ASYMPTOTIC RECURRANCE AND WAITING TIMES
FOR STATIONARY PROCESSES

BY
I. KONTOYIANNIS
STANFORD UNIVERSITY

TECHNICAL REPORT NO. 92
JULY 1996

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

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
ASYMPTOTIC RECURRENCE AND WAITING TIMES FOR STATIONARY PROCESSES

BY IOANNIS KONTOYIANNIS
Stanford University

Let $X = \{X_n ; n \in \mathbb{Z}\}$ 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 $X$. We investigate the asymptotic behavior of the recurrence time $R_n$ defined as the first time that the initial $n$-block $x^n_0 = (x_1, x_2, \ldots, x_n)$ recurs in $(x_2, x_3, \ldots)$. We identify an associated random walk, $-\log P(x^n_0)$, on the same probability space as $X$ and we prove a strong approximation theorem between $\log R_n$ and $-\log P(x^n_0)$. This provides a natural probabilistic framework for deducing the exact asymptotics of $\log R_n$. As a byproduct of our analysis we get unified simple 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 of the same process.

---

1This work was partially supported by grants NSF #NCR-9205663, JSEP #DAAH04-94-G-0058, ARPA #J-FBI-94-218-2.

2AMS 1991 subject classification. Primary—60F15, 94A17; secondary—60F17, 60F20, 60F05.

3Key words and phrases. Recurrence times, waiting times, string matching, entropy rate, strong approximation, strong, weak convergence, functional central limit theorem, functional law of the iterated logarithm, mixing processes.
1. Introduction. The purpose of this paper is to introduce a natural probabilistic approach to the study of asymptotic recurrence and waiting times for finite-valued stationary processes, under various mixing conditions.

Let $X = \{X_n : n \in \mathbb{Z}\}$ be a stationary ergodic process on the space of infinite sequences $(S^\infty, \mathcal{B}, \mathbb{P})$, where $S$ is a finite set (the state-space of $X$), $\mathcal{B}$ is the $\sigma$-field generated by finite-dimensional cylinders and $\mathbb{P}$ is a shift-invariant ergodic probability measure. By $x \in S^\infty$, $x = (\ldots, x_{-1}, x_0, x_1, \ldots)$ we denote an infinite realization of $X$, and for $i \leq j$ we write $x^j_i$ for the finite substring $(x_i, x_{i+1}, \ldots, x_j)$ and $P(x^j_i)$ for the probability of the cylinder \{ $y : y^j_i = x^j_i$ \}. Similarly we write $X^j_i$ for the vector $(X_i, \ldots, X_j)$, $x^j_{-\infty}$ 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 $x^n_1$ recurs in $x$, and the waiting time, $W_n$, until the the opening string $x^n_1$ from the realization $x$ first appears in the independent realization $y$:

\[
R_n = R_n(x) = \inf \{ k \geq 1 : x^n_{k+1} = x^n_1 \} \\
W_n = W_n(x^n_1, y) = \inf \{ k \geq 1 : y^n_{k+1} = x^n_1 \}.
\]

There has been a lot of work on calculating the exact asymptotic behavior of $R_n$ and $W_n$. Wyner and Ziv (1989), motivated by coding problems in information theory, drew a deep connection between these quantities and the entropy rate of the underlying process. They proved that $R_n$ and $W_n$ both grow exponentially with $n$ and that the limiting rate is equal to the entropy rate $H = \lim_n E[- \log P(X_0 | X_{-n})]$. (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, and for stationary ergodic Markov chains $(1/n) \log W_n \to H$ in probability, with respect to the product measure $\mathbb{P} = P \times P$. Wyner and Ziv also suggested that these results hold in the almost sure sense. Indeed this was later established by Ornstein and Weiss (1993) who showed that for stationary ergodic processes

\[
\frac{1}{n} \log R_n \to H, \quad \mathbb{P} - \text{a.s.}, \quad (1)
\]

and by Shields (1993) who showed that for functions of stationary ergodic Markov chains

\[
\frac{1}{n} \log W_n \to H, \quad \mathbb{P} - \text{a.s.} \quad (2)
\]

Shields (1993) also provided a counter-example to show that (2) is not true in the general ergodic case. Finally, a more precise result was found by Wyner (1993) who proved a central limit theorem (CLT) refinement to (2). He showed that when $X$ is a stationary ergodic Markov chain, the quantity $(\log W_n - nH)/\sqrt{n}$ has an asymptotically normal (possibly degenerate) distribution.

