ASYMPTOTIC RECURRENCE AND WAITING TIMES FOR STATIONARY PROCESSES

Ioannis Kontoyiannis[1]

Appeared in *Journal of Theoretical Probability*, **11**, pp. 795 - 811, July 1998.

Let $\boldsymbol{X} = \{X_n\ ;\ n \in \mathbb{Z}\}$ be a discrete-valued stationary ergodic process distributed according to P and let $x = (\ldots, x_{-1}, x_0, x_1, \ldots)$ denote a realization from $\boldsymbol{X}$. We investigate the asymptotic behavior of the *recurrence time* R_n defined as the first time that the initial n-block $x_1^n = (x_1, x_2, \ldots, x_n)$ re-appears in the past of x. We identify an associated random walk, $-\log P(X_1^n)$, on the same probability space as $\boldsymbol{X}$, and we prove a strong approximation theorem between $\log R_n$ and $-\log P(X_1^n)$. From this we deduce an almost sure invariance principle for $\log R_n$. As a byproduct of our analysis we get unified proofs for several recent results that were previously established using methods from ergodic theory, the theory of Poisson approximation and the analysis of random trees.

Similar results are proved for the *waiting time* W_n defined as the first time until the initial n-block from one realization first appears in an independent realization generated by the same (or by a different) process.

KEY WORDS: Recurrence; strong approximation; almost-sure invariance principles.

1. INTRODUCTION

Recurrence properties are important in the study of stationary processes in probability theory, and dynamical systems in ergodic theory. In this paper we investigate the asymptotic behavior of recurrence and waiting times for finite-valued stationary processes, under various mixing conditions.

Let $\boldsymbol{X} = \{X_n\ ;\ n \in \mathbb{Z}\}$ be a stationary ergodic process on the space of infinite sequences $(S^\infty, \mathcal{B}, P)$, where S is a finite set, $\mathcal{B}$ is the σ-field generated by finite-dimensional cylinders, and P is a shift-invariant ergodic probability measure. By $x \in S^{\mathbb{Z}}$, $x = (\ldots, x_{-1}, x_0, x_1, \ldots)$ we denote an infinite realization of $\boldsymbol{X}$, and for $i \leq j$ we write x_i^j for the finite substring $(x_i, x_{i+1}, \ldots, x_j)$ and $P(x_i^j)$ for the probability of the cylinder $\{y\ :\ y_i^j = x_i^j\}$. Similarly we write X_i^j for the vector $(X_i, \ldots, X_j)$, $x_{-\infty}^j$ for the semi-infinite string $(\ldots, x_{j-1}, x_j)$, and so on. Given two *independent* realizations x, y, our main quantities of interest are the *recurrence time* R_n defined as the first time until the opening string x_1^n recurs in the past of x, and

[1]Durand 141A, Information Systems Lab, Electrical Engineering Department, Stanford University, Stanford CA 94305-4055, USA. Tel: (650) 723-4544. Email: yiannis@isl.stanford.edu

[2]This work was partially supported by grants NSF #NCR-9205663, JSEP #DAAH04-94-G-0058, ARPA #J-FBI-94-218-2.

the *waiting time* W_n until the opening string x_1^n from the realization x first appears in the independent realization y:

$$\begin{aligned} R_n &= R_n(x) = \inf\{k \geq 1 : x_{-k+1}^{-k+n} = x_1^n\} \\ W_n &= W_n(x_1^n, y) = \inf\{k \geq 1 : y_k^{k+n-1} = x_1^n\}. \end{aligned}$$

There has been a lot of work on calculating the exact asymptotic behavior of R_n and W_n. Wyner and Ziv, motivated by coding problems in information theory, drew a deep connection between these quantities and the entropy rate of the underlying process.[22] They proved that R_n and W_n both grow exponentially with n and that the limiting rate is equal to the entropy rate $H = H(P) = \lim_n E[-\log P(X_0 \mid X_{-n}^{-1})]$. (Here and throughout the paper 'log' will denote the logarithm to base 2, and 'ln' the logarithm to base e.) Specifically, they showed that for stationary ergodic processes $(1/n)\log R_n$ converges to H in probability, that for stationary ergodic Markov chains $(1/n)\log W_n \to H$ in probability, and they also suggested that these results hold in the almost sure sense. Indeed this was later established by Ornstein and Weiss[13] who showed that for stationary ergodic processes

$$\frac{1}{n}\log R_n \to H \quad P - \text{a.s.}, \tag{1.1}$$

and by Shields[17] who showed that for functions of stationary ergodic Markov chains

$$\frac{1}{n}\log W_n \to H \quad P \times P - \text{a.s.} \tag{1.2}$$

The waiting time results were further extended, first by Nobel and Wyner[12] who showed that the convergence in probability holds for processes that are α-mixing with a certain rate, and then by Marton and Shields[11], who extended (1.2) to weak Bernoulli processes. Shields also provided a counter-example to show that (1.2) does not hold in the general ergodic case.[17]

Wyner and Ziv used in their analysis a theorem of Kac,[10] which can be phrased as follows: If $\boldsymbol{X}$ is stationary ergodic, then for any opening string x_1^n we have $E(R_n \mid X_1^n = x_1^n) = 1/P(x_1^n)$. This provides a strong formal connection between R_n and H: Taking logarithms of both sides in Kac's theorem, dividing by n and applying the Shannon-McMillan-Breiman theorem[5] yields

$$\lim_n \frac{1}{n}\log E(R_n \mid X_1^n) = \lim_n \frac{1}{n}\log[1/P(X_1^n)] = H \quad \text{a.s.} \tag{1.3}$$

We can therefore rephrase the Wyner-Ziv-Ornstein-Weiss result (1.1) by saying that they strengthened (1.3) by removing the conditional expectation

$$\lim_n \frac{1}{n}\log R_n = \lim_n \frac{1}{n}\log[1/P(X_1^n)] = H \quad \text{a.s.} \tag{1.4}$$

The crucial observation here is that Eq. (1.4) can be thought of as a strong approximation result between $\log R_n$ and $-\log P(X_1^n)$:

$$\log[R_n P(X_1^n)] \;=\; o(n) \quad \text{a.s.} \tag{1.5}$$

Our first result is a sharper form of (1.5), and a corresponding result for W_n.

