \frac{1}{H} \quad \text{a.s.} \tag{9}
\]
If the integrability condition (8) is satisfied (in particular, if \( (dDC) \) holds), then
\[
\frac{1}{n^d} \sum_{u \in C(n) \atop \min_j u_j \geq 2} \frac{(\Lambda_u^w)^d}{\log \pi(u)} \to \frac{1}{H} \quad \text{a.s. and in } L^1, \tag{10}
\]
\[
\frac{\sum_{u \in C(n)} (\Lambda_u^w)^d}{n^d \log n^d} \to \frac{1}{H} \quad \text{a.s. and in } L^1. \tag{11}
\]

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 \( S \). Assertions (9) and (10) follow by setting \( g_{n,u} = L_u^d/\log \pi(u) \) and invoking Theorem 5 in the appendix. Finally, (11) follows from (10) by a multivariate generalization of the technique in Theorem 1. \qed
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 Algoit [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 [10]. We prove the one-sided version and then generalize to random fields. See also Theorem 7.5 on p. 66 of Krengel [8].

Theorem 4. (Maker) Let T be a measure preserving transformation of a probability space \((\mathcal{X}, \mathcal{B}, \mathcal{P})\) and let \(\mathcal{I}\) denote the \(\sigma\)-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 \(g = \lim \inf_{n,i \to \infty} g_{n,i}\), then
\[
\liminf_{n \to \infty} \frac{1}{n} \sum_{1 \leq i \leq n} g_{n,i}(T^i x) \geq E\{g|\mathcal{I}\} \quad \text{a.s.}
\]

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

Dividing both sides by \(n\) and taking the \(\liminf\) as \(n \to \infty\), we see that
\[
\liminf_{n \to \infty} \frac{1}{n} \sum_{1 \leq i \leq n} g_{n,i}(T^i x) \geq E\{g_k|\mathcal{I}\} \quad \text{a.s.,}
\]

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 \(k \geq 1\) 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,
\[
\frac{1}{n} \sum_{1 \leq i \leq n} g(T^i x) \to E\{g|\mathcal{I}\} \quad \text{in } L^1.
\]

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
\[
E\left|\frac{1}{n} \sum_{1 \leq i \leq n} g_{n,i}(T^i x) - g(T^i x)\right| \leq \frac{1}{n} \sum_{1 \leq i \leq n} E|g_{n,i}(T^i x) - g(T^i x)|
\]
\[
= \frac{1}{n} \sum_{1 \leq i \leq n} E|g_{n,i} - g| \to 0 \quad \text{as } n \to \infty.
\]
The $L^1$ convergence follows from (13) and (14) 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 [8]. 

Let $\{T_u : u \in \mathbb{Z}_+^d\}$ 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 $\{X_u : u \in \mathbb{Z}_+^d\}$ 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 $\sigma$-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 $D^\epsilon(n) = C(n) \setminus C^\epsilon(n)$. Note that $C^\epsilon(n)$ is contained in the sector

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

For any integer $n \geq 0$ and any nonnegative integer vector $u \in \mathbb{Z}_+^d$ 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

$$
g_k^\epsilon = \inf_{u \in S^\epsilon} g_{n,u}.
$$

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 $g^\epsilon$ decreases to a limit $g = \lim_{\epsilon \downarrow 0} g^\epsilon$.

**Theorem 5.** (a) Suppose the family $\{g_{n,u} : n \in \mathbb{Z}_+, u \in \mathbb{Z}_+^d\}$ 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 $S^\epsilon$. 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 lim inf as $n \to \infty$, we obtain

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

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