A strong formal connection between $R_n$ and $H$ is provided by a theorem of Kac (1947) which can be phrased as follows: If $X$ is stationary ergodic, then for any opening string $x^n_1$ we have $E(R_n | X^n_1 = x^n_1) = 1/P(x^n_1)$. Taking logarithms of both sides, dividing by $n$ and applying the Shannon-McMillan-Breiman theorem (Breiman (1957/1960)) yields

\[
\lim_{n \to \infty} \frac{1}{n} \log E(R_n | X^n_1) = \lim_{n \to \infty} \frac{1}{n} \log[1/P(X^n_1)] = H, \quad \text{a.s.} \quad (3)
\]
We can therefore rephrase the Wyner-Ziv-Ornstein-Weiss result (1) by saying that they strengthened (3) by removing the conditional expectation

$$\lim_{n} \frac{1}{n} \log R_n = \lim_{n} \frac{1}{n} \log[1/P(X_1^n)] = H, \quad \text{a.s.} \quad (4)$$

The crucial observation here is that (4) can be thought of as a strong approximation result between $\log R_n$ and $-\log P(X_1^n)$:

$$\log[R_n P(X_1^n)] = o(n), \quad \text{a.s.} \quad (5)$$

Our main result is a sharper form of (5), and a corresponding result for $W_n$.

For $-\infty \leq i \leq j \leq \infty$ let $B_i^j$ denote the $\sigma$-field generated by $X_i^j$ and define, for $d \geq 1$:

$$\psi(d) = \sup_{A \in B_{-\infty}^{d}, B \in B_{d}^{\infty}} \frac{|P(B \cap A) - P(B)P(A)|}{P(B)P(A)}$$

$$\rho(d) = \max_{s \in S} \mathbb{E} \left[ \log P(X_0 = s \mid X_{-\infty}^{-1}) - \log P(X_0 = s \mid X_{-d}^{-1}) \right],$$

where $0/0$ is interpreted as 0. $X$ is called $\psi$-mixing if $\psi(d) \to 0$ as $d \to \infty$.

**Theorem A.** Let $X$ be a finite-valued stationary ergodic process.

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

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

(b) If $X$ is $\psi$-mixing then for any $\beta > 0$

$$\log[W_n P(X_1^n)] = o(n^\beta), \quad \text{a.s.}$$

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

(c) In the general ergodic case we have

$$\log[R_n P(X_1^n)] = o(n), \quad \mathbb{P} - \text{a.s.}$$

The coefficients $\rho(d)$ were introduced by Ibragimov (1962). If $X$ is a Markov chain of order $m \geq 1$ then $\rho(d) = 0$ for all $d \geq m$, so the rate at which $\rho(d)$ decay may be interpreted as a measure of how well $X$ can be approximated by finite-order Markov processes.

Theorem A will be seen to follow from a stronger result (Theorem 1), which is stated and proved in section 2 below. We now use Theorem A 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 is interpreted in information theoretic terms as the ideal Shannon codeword length for the string $X_1^n$. If $X$ is ergodic, the Shannon-McMillan-Breiman theorem (SMBT) says that $(-1/n) \log P(X_1^n)$ converges almost surely to $H$, and combining this with Theorem A we immediately get (1) and (2). In fact, since functions of ergodic Markov chains are $\psi$-mixing, this extends Shields' result (2) to the somewhat larger class of $\psi$-mixing processes.
If $X$ is Markov then $-\log P(X^n_i)$ behaves like the random walk $\sum_{i=1}^{n-1}[-\log P(X_{i+1} \mid X_i)]$, so that when $X$ is stationary irreducible and aperiodic $-\log P(X^n_i)$ satisfies a CLT and a law of the iterated logarithm (LIL), as well as their infinite dimensional, functional counterparts (Theorem 2, section 3). Combining this with Theorem A gives us a complete description of the asymptotics of $\log R_n$ and $\log W_n$ (recall that $\log \equiv \log_2$ and $\ln \equiv \log_e$):

**THEOREM B.** If $X$ is a finite-valued stationary irreducible aperiodic Markov chain, then

$$\frac{\log R_n - nH}{\sqrt{n}} \xrightarrow{d} N(0, \sigma^2) \quad (6)$$

and, with probability one, the set of limit points of the sequence

$$\frac{\log R_n - nH}{\sqrt{2n \ln \ln n}} \quad (7)$$

coincides with the interval $[-\sigma, \sigma]$, where $\sigma^2 = \lim_n (1/n) \text{Var}(-\log P(X^n_i))$. Corresponding results hold for the waiting times $W_n$ in place of $R_n$.