1.1 Main Results

For $-\infty \le i \le j \le \infty$ let $\mathcal{B}_i^j$ denote the σ-field generated by X_i^j and define, for $d \ge 1$:

$$\begin{aligned}
\gamma(d) \;&=\; \max_{s\in S} E\left|\log P(X_0 = s \,|\, X_{-\infty}^{-1}) - \log P(X_0 = s \,|\, X_{-d}^{-1})\right| \\
\psi(d) \;&=\; \sup\left\{\frac{|P(B\cap A) - P(B)P(A)|}{P(B)P(A)} \;:\; A \in \mathcal{B}_{-\infty}^0,\; B \in \mathcal{B}_d^\infty\right\} \\
\alpha(d) \;&=\; \sup\left\{|P(B\cap A) - P(B)P(A)| \;:\; A \in \mathcal{B}_{-\infty}^0,\; B \in \mathcal{B}_d^\infty\right\},
\end{aligned}$$

where $0/0$ is interpreted as 0. $\boldsymbol{X}$ is called *ψ-mixing* if $\psi(d) \to 0$ as $d \to \infty$, and *strongly mixing* if $\alpha(d) \to 0$ as $d \to \infty$.

Theorem 1. *Let $\boldsymbol{X}$ be a finite-valued stationary ergodic process, and $\{c(n)\}$ an arbitrary sequence of non-negative constants such that $\sum n2^{-c(n)} < \infty$. For the recurrence times R_n we have*

$$\begin{aligned}
&(i) \quad \log[R_n P(X_1^n)] \;\le\; c(n), \quad \textit{eventually a.s.} \\
&(ii) \quad \log[R_n P(X_1^n \,|\, X_{-\infty}^0)] \;\ge\; -c(n), \quad \textit{eventually a.s.}
\end{aligned}$$

For the waiting times W_n we have

$$(iii) \quad \log[W_n P(X_1^n)] \;\ge\; -c(n), \quad \textit{eventually a.s.}$$

and if, in addition, $\boldsymbol{X}$ is ψ-mixing, then

$$(iv) \quad \log[W_n P(X_1^n)] \;\le\; c(n), \quad \textit{eventually a.s.}$$

with respect to the product measure $\mathbb{P} = P\times P$.

From Theorem 1 we can easily deduce:

Corollary 1. Strong approximation: *Let $\boldsymbol{X}$ be a finite-valued stationary ergodic process.*

(a) If $\sum \gamma(d) < \infty$ then for any $\beta > 0$

$$\log[R_n P(X_1^n)] \;=\; o(n^\beta) \quad \textit{a.s.}$$

(*b*) *If* $\boldsymbol{X}$ *is* ψ*-mixing then for any* $\beta > 0$

$$\log[W_n P(X_1^n)] \;=\; o(n^\beta) \quad a.s.$$

with respect to the product measure $\mathbb{P} = P \times P$.

(*c*) *In the general ergodic case we have*

$$\log[R_n P(X_1^n)] \;=\; o(n) \quad a.s.$$

The coefficients $\gamma(d)$ were introduced by Ibragimov in Ref. 8. If $\boldsymbol{X}$ is a Markov chain of order $m \geq 1$ then $\gamma(d) = 0$ for all $d \geq m$, so the rate at which $\gamma(d)$ decays may be interpreted as a measure of how well $\boldsymbol{X}$ can be approximated by finite-order Markov chains. Theorem 1 and Corollary 1 are proved in section 2.

We can now use Corollary 1 to read off the exact asymptotic behavior of $\log R_n$ and $\log W_n$ from that of $-\log P(X_1^n)$. This quantity can be interpreted in information theoretic terms as the ideal Shannon codeword length for the string X_1^n, and its asymptotics are well-understood. If $\boldsymbol{X}$ is ergodic, the Shannon-McMillan-Breiman theorem says that $(-1/n)\log P(X_1^n)$ converges almost surely to H, and combining this with Corollary 1 we get (1.1), and also (1.2) in the case when $\boldsymbol{X}$ is ψ-mixing.

If $\boldsymbol{X}$ is a Markov chain or, more generally, if it satisfies certain conditions on the rate of decay of $\alpha(d)$ and $\gamma(d)$, then $-\log P(X_1^n)$ behaves like the partial sum sequence of a strongly mixing stationary process (Ref. 15, chapter 9), and so it satisfies satisfies a central limit theorem, a law of the iterated logarithm, their infinite dimensional (functional) counterparts, as well as an almost sure invariance principle. Combining this with Corollary 1 gives us almost sure invariance principles for $\log R_n$ and $\log W_n$: Let $\{R(t)\,;\; t \geq 0\}$ denote the continuous-time path obtained by letting $R(t) = 0$ for $t < 1$, and $R(t) = [\log R_{\lfloor t \rfloor} - \lfloor t \rfloor H]$ for $t \geq 1$, where $\lfloor t \rfloor$ denotes the largest integer not exceeding t.