Moreover, and essentially with no additional effort, we get a functional central limit theorem (FCLT) and a functional law of the iterated logarithm (FLIL) for $\log R_n$ and $\log W_n$ (Theorem 4, section 4). [Notation: The reason that we have logarithms to two different bases appearing in expression (7) above is that the natural units for measuring entropy are bits, whereas the deviation of partial sums from their mean is described in terms of natural logarithms. One could, for the sake of uniformity, either redefine $H$ in terms of $\ln$ so that all logarithms in (7) would be natural logarithms, or replace $\ln \ln$ by $\log \log$ and multiply (7) by an extra factor of $\sqrt{\log e}$.]

Some remarks about the history of these results are in order here. 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 $x^n_i$ that does not appear starting anywhere else in $x^n_i$:

$$L_m = L_m(x) = \inf\{n \geq 1: x^n_i \neq x^{j+m}_i, \text{ for all } j = 1, 2, \ldots, m-1\}.\quad (8)$$

Following Wyner and Ziv (1989) we observe that $R_n \geq m$ 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 $H$ (1) is equivalent to

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

the CLT for $R_n$ of Theorem B translates to

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

and so on. In the case of the waiting time, given two independent realizations $x, y$ from $X$, the dual quantity is the length $n$ of the shortest prefix $x^n_i$ of $x$ that does not appear in $y$ starting anywhere in $y^n_m$:

$$M_m = M_m(x, y) = \inf\{n \geq 1: x^n_i \neq y^{j+n-1}_j, \text{ for all } j = 1, 2, \ldots, m\}.$$
Here $W_n > m$ if and only if $M_m \leq n$ and results about $W_n$ can be equivalently stated in terms of $M_m$.

The first explicit connection between match lengths and entropy seems to have been made by Pittel (1985), whose results are phrased in terms of path lengths in random trees. For $l_n$ being the length of a path in a random tree, he investigated the convergence of $l_n / \log n$ to $1/h$ (cf. (8)), where $h$ is the entropy rate of an associated stationary process that satisfies certain mixing conditions. In the independent identically distributed (IID) case, Pittel (1986) also proved a CLT for $l_n$. Aldous and Shields (1988) first pointed out the relationship between the random tree interpretations of these results and coding algorithms. In the IID case of symmetric binary trees they proved a functional limit theorem for the number of available nodes in a random tree generated according to a coding algorithm. Their methods were very involved and, as they remarked, make it hard to see how these results would generalize to the Markov case or even the non-symmetric IID case.

Recurrence times in relation to coding theory first appeared in Willems (1989) and Wyner and Ziv (1989). Wyner and Ziv discovered the results (1) and (2), which were formally established by Ornstein and Weiss (1993) and Shields (1993), using methods from ergodic theory. In the Markov case, Wyner (1993) proved the CLT for $W_n$ using heavily technical arguments involving the Chen-Stein method for Poisson approximation and Markov coupling. Szpankowski (1993) made explicit the equivalence between match lengths along random sequences and feasible paths in random trees; also in the Markov case, using methods similar to those employed by Pittel (1985), he analyzed the asymptotics of shortest paths and used them to demonstrate the optimality of certain coding algorithms. The same point of view was adopted by Jacquet and Szpankowski (1995) to refine these results to CLTs in the IID case.

Finally we mention that ideas related to the use of $- \log P(X^n)$ (or a similar random walk) as an approximating sequence were used by Ibragimov (1962) in proving a CLT refinement to the SMBT, by Barron (1985a) in proving the Shannon source coding theorem in the almost sure sense, and by Algoet and Cover (1988) in an elementary proof of the SMBT.

The approach introduced in this paper provides a natural probabilistic framework for studying the asymptotics of $\log R_n$ and $\log W_n$. Theorem A provides a uniform basis from which 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), it tells us why these results are true: Because, in a strong pointwise sense, $\log R_n$ and $- \log P(X^n)$ are asymptotically equal.