Theorem 2. Almost sure invariance principle: *Let* $\boldsymbol{X}$ *be a finite-valued stationary process such that* $\alpha(d) = O(d^{-336})$ *and* $\gamma(d) = O(d^{-48})$. *The following series converges:*

$$\sigma^2 = E[-\log P(X_0 \,|\, X_{-\infty}^{-1}) - H]^2 + 2\sum_{k=1}^{\infty} E[(-\log P(X_0 \,|\, X_{-\infty}^{-1}) - H)(-\log P(X_k \,|\, X_{-\infty}^{k-1}) - H)].$$

If $\sigma^2 > 0$, *then without loss of generality in the sense of Strassen, there exists a Brownian motion* $\{B(t)\,;\; t \geq 0\}$ *such that for any* $\lambda < 1/294$,

$$R(t) - \sigma B(t) \;=\; O(t^{1/2-\lambda}) \quad a.s.$$

Corresponding results hold for the waiting times W_n *in place of* R_n, *under the additional assumption that* $\boldsymbol{X}$ *is* ψ*-mixing.*

The phrase "without loss of generality in the sense of Strassen" means, as usual, that without changing its distribution, $R(t)$ can be redefined on a richer probability space, that contains a Brownian motion such that Theorem 2 holds. Theorem 2 is an immediate consequence of combining our Corollary 1 with Theorem 9.1 of Philipp and Stout.[15]

1.2 Consequences

The numerous corollaries that can be derived from the almost sure invariance principles of Theorem 2 are well-known:[18] $\log R_n$ and $\log W_n$ satisfy a central limit theorem and a law of the iterated logarithm (and also their functional counterparts – see Corollary 2, below), as well as stronger results such as

$$\limsup_{n\to\infty} \frac{\sum_{k=1}^{n} |\log R_k - kH|}{\sigma\sqrt{2n^3\,\mathrm{LLg}(n)}} \;=\; 3^{-1/2} \quad \text{a.s.},$$

where $\mathrm{LLg}(\cdot)$ denotes the function $\mathrm{LLg}(\cdot) = \ln\ln(\cdot \wedge e)$.

Corollary 2. *Under the assumptions of Theorem 2, if $\sigma^2 > 0$:*

(i) CLT:

$$\frac{\log R_n - nH}{\sigma\sqrt{n}} \;\xrightarrow{\mathcal{D}}\; N(0,1)$$

Moreover, the sequence of processes

$$\left\{ \frac{R(nt)}{\sigma\sqrt{n}} \;;\; t \in [0,1] \right\}, \quad n \geq 1,$$

converges in distribution to standard Brownian motion.

(ii) LIL:

$$\limsup_{n\to\infty} \frac{\log R_n - nH}{\sigma\sqrt{2n\mathrm{LLg}(n)}} \;=\; 1 \quad a.s.$$

Moreover, with probability one, the sequence of sample paths

$$\left\{ \frac{R(nt)}{\sigma\sqrt{2n\,\mathrm{LLg}(n)}} \;;\; t \in [0,1] \right\}, \quad n \geq 1,$$

is relatively compact in the topology of uniform convergence, and the set of its limit points is the collection of all absolutely continuous functions $r : [0,1] \to \mathbb{R}$, such that $r(0) = 0$ and $\int_0^1 (dr/dt)^2 dt \leq 1$.

(iii) Corresponding results hold for the waiting times W_n in place of R_n, under the additional assumption that $\boldsymbol{X}$ is ψ-mixing.

In the special case when $\boldsymbol{X}$ is an irreducible, aperiodic Markov chain, the one-dimensional version of the central limit theorem for W_n was first proved by Wyner,[21] who remarked that his methods can be modified to handle the case of R_n as well.

1.3 Match Lengths

The story of the asymptotics of R_n and W_n can equivalently be told in terms of match lengths along a realization. Given a realization x we define L_m as the length n of the shortest prefix x_1^n that does not appear starting anywhere else in the previous m positions x_{-m+1}^0:

$$L_m = L_m(x) = \inf\{n \geq 1 \ : \ x_1^n \neq x_{-j+1}^{-j+m}, \ \text{ for all } \ j = 1, 2, \ldots, m\}.$$

Following Wyner and Ziv[22] we observe that $R_n > m$ if and only if $L_m \leq n$, and consequently all asymptotic results about R_n can be translated into results about L_m: The almost sure convergence of $(1/n)\log R_n$ to H is equivalent to

$$\frac{L_m}{\log m} \to \frac{1}{H} \quad \text{a.s.}, \tag{1.6}$$

and the central limit theorem and law of the iterated logarithm for $\log R_n$ (Corollary 2) translate to:

Corollary 3. *Under the assumptions of Corollary 2:*

(*i*) *CLT:*

$$\frac{L_m - \frac{\log m}{H}}{\sigma H^{-3/2}\sqrt{\log m}} \xrightarrow{\mathcal{D}} N(0,1)$$

(*ii*) *LIL:*

$$\limsup_{n\to\infty} \frac{L_m - \frac{\log m}{H}}{\sigma H^{-3/2}\sqrt{2\log m \,\mathrm{LLg}(\log m)}} = 1 \quad a.s.$$

In the case of the waiting time, given two independent realizations x, y from $\boldsymbol{X}$, the dual quantity is the length n of the shortest prefix x_1^n of x that does not appear in y starting anywhere in y_1^m:

$$M_m = M_m(x,y) = \inf\{n \geq 1 \ : \ x_1^n \neq y_j^{j+n-1}, \ \text{ for all } \ j = 1, 2, \ldots, m\}.$$

Here $W_n > m$ if and only if $M_m \leq n$ and results about W_n can be equivalently stated in terms of M_m. In particular, (1.6) and the results of Corollary 3 hold with M_m in place of L_m when $\boldsymbol{X}$ is ψ-mixing.

1.4 History

Some brief remarks about the history of these results are in order here. The first explicit connection between match lengths and entropy seems to have been made by Pittel,[16] whose results are phrased in terms of path lengths in random trees. Aldous and Shields[1] first pointed out the relationship between the random tree interpretations of these results and coding algorithms. Recurrence times in relation to coding theory first appeared in Willems[20] and Wyner and Ziv.[22] Wyner and Ziv discovered the results (1.1) and (1.2), which were formally established by Ornstein and Weiss[13] and Shields,[17] using methods

from ergodic theory. The waiting time results were further extended to mixing processes by Nobel and Wyner[12] and Marton and Shields.[11] In the Markov case, Wyner[21] used the Chen-Stein method for Poisson approximation and Markov coupling to prove the one-dimensional central limit theorem for W_n. Szpankowski[19] made explicit the equivalence between match lengths along random sequences and feasible paths in random trees.