Apart from their theoretical interest, these results have a wide range of applications in several areas including coding theory (Wyner and Ziv (1989), Wyner and Wyner (1995)), DNA sequence analysis (Pevzner, Borodovsky and Mironov (1991), Regnier and Szpankowski (1995)) and string searching algorithms (Guibas and Odlyzko (1981), Jacquet and Szpankowski (1994)).

In section 2 we state and prove our main result, Theorem 1. Section 3 discusses the asymptotic behavior of the Shannon codeword lengths $- \log P(X^n)$. In section 4 we combine the results from the previous two sections to obtain the asymptotic description of $R_n$ and $W_n$. Section 5 contains a discussion of extensions of our results: We consider waiting times between independent realizations produced by different processes, and we discuss the extent to which the Markovian assumption can be relaxed. Finally section 6 contains the proof of Theorem 3.
2. Strong approximation. In this section we state and prove our main result:

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

(i) \( \log[R_n P(X^n)] \leq c(n), \) eventually a.s.

(ii) \( \log[R_n P(X^n | X_{n+1}^{\infty})] \geq -c(n), \) eventually a.s.

If \( X \) is \( \psi \)-mixing then

(iii) \( |\log[W_n P(X^n)]| \leq c(n), \) eventually a.s.,

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

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

**Proof of Theorem A.**

Part (a). Let \( \beta > 0 \), arbitrary. Since \( \sum n2^{-\epsilon \beta} < \infty \) for any \( \epsilon > 0 \), from (i) and (ii) of Theorem 1 we get

\[
\limsup_{n \to \infty} \frac{1}{n^\beta} \log [R_n P(X^n)] \leq 0, \quad \text{a.s.} (9)
\]

\[
\liminf_{n \to \infty} \frac{1}{n^\beta} \log [R_n P(X^n | X_{n+1}^{\infty})] \geq 0, \quad \text{a.s.} (10)
\]

Therefore to prove (a) it suffices to show

\[
\log P(X^n) - \log P(X^n | X_{n+1}^{\infty}) = O(1), \quad \text{a.s.} (11)
\]

Observe that \( |\log P(X^n) - \log P(X^n | X_{n+1}^{\infty})| \leq \sum_{i=1}^{n} |\log P(X_i | X_{i+1}^{\infty}) - \log P(X_i | X_i^{\infty})| \). Taking expectations of both sides [and noting that by the stationarity of \( X \) the \( \rho \) coefficients of \( X \) are the same as those for the time reversed process \( \{X_{-n}\} \)] we get \( E|\log P(X^n) - \log P(X^n | X_{n+1}^{\infty})| \leq \sum_{i=1}^{n} \rho(i) \). Since, by assumption, \( \sum_{i=1}^{\infty} \rho(i) < \infty \), this implies (11).

Part (b). This follows immediately from (iii) of Theorem 1 upon noticing that \( \sum n2^{-\epsilon \beta} < \infty \) for any \( \epsilon, \beta > 0 \).

Part (c). Taking \( \beta = 1 \) in (9) and (10) above we see that to prove (c) it suffices to show that

\[
\frac{1}{n} [\log P(X^n) - \log P(X^n | X_{n+1}^{\infty})] \to 0, \quad \text{a.s.} (12)
\]

By the SMBT the first term above converges to \(-H\), with probability one. For the second term we expand \(- \log P(X_i | X_{i+1}^{\infty}) = \sum_{i=1}^{n} - \log P(X_i | X_{i+1}^{\infty})\), and we observe that the process \( \{- \log P(X_i | X_{i+1}^{\infty}) ; \; i \in \mathbb{Z}\} \) is stationary and ergodic since \( X \) is. Therefore, by the ergodic theorem, the second term converges to \( E(- \log P(X_0 | X_{\infty}^{-1}))\), with probability one, and by the stationarity of \( X \), \( E(- \log P(X_0 | X_{\infty}^{-1})) = E(- \log P(X_0 | X_{-\infty}^{-1})) = H \). This proves (12) and completes the proof of Theorem A. \( \square \)
Proof of Theorem 1.