The approach introduced in this paper provides a probabilistic framework for studying the asymptotic behavior of R_n and W_n. From Theorem 2 we can deduce strong results that were not previously known, as well as several known results that were previously established using involved arguments and methods from other areas. Moreover, and, perhaps, more importantly, Theorem 1 tells us why these results are true: Because, in a strong pointwise sense, the recurrence time is asymptotically equal to reciprocal of the probability of the recurring string. We should also mention that this approach can be extended to random fields on $\mathbb{Z}^d$, though, of course, new subtleties arise in this case regarding the conditional structure of the measures and their mixing rates.

Some of the ideas in the proof of Theorem 1 can be traced in the work of Wyner and Ziv,[22] Ornstein and Weiss,[13] and Shields.[17] We also mention that ideas related to the use of $-\log P(X_1^n)$ (or a similar random walk) as an approximating sequence were used by Ibragimov[8] in proving a refinement to the Shannon-McMillan-Breiman theorem, by Barron[3] in proving the Shannon source coding theorem in the almost sure sense, and by Algoet and Cover[2] in an elementary proof of the Shannon-McMillan-Breiman theorem.

Apart from their theoretical interest, the results in this paper may be relevant to several areas of applications such as coding theory,[20,22] DNA sequence analysis,[14] and string searching algorithms.[7,9]

In section 2 we prove our main result, Theorem 1, and we deduce Corollary 1 from it. In section 3 we briefly discuss the special case when $\boldsymbol{X}$ is a Markov chain, and we give an explicit characterization (Theorem 3) of the degenerate case when the asymptotic variance in Theorem 2 is zero. Section 4 contains an extension of our waiting time results to the case when the independent realizations x and y are produced by different processes, and section 5 contains the proof of Theorem 3.

2. STRONG APPROXIMATION

We first deduce Corollary 1 from Theorem 1 and then we give the proof of Theorem 1.

Proof of Corollary 1. For part (a) let $\beta > 0$ arbitrary; since $\sum n2^{-\epsilon n^\beta} < \infty$ for any $\epsilon > 0$, from (i)

and (ii) of Theorem 1 we get

$$\limsup_{n\to\infty} \frac{1}{n^\beta} \log[R_n P(X_1^n)] \;\le\; 0 \quad \text{a.s.} \tag{2.1}$$

$$\liminf_{n\to\infty} \frac{1}{n^\beta} \log[R_n P(X_1^n \,|\, X_{-\infty}^0)] \;\ge\; 0 \quad \text{a.s.} \tag{2.2}$$

Therefore to prove (a) it suffices to show

$$\log P(X_1^n) - \log P(X_1^n \,|\, X_{-\infty}^0) = O(1) \quad \text{a.s.} \tag{2.3}$$

Observe that $|\log P(X_1^n) - \log P(X_1^n \,|\, X_{-\infty}^0)| \le \sum_{i=1}^n |\log P(X_i \,|\, X_1^{i-1}) - \log P(X_i \,|\, X_{-\infty}^{i-1})|$. Taking expectations of both sides we get $E|\log P(X_1^n) - \log P(X_1^n \,|\, X_{n+1}^\infty)| \le \sum_{i=1}^n \gamma(i)$, and since $\sum_{i=1}^\infty \gamma(i) < \infty$, this implies Eq. (2.3).

Part (b) follows immediately from (iii) and (iv) of Theorem 1, upon noticing that $\sum n 2^{-\epsilon n^\beta} < \infty$ for any $\epsilon, \beta > 0$. For part (c), taking $\beta = 1$ in Eqs. (2.1) and (2.2) we see that to prove (c) it suffices to show that

$$\frac{1}{n}\left[\log P(X_1^n) - \log P(X_1^n \,|\, X_{-\infty}^0)\right] \to 0 \quad \text{a.s.} \tag{2.4}$$

By the Shannon-McMillan-Breiman theore, the first term converges almost surely to $-H$, and the second term is equal to $(1/n)\sum_{i=1}^n [-\log P(X_i \,|\, X_{-\infty}^{i-1})]$ which converges to $H = E[-\log P(X_0 \,|\, X_1^\infty)]$, almost surely, by the ergodic theorem. This proves (2.4) and completes the proof of Corollary 1. □

Proof of Theorem 1. Part (i). Given an arbitrary positive constant K, by Markov's inequality and Kac's theorem,

$$P(R_n > K \,|\, X_1^n = x_1^n) \;\le\; \frac{E(R_n \,|\, X_1^n = x_1^n)}{K} \;=\; \frac{1}{K\, P(x_1^n)},$$

for any opening sequence x_1^n with non-zero probability. Since $P(x_1^n)$ is constant with respect to the conditional measure $P(\cdot \,|\, X_1^n = x_1^n)$ we can let $K = 2^{c(n)}/P(x_1^n)$ to get

$$P(\log[R_n P(X_1^n)] > c(n) \,|\, X_1^n = x_1^n) \;=\; P\left(R_n > 2^{c(n)}/P(x_1^n) \,|\, X_1^n = x_1^n\right) \;\le\; 2^{-c(n)}.$$

Averaging over all opening patterns $x_1^n \in S^n$, $P(\log[R_n P(X_1^n)] > c(n)) \le 2^{-c(n)}$, and the Borel-Cantelli lemma gives (i).

Part (ii). We now condition on the infinite past $X_{-\infty}^0$ instead of the opening string X_1^n. Fix any $x_{-\infty}^0$ and consider

$$P\left(\log[R_n(X) P(X_1^n \,|\, X_{-\infty}^0)] < -c(n) \,|\, X_{-\infty}^0 = x_{-\infty}^0\right) =$$
$$P\left(z_1^n \in S^n \,:\, P(X_1^n = z_1^n \,|\, X_{-\infty}^0) < \frac{2^{-c(n)}}{R_n(x_{-\infty}^0 * z_1^n)} \;\middle|\; X_{-\infty}^0 = x_{-\infty}^0\right),$$

where $*$ denotes concatenation of strings. If we let $G_n = G_n(x_{-\infty}^0)$ denote the set

$$\left\{ z_1^n \in S^n \ : \ P(z_1^n \,|\, x_{-\infty}^0) < 2^{-c(n)}/R_n(x_{-\infty}^0 * z_1^n) \right\},$$

then the above probability can be written as

$$\sum_{z_1^n \in G_n} P(z_1^n \,|\, x_{-\infty}^0) \ \leq \ \sum_{z_1^n \in G_n} 2^{-c(n)}/R_n(x_{-\infty}^0 * z_1^n) \ \leq \ 2^{-c(n)} \sum_{z_1^n \in S^n} 1/R_n(x_{-\infty}^0 * z_1^n). \tag{2.5}$$

Since $x_{-\infty}^0$ is fixed, for each $j \geq 1$ there is exactly one string z_1^n from S^n with $R_n(x_{-\infty}^0 * z_1^n) = j$, so the sum in Eq. (2.5) is bounded above by

$$\sum_{z_1^n \in S^n} 1/R_n(x_{-\infty}^0 * z_1^n) \ \leq \ \sum_{j=1}^{s^n} 1/j \ \leq \ Dn,$$

for some positive constant D, where $s = |S|$ is the cardinality of S. Therefore

$$P(\log[R_n P(X_1^n \,|\, X_{-\infty}^0)] < -c(n) \,|\, X_{-\infty}^0 = x_{-\infty}^0) \leq Dn2^{-c(n)},$$

and since this bound is independent of $x_{-\infty}^0$ and summable over n, from the Borel-Cantelli lemma we deduce (ii).

Part (iii). Consider the joint process $(\boldsymbol{X}, \boldsymbol{Y})$ distributed according to the product measure $\mathbb{P} = P \times P$. Given an arbitrary constant $K > 1$, for any opening string x_1^n with non-zero probability we have

$$\mathbb{P}(W_n < K \,|\, X_1^n = x_1^n) \ \leq \ \sum_{j=1}^{\lfloor K \rfloor} \mathbb{P}(W_n = j \,|\, X_1^n = x_1^n) \ \leq \ \sum_{j=1}^{\lfloor K \rfloor} P(Y_j^{j+n-1} = x_1^n) \ \leq \ K\, P(x_1^n).$$

Setting $K = 2^{-c(n)}/P(x_1^n)$ gives

$$\mathbb{P}(\log[W_n P(X_1^n)] < -c(n) \,|\, X_1^n = x_1^n) \leq 2^{-c(n)}.$$

(If $K = 2^{-c(n)}/P(x_1^n) \leq 1$ then $\mathbb{P}(W_n < K \,|\, X_1^n = x_1^n) = 0$ since $W_n \geq 1$ by definition so the above bound will trivially hold.) This is independent of x_1^n, so by the Borel-Cantelli lemma we get (iii).

Part (vi). Let $\delta \in (0,1)$ arbitrary and choose d such that $\psi(d) < \delta$. Fix an integer N large enough so that $2^{c(n)} \geq 2(n+d)$ for all $n \geq N$, fix an $n \geq N$, and let $K \geq 2(n+d)$ arbitrary. Then for any sequence x_1^n with non-zero probability, we can expand

$$\begin{aligned}
\mathbb{P}(W_n > K \,|\, X_1^n = x_1^n) \ &= \ P(Y_1^n \neq x_1^n, Y_2^{n+1} \neq x_1^n, \ldots Y_K^{K+n-1} \neq x_1^n) \\
&\leq \ [1 - P(x_1^n)] \prod_{j=1}^{\lfloor K/(n+d) \rfloor - 1} P\Big(Y_{j(n+d)+1}^{j(n+d)+n} \neq x_1^n \,\Big|\, Y_{i(n+d)+1}^{i(n+d)+n} \neq x_1^n, \ 0 \leq i < j \Big) \\
&= \ [1 - P(x_1^n)] \prod_{j=1}^{\lfloor K/(n+d) \rfloor - 1} [1 - P(B_j \,|\, A_j)],
\end{aligned} \tag{2.6}$$

where A_j and B_j are the events,

$$\begin{aligned} A_j &= \{Y_{i(n+d)+1}^{i(n+d)+n} \neq x_1^n,\ i = 0, 1, \ldots, j-1\} \ \in \ \mathcal{B}_1^{j(n+d)-d} \\ B_j &= \{Y_{j(n+d)+1}^{j(n+d)+n} = x_1^n\} \ \in \ \mathcal{B}_{j(n+d)+1}^{\infty}. \end{aligned}$$

By the choice of d and stationarity we have $P(B_j \,|\, A_j) \geq (1-\delta)P(B_j) = (1-\delta)P(x_1^n)$, for all j, and substituting in Eq. (2.6) we get

$$\begin{aligned} \mathbb{P}(W_n > K \,|\, X_1^n = x_1^n) &\leq [1-(1-\delta)P(x_1^n)]^{[K/(n+d)]} \\ &\leq \frac{1}{\delta}\,[1-(1-\delta)P(x_1^n)]^{\frac{K}{n+d}}. \end{aligned}$$

For any $n \geq N$ let $K = 2^{c(n)}/P(x_1^n) \geq 2(n+d)$ to obtain:

$$\begin{aligned} \mathbb{P}(\log[W_n P(X_1^n)] > c(n) \,|\, X_1^n = x_1^n) &\leq \frac{1}{\delta}\,[1-(1-\delta)P(x_1^n)]^{\frac{1}{P(x_1^n)}\,\frac{2^{c(n)}}{n+d}} \\ &\leq \frac{1}{\delta}\,\phi^{\frac{(1-\delta)2^{c(n)}}{n+d}}, \end{aligned}$$

where $\phi = \sup\{(1-z)^{1/z}\ ;\ 0 < z \leq 1-\delta\} < 1$. Since $\sum n2^{-c(n)} < \infty$, $(1/n)2^{c(n)} \to \infty$ as $n \to \infty$ and we can choose N' large enough such that $(\phi^{(1-\delta)})^{2^{c(n)}/(n+d)} \leq 2n2^{-c(n)}$ for all $n \geq N'$. Consequently,

$$\mathbb{P}(\log[W_n P(X_1^n)] > c(n) \,|\, X_1^n = x_1^n) \ \leq \ \frac{2}{\delta}\,n2^{-c(n)},$$

for all $n \geq M = \max\{N, N'\}$. Since this bound is independent of x_1^n,

$$\sum_{n\geq 1} \mathbb{P}(\log[W_n P(X_1^n)] > c(n)) \ \leq \ M + \frac{2}{\delta}\sum_{n\geq M} n2^{-c(n)} \ < \ \infty,$$

and the Borel-Cantelli lemma gives (iv) and completes the proof of the Theorem. □

Remark. In the proofs of (ii) and (iii) only the stationarity (and not the ergodicity) of $\boldsymbol{X}$ was used.