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

$$P(R_n > K \mid X^n_1 = x^n_1) \leq \frac{E(R_n \mid X^n_1 = x^n_1)}{K} = \frac{1}{K P(x^n_1)},$$

for any opening sequence $x^n_1$ with non-zero probability. Since $P(x^n_1)$ is constant with respect to the conditional measure $P(\cdot \mid X^n_1 = x^n_1)$ we can let $K = 2^{-c(n)}/P(x^n_1)$ to get

$$P(\log[R_n P(X^n_1)] > c(n) \mid X^n_1 = x^n_1) = P[R_n > 2^{-c(n)}/P(x^n_1) \mid X^n_1 = x^n_1] \leq 2^{-c(n)},$$

and averaging over all opening patterns $x^n_1 \in S^n$, $P(\log[R_n P(X^n_1)] > c(n)) \leq 2^{-c(n)}$. Since $\sum 2^{-c(n)} \leq \sum n2^{-c(n)} < \infty$ the Borel-Cantelli lemma gives us (i).

Part (ii). We now condition on the infinite future $X^\infty_{n+1}$ instead of the opening string $X^n_1$. Fix any $X^\infty_{n+1}$ and consider

$$P(\log[R_n P(X^n_1) \mid X^\infty_{n+1}] < -c(n) \mid X^\infty_{n+1} = x^\infty_{n+1}) = \frac{2^{-c(n)}}{R_n(x^n_1 \ast x^\infty_{n+1})} \left\{ \begin{array}{l}
\frac{2^{-c(n)}}{R_n(x^n_1 \ast x^\infty_{n+1})} \left\{ z^n_1 \in S^n : P(X^n_1 = z^n_1 \mid X^\infty_{n+1} = x^\infty_{n+1}) < 2^{-c(n)} \right\},
\end{array} \right.$$  

where $\ast$ denotes concatenation of strings. If we let $G_n = G_n(x^\infty_{n+1})$ denote the set

$$\{ z^n_1 \in S^n : P(z^n_1 \mid x^\infty_{n+1}) < 2^{-c(n)} R_n(x^n_1 \ast x^\infty_{n+1}) \},$$

then the above probability can be written as

$$\sum_{z^n_1 \in G_n} P(z^n_1 \mid x^\infty_{n+1}) \leq \sum_{z^n_1 \in G_n} 2^{-c(n)}/R_n(x^n_1 \ast x^\infty_{n+1}) \leq 2^{-c(n)} \sum_{z^n_1 \in S^n} 1/R_n(x^n_1 \ast x^\infty_{n+1}). \quad (13)$$

Since $x^\infty_{n+1}$ is fixed, for each $j \geq 1$ there is exactly one string $z^n_1$ from $S^n$ with $R_n(z^n_1 \ast x^\infty_{n+1}) = j$. But the sum in (13) has at most $s^n$ non-zero terms (where $s = |S|$ is the cardinality of $S$), so it is bounded above by

$$\sum_{z^n_1 \in S^n} 1/R_n(z^n_1 \ast x^\infty_{n+1}) \leq \sum_{j=1}^{s^n} 1/j \leq Dn,$$

for some positive constant $D$ and therefore

$$P(\log[R_n P(X^n_1 \mid X^\infty_{n+1}] < -c(n) \mid X^\infty_{n+1} = x^\infty_{n+1}) \leq Dn2^{-c(n)}.$$

Since this bound is independent of $x^\infty_{n+1}$ the same inequality holds for the unconditional probability, and since the right-hand-side is summable over $n \geq 1$ from the Borel-Cantelli lemma we deduce (ii).

Part (iii). For the analysis of the waiting times $W_n(x^n_1, y)$ we consider the process $(X, Y)$ distributed according to the product measure $P = P \times P$. Given an arbitrary constant $K > 1$ we have

$$P(W_n < K \mid X^n_1 = x^n_1) = \sum_{j=1}^{[K-1]} P(W_n = j \mid X^n_1 = x^n_1) \leq \sum_{j=1}^{[K-1]} P(Y_j^{j+n-1} = x^n_1) \leq K P(x^n_1),$$

7
for any opening string $x^n_1$ with non-zero probability. Setting $K = 2^{-c(n)}/P(x^n_1)$ gives

$$P(\log[W_n P(X^n_1)] < -c(n) \mid X^n_1 = x^n_1) \leq 2^{-c(n)}.$$

(If $K = 2^{-c(n)}/P(x^n_1) \leq 1$, the probability $P(W_n < K \mid X^n_1 = x^n_1)$ will be zero since $W_n$ is greater than or equal to one by definition, and the above bound will still trivially hold.) This is independent of $x^n_1$ and summable over $n$ so by the Borel-Cantelli lemma

$$\log[W_n P(X^n_1)] \geq -c(n), \quad \text{eventually } P - \text{a.s.} \tag{14}$$

It remains to prove the corresponding asymptotic upper bound for $W_n$. Although we do not have Kac’s theorem for the waiting times, we can use the $\psi$-mixing structure of $P$ to obtain a different argument from the one used in the corresponding proof for $R_n$. 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 $x^n_1$ with non-zero probability, we can expand

$$P(W_n > K \mid X^n_1 = x^n_1) = P(Y^n_1 \neq x^n_1, Y^n_2 \neq x^n_1, \ldots, Y^K \neq x^n_1)$$

$$\leq [1 - P(x^n_1)] \prod_{j=1}^{K-d} P(Y^{j+n} \neq x^n_1 \mid Y^{n+1} \neq x^n_1, \ldots, j-1)$$

$$= [1 - P(x^n_1)] \prod_{j=1}^{K-d} [1 - P(B_j \mid A_j)], \tag{15}$$

where $A_j$ and $B_j$ are the events,

$$A_j = \{Y^{j+n} \neq x^n_1, i = 0, 1, \ldots, j-1\} \in \mathcal{B}_1^{(n+d)-d}$$

$$B_j = \{Y^{j+n} = x^n_1\} \in \mathcal{B}_j^{n+d+1}.$$

By the choice of $d$ and stationarity we have $P(B_j \mid A_j) \geq (1 - \delta)P(B_j) = (1 - \delta)P(x^n_1)$, for all $j$, and substituting in (15) we get

$$P(W_n > K \mid X^n_1 = x^n_1) \leq [1 - (1 - \delta)P(x^n_1)]^{\frac{K-d}{K}}$$

$$\leq \frac{1}{\delta} [1 - (1 - \delta)P(x^n_1)]^{\frac{K}{K+d}}. \tag{16}$$

Now for any $n \geq N$ we can let $K = 2^{c(n)}/P(x^n_1) \geq 2(n + d)$ to obtain:

$$P(\log[W_n P(X^n_1)] > c(n) \mid X^n_1 = x^n_1) \leq \frac{1}{\delta} [1 - (1 - \delta)P(x^n_1)]^{\frac{c(n)}{n+d}}$$

$$\leq \frac{1}{\delta} \gamma^{(1-\delta)c(n)/(n+d)}$$

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

$$P(\log[W_n P(X^n_1)] > c(n) \mid X^n_1 = x^n_1) \leq \frac{2}{\delta} n2^{-c(n)},$$
for all \( n \geq M = \max\{N,N'\} \). Since this bound is independent of \( x_i^n \),

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

and by the Borel-Cantelli lemma

\[
\log[W_n P(X_i^n)] \leq 1, \quad \text{eventually } P - \text{a.s.} \tag{17}
\]

Combining (14) and (17) gives (iii) and completes the proof of the Theorem. \( \Box \)

3. Asymptotics for \(- \log P(X_i^n)\). The quantity \(- \log P(X_i^n)\) is the ideal description length for the string \( X_i^n \) corresponding to the Shannon code. From the SMBT we know that the asymptotic mean of \(- \log P(X_i^n)\) is \( nH \). In this section we describe the asymptotics of \([ - \log P(X_i^n) - nH ]\) when \( X \) is a Markov chain. As reported in Yushkevich (1953), the question for such refinements of the SMBT was raised by Kolmogorov in the early 1950's. A first partial answer was offered by Yushkevich (1953) who proved a (one-dimensional) CLT for Markov chains; Ibragimov (1962) extended it to a class of stationary processes satisfying certain strong mixing conditions, and Philipp and Stout (1975) proved an almost sure invariance principle for \(- \log P(X_i^n)\) under conditions similar to the ones introduced by Ibragimov. These extensions are discussed in detail in section 5; throughout sections 3 and 4 we will assume that \( X \) is a finite-valued stationary irreducible aperiodic Markov chain.