3. MARKOV CHAINS

If $\boldsymbol{X}$ is a stationary irreducible aperiodic Markov chain, then $\alpha(d)$ and $\psi(d)$ both decay exponentially fast, and $\gamma(d) = 0$ for all $d \geq 1$, so $\boldsymbol{X}$ satisfies all the conditions of Theorem 1 and Corollary 1. Consider the chain $\tilde{\boldsymbol{X}} = \{\tilde{X}_n = (X_n, X_{n+1})\ ;\ n \in \mathbb{Z}\}$ with state-space $T = \{(s,t) \in S\times S\ :\ P(X_{i+1} = t \,|\, X_i = s) > 0\}$; $\tilde{\boldsymbol{X}}$ is also stationary irreducible and aperiodic.

Let $f(s,t) = [-\log P(X_{i+1} = t \,|\, X_i = s)]$, $f : T \to \mathbb{R}$, and observe that here the entropy rate H of $\boldsymbol{X}$ is equal to $Ef(\tilde{X}_i)$, so that, with probability one, $[-\log P(X_1^n) - nH]$ behaves like the sequence of partial

sums of a centered bounded function of a Markov chain, up to a bounded term:

$$-\log P(X_1^n) - nH \;=\; \sum_{i=1}^{n-1}[f(\tilde{X}_i) - Ef(\tilde{X}_i)] \;+\; [-\log P(X_1) - H]. \tag{3.1}$$

In this case the asymptotic variance of Theorem 2 reduces to

$$\sigma^2 = \lim_{n\to\infty} \frac{1}{n}\mathrm{Var}(-\log P(X_1^n)). \tag{3.2}$$

The following characterization of the degenerate case $\sigma^2 = 0$ was stated by Yushkevich.[23] We supply a proof of a slightly stronger result in section 5.

Theorem 3. *Let $\boldsymbol{X}$ be a finite-valued stationary irreducible aperiodic Markov chain with entropy rate H, and let σ^2 be defined by Eq. (3.2). Then $\sigma^2 = 0$ if and only if every string X_1^{n+1} that starts and ends in some fixed state $j \in S$, has probability (given that $X_1 = j$) either zero or q^n, for some constant q depending on $\boldsymbol{X}$.*

4. WAITING TIMES BETWEEN DIFFERENT PROCESSES

Let $\boldsymbol{X}$, $\boldsymbol{Y}$ be two independent stationary processes distributed according to the measures P and Q, respectively, with values in the finite set S. We consider the waiting time $W_n(x_1^n, y)$ until the opening string x_1^n in the realization x of the $\boldsymbol{X}$-process first appears in an independent realization y produced by the $\boldsymbol{Y}$-process. We generalize our earlier results about W_n to this case, under the additional natural assumption that all finite-dimensional marginals P_n of P are dominated by the corresponding marginals Q_n of Q. If this is not satisfied then there will exist finite strings x_1^n such that $P(x_1^n) > 0$ but $Q(x_1^n) = 0$, and W_n will be infinite with positive probability.

The analog of Theorem 1 (parts (*iii*) and (*iv*)), reads:

Theorem 4. *Let $\boldsymbol{X}$, $\boldsymbol{Y}$ be independent stationary processes, distributed according to P, Q, respectively. Assume that $P_n \ll Q_n$ for all n, and write $\mathbb{P}$ for the product measure $P\times Q$. For any sequence $\{c(n)\}$ of non-negative constants such that $\sum n2^{-c(n)} < \infty$,*

$$\log[W_n Q(X_1^n)] \;\geq\; -c(n), \quad \textit{eventually } \mathbb{P}-a.s.,$$

and if, in addition, $\boldsymbol{Y}$ is ψ-mixing, then

$$\log[W_n Q(X_1^n)] \;\leq\; c(n), \quad \textit{eventually } \mathbb{P}-a.s.$$

The proof of Theorem 4 is a simple modification of the proof of the corresponding waiting time results in Theorem 1. Simply replace P by Q throughout the arguments that lead to (*iii*) and (*iv*), and note

that under the additional assumption that $P_n \ll Q_n$ for all n it suffices to condition on opening strings x_1^n of non-zero P_n-probability.

Now assume $\boldsymbol{X}$ is ergodic and $\boldsymbol{Y}$ is a Markov chain and define the relative entropy rate between P and Q as

$$D(P\|Q) = \lim_{n\to\infty} E_P\left[\log \frac{P(X_0 \,|\, X_{-n}^{-1})}{Q(X_0 \,|\, X_{-n}^{-1})}\right].$$

If for any $\epsilon > 0$ we let $c(n) = \epsilon n$ in Theorem 4, apply the generalized Shannon-McMillan-Breiman theorem[4] and let ϵ decrease to zero, we get:

Corollary 4. *If $\boldsymbol{X}$ is ergodic and $\boldsymbol{Y}$ is a Markov chain*

$$\lim_{n\to\infty} \frac{1}{n} \log W_n \;=\; H(P) + D(P\|Q) \quad \mathbb{P}-a.s.$$

Next, suppose that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are both stationary Markov chains and that $\boldsymbol{X}$ is irreducible and aperiodic. Define a continuous-time process $\{q(t)\ ;\ t \geq 0\}$ by letting $q(t) = 0$, for $t < 1$, and $q(t) = [-\log Q(X_1^{\lfloor t\rfloor}) - \lfloor t\rfloor(H(P) + D(P\|Q))]$, for $t \geq 1$. Consider the chain $\tilde{\boldsymbol{X}}$ of section 3, let $g : T \to \mathbb{R}$ be defined by $g(s,t) = [-\log Q(X_{i+1} = t \,|\, X_i = s)]$, and notice that $Eg(\tilde{X}_i) = H(P) + D(P\|Q)$. As in Eq. (3.1), we can think of $[-\log Q(X_1^n) - n(H(P) + D(P\|Q))]$ as the sequence of partial sums of a centered bounded function of a Markov chain (up to a bounded term):

$$\sum_{i=1}^{n-1}[g(\tilde{X}_i) - Eg(\tilde{X}_i)] \;+\; [-\log Q(X_1) - (H(P) + D(P\|Q))].$$

From well-known Markov chain results (Ref. 15, Theorem 10.1) it follows that $\{q(t)\}$ satisfies an almost sure invariance principle with asymptotic variance $\sigma^2 = \lim_n \mathrm{Var}(-\log Q(X_1^n))$. We can therefore combine this with Theorem 4 to obtain an analog of Theorem 2 in the case $P \neq Q$: Let $W(t) = 0$ for $t < 1$ and $W(t) = [\log W_{\lfloor t\rfloor} - \lfloor t\rfloor(H(P) + D(P\|Q))]$, for $t \geq 1$.

Corollary 5. *Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be stationary Markov chains and suppose that $\boldsymbol{X}$ is irreducible and aperiodic.*

If $\sigma^2 = \lim_n \mathrm{Var}(-\log Q(X_1^n)) > 0$, then, without loss of generality in the sense of Strassen, there exists a Brownian motion $\{B(t)\ ;\ t \geq 0\}$ such that for any $\beta > 1/4$

$$W(t) - \sigma B(t) \;=\; O(t^\beta), \quad a.s.$$

In the special case where both $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent and identically distributed, Wyner[21] proved Corollary 4 and the one-dimensional version of the central limit theorem in Corollary 5.

5. PROOF OF THEOREM 3

In this section we prove the following strengthened version of Theorem 3: $\sigma^2 = 0$ if and only if all the nonzero transition probabilities from state i to state j are of the form $2^{-H}v_i/v_j$, for some positive constants v_i, $i \in S$. Theorem 3 now follows with $q = 2^{-H}$.

We begin by deriving a generalization of a formula due to Fréchet,[6] for the asymptotic variance of Markov chains. Let $\boldsymbol{Z} = \{Z_n \ ; \ n \in \mathbb{Z}\}$ be a stationary irreducible aperiodic Markov chain with finite state-space T, stationary distribution $(q_i)_{i\in T}$, and kth order transition probabilities $(q_{ij}^{(k)})_{i,j\in T}$. Let f be a real-valued function on T and write $\bar{f}(\cdot)$ for $f(\cdot) - Ef(X_1)$. Define

$$\begin{aligned}\Sigma^2 = \lim_{n\to\infty} \frac{1}{n} \operatorname{Var}\left(\sum_{i=1}^{n} \bar{f}(Z_i)\right) &= E(\bar{f}(Z_1))^2 + 2\sum_{k=1}^{\infty} E(\bar{f}(Z_1)\bar{f}(Z_{k+1})) \\ &= \sum_{j\in T} \bar{f}(j)^2 q_j + 2\sum_{k=1}^{\infty}\sum_{i,j\in T} q_i q_{ij}^{(k)} \bar{f}(i)\bar{f}(j). \qquad (5.1)\end{aligned}$$

Letting $s_{ij} = \sum_{k=1}^{\infty}\left[q_{ij}^{(k)} - q_j\right] < \infty$ (for $i, j \in T$) the second term above becomes

$$2\sum_{i,j} q_i s_{ij} \bar{f}(i)\bar{f}(j) = 2\sum_i q_i \bar{f}(i)\theta_i,$$

where $\theta_i = \sum_j s_{ij}\bar{f}(j)$ (for $j \in T$), and substituting this in Eq. (5.1) gives

$$\Sigma^2 \ = \ \sum_i q_i \left[\bar{f}(i) + \theta_i\right]^2 - \sum_i q_i\theta_i^2 \ = \ \sum_j q_j \left[\sum_i q_{ji}(\bar{f}(i) + \theta_i)^2 - \theta_j^2\right]. \qquad (5.2)$$

Expanding

$$\begin{aligned}\sum_i q_{ji}\theta_i &= \sum_i q_{ji}\sum_m s_{im}\bar{f}(m) \\ &= \sum_m \bar{f}(m)\sum_i q_{ji}\sum_{k\geq 1}(q_{im}^{(k)} - q_m) \\ &= \sum_m \bar{f}(m)\sum_{k\geq 1}(q_{jm}^{(k+1)} - q_m) \\ &= \sum_m \bar{f}(m)\left[\sum_{k\geq 1}(q_{jm}^{(k)} - q_m) - (q_{jm} - q_m)\right] \\ &= \sum_m s_{jm}\bar{f}(m) - \sum_m q_{jm}\bar{f}(m) \\ &= \theta_j - \sum_m q_{jm}\bar{f}(m), \qquad (5.3)\end{aligned}$$

and so

$$\sum_i q_{ji}(\bar{f}(i) + \theta_i)^2 \ = \ \sum_i q_{ji}[(\bar{f}(i) + \theta_i - \theta_j) + \theta_j]^2 \ = \ \sum_i q_{ji}[(\bar{f}(i) + \theta_i - \theta_j)^2 + \theta_j^2], \qquad (5.4)$$

since by Eq. (5.3) the cross terms vanish

$$\begin{aligned}\sum_i q_{ji} 2\theta_j(\bar{f}(i) + \theta_i - \theta_j) &= 2\theta_j \left(\sum_i q_{ji}\bar{f}(i) - \theta_j + \sum_i q_{ji}\theta_i \right) \\ &= 2\theta_j \left(\sum_i q_{ji}\bar{f}(i) - \theta_j + \theta_j - \sum_m q_{jm}\bar{f}(m) \right) \;=\; 0.\end{aligned}$$

Substituting Eq. (5.4) into (5.2) and interchanging i and j yields

$$\Sigma^2 = \sum_j q_j \sum_i q_{ji}(\bar{f}(i) + \theta_i - \theta_j)^2, \tag{5.5}$$

which is the generalization of Fréchet's formula for the variance.

Now consider the chain $\tilde{\boldsymbol{X}}$ defined in section 3. For $i, j \in S$ we write $p_i = P(X_1 = i)$ and $p_{ij} = P(X_2 = j \,|\, X_1 = i)$, so that $\tilde{\boldsymbol{X}}$ has stationary distribution $(q_{ij}) = (p_i p_{ij})$ and transition probabilities $(q_{ij,kl}) = (\delta_{jk} p_{kl})$. Let f be defined as in section 3. Since here $\theta_{ij} = \theta_j$ is independent of i, using Eq. (5.5) we get

$$\begin{aligned}\sigma^2 &= \sum_{(i,j)\in T} p_i p_{ij} \sum_{(k,l)\in T} \delta_{jk} p_{kl} (\bar{f}(k,l) + \theta_{kl} - \theta_{ij})^2 \\ &= \sum_{(i,j)\in T} p_i p_{ij} \sum_{l \in S : p_{jl} > 0} p_{jl} (\bar{f}(j,l) + \theta_l - \theta_j)^2.\end{aligned}$$

For any $(j,l) \in T$ we have $p_{jl} > 0$ and the result stated in the beginning of this section follows, with $v_i = 2^{-\theta_i}$, $i \in S$.

The converse is obvious. □

ACKNOWLEDGEMENTS

The author wishes to thank Tom Cover and Tze Leung Lai for their helpful comments and suggestions, Amir Dembo for his encouragement and Benjamin Weiss for pointing out an error in an earlier version of our statement of Yushkevich's Theorem.

REFERENCES

1. Aldous, D. and Shields, P. C. (1988). A diffusion limit for a class of randomly-growing binary trees. *Prob. Th. Rel. Fields* **79**, 509–542.

2. Algoet, P. H. and Cover, T. M. (1988). A sandwich proof of the Shannon-McMillan-Breiman theorem. *Ann. Probab.* **16**, 876–898.

3. Barron, A. R. (1985a). *Logically smooth density estimation.* Ph.D. Thesis, Dept. of Electrical Engineering, Stanford University.

4. Barron, A. R. (1985b). The strong ergodic theorem for densities: Generalized Shannon-McMillan-Breiman theorem. *Ann. Probab.* **13**, 1992–1303.

5. Breiman, L. (1957). The individual ergodic theorem of information theory. *Ann. Math. Statist.* **28**, 809–811. See also, Breiman, L. (1960). A correction to "The individual ergodic theorem of information theory". *Ann. Math. Statist.* **31**, 809–810.

6. Fréchet, M. (1938). *Reserches théoriques modernes sur le calcul des probabilités*, vol. II, *Théorie des événements en chaîne dans le cas d'un nombre fini d'états possibles.* Gauthier-Villars, Paris (French).

7. Guibas, L. and Odlyzko, A. M. (1981). Periods in strings. *J. Combin. Theory Ser. A* **31**, 19–42.

8. Ibragimov, I. A. (1962). Some limit theorems for stationary processes. *Theory Prob. Appl.* **7**, 349–382.

9. Jacquet, P. and Szpankowski, W. (1994). Autocorrelation of words and its applications. *J. Combin. Theory Ser. A* **66**, 237–269.

10. Kac, M. (1947). On the notion of recurrence in discrete stochastic processes. *Bull. Amer. Math. Soc.* **53**, 1002–1010.

11. Marton, K. and Shields, P. C. (1995). Almost-sure waiting time results for weak and very weak Bernoulli processes. *Ergod. Th. & Dynam. Sys.* **15**, 951–960.

12. Nobel, A. and Wyner, A. D. (1992). A recurrence theorem for dependent processes with applications to data compression. *IEEE Trans. Inform. Theory* **38**, 1561–1564.

13. Ornstein, D. S. and Weiss, B. (1993). Entropy and data compression schemes. *IEEE Trans. Inform. Theory* **39**, 78–83.

14. Pevzner, P., Borodovsky, M. and Mironov, A. (1991). Linguistic of nucleotide sequences: the significance of deviations from mean statistical characteristics and prediction of the frequency of occurrence of words. *J. Biomol. Struct. Dynam.* **6**, 1013–1026.

15. Philipp, W. and Stout, W. (1975). *Almost sure invariance principles for partial sums of weakly dependent random variables.* Memoirs of the AMS, vol. 2, issue 2, no. 161.

16. Pittel, B. (1985). Asymptotical growth of a class of random trees. *Ann. Probab.* **13**, 414–427.

17. Shields, P. C. (1993). Waiting times: positive and negative results on the Wyner-Ziv problem. *J. Theor. Probab.* **6**, 499–519.

18. Strassen, V. (1964). An almost sure invariance principle for the law of the iterated logarithm. *Z. Wahrsch. Verw. Gabiete* **3**, 23–32.

19. Szpankowski, W. (1993). Asymptotic properties of data compression and suffix trees. *IEEE Trans. Inform. Theory* **39**, 1647–1659.

20. Willems, F. M. J. (1989). Universal data compression and repetition times. *IEEE Trans. Inform. Theory* **35**, 54–58.

21. Wyner, A. J. (1993). *String matching theorems and applications to data compression and statistics.* Ph.D. Thesis, Dept. of Statistics, Stanford University.

22. Wyner, A. D. and Ziv, J. (1989). Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data compression. *IEEE Trans. Inform. Theory* **35**, 1250–1258.

23. Yushkevich, A. A. (1953). On limit theorems connected with the concept of the entropy of Markov chains. *Uspehi Mat. Nauk* **8**, 177–180 (Russian).