We consider the chain \( \tilde{X} \) on the state-space \( T = \{ (s,t) \in S \times S : P(X_{i+1} = t | X_i = s) > 0 \} \) defined by \( \{ \tilde{X}_n = (X_n, X_{n+1}) ; n \in \mathbb{Z} \} \), so that \( \tilde{X} \) is also stationary irreducible and aperiodic. Then

\[
- \log P(X_i^n) = \sum_{i=1}^{n-1} [- \log P(X_{i+1} | X_i)] - \log P(X_1) \tag{18}
\]

\[
= \sum_{i=1}^{n-1} f(\tilde{X}_i) - \log P(X_1),
\]

where \( f : T \to \mathbb{R} ; (s,t) \mapsto - \log P(X_{i+1} = t | X_i = s) \). Observe that here the entropy rate \( H \) of \( X \) is equal to \( Ef(\tilde{X}_i) \), so that, with probability one, \([ - \log P(X_i^n) - nH ]\) behaves like the sequence of partial sums of a centered bounded function of a Markov chain, up to a bounded term:

\[
- \log P(X_i^n) - nH = \sum_{i=1}^{n-1} [f(\tilde{X}_i) - Ef(\tilde{X}_i)] + [- \log P(X_1) - H]. \tag{19}
\]

Let \( p_n(t) \) be the continuous path obtained by setting \( p_n(0) = 0 \) and linearly interpolating the function \( ( - \log P(X_i^n) - ntH ) \) at \( t = i/n, \ i = 1,2,\ldots,n, \ 0 \leq t \leq 1 \). This defines a family of continuous sample-path processes \( \{ p_n(t) ; 0 \leq t \leq 1 \}_{n \geq 1} \). Let \( \lg^2(n) \) denote the function \( \ln \ln(n), \ n \geq 3 \) and let \( K \) be the set of all real-valued absolutely continuous functions \( q \) on \([0,1]\) such that \( q(0) = 0 \) and \( \int_0^1 (dq/dt)^2 dt \leq 1 \).
Theorem 2. Let $X$ be a finite-valued stationary irreducible aperiodic Markov chain.

(i) Variance: The limit $\sigma^2 = \lim_n (1/n) \text{Var}( -\log P(X^n_t))$ exists and is finite.

(ii) FCLT: If $\sigma^2 > 0$, the sequence

$$\left\{ \frac{p_n(t)}{\sigma \sqrt{n}} ; 0 \leq t \leq 1 \right\}_{n \geq 1}
$$

converges in distribution to standard Brownian motion.

(iii) FLIL: If $\sigma^2 > 0$, with probability one the sequence

$$\left\{ \frac{p_n(t)}{\sigma \sqrt{2n \log^2 n}} ; 0 \leq t \leq 1 \right\}_{n \geq 3}
$$

is relatively compact and the set of its limit points coincides with $K$.

Proof. (i) The limit

$$\lim_n \frac{1}{n} \text{Var} \left( \sum_{i=1}^{n-1} -\log P(X_{i+1} \mid X_i) \right) = \lim_n \frac{1}{n} \text{Var} \left( -\log P(X^n_2 \mid X_1) \right)
$$

exists and is finite by the discussion preceding the theorem and the CLT for functions of Markov chains (see Chung (1967), for example). That it equals $\sigma^2$ follows from (19). Parts (ii) and (iii) follow from part (i), (19) and standard results from Markov chain theory (see Philipp and Stout (1975), chapter 10).

The following characterization of the degenerate case $\sigma^2 = 0$ was stated without proof in Yushkevich (1953). We supply a proof in section 6.

Theorem 3. Let $X$ be a finite-valued stationary irreducible aperiodic Markov chain with entropy rate $H$, and let $\sigma^2$ be defined as in Theorem 2. Then $\sigma^2 = 0$ if and only if all the non-zero transition probabilities of the chain are equal to $2^{-H}$.

4. Functional Limit Theorems for $R_n$ and $W_n$. In this section we combine the results of the previous two sections to get a FCLT and a FLIL for $\log R_n$ and $\log W_n$. Theorem 4 below is the infinite-dimensional generalization of Theorem B stated in the introduction.

Let $r_n(t)$ be the continuous path obtained by setting $r_n(0) = 0$ and linearly interpolating the function $(\log R_{nt-ntH})$ at $t = i/n$, $i = 0, 1, \ldots, n$, $0 \leq t \leq 1$, and similarly define $w_n(t)$ in terms of $(\log W_{nt-ntH})$. This defines two families of continuous sample-path processes, $\{r_n(t) ; 0 \leq t \leq 1\}$ and $\{w_n(t) ; 0 \leq t \leq 1\}$, $n \geq 1$, with distributions determined by the measures $P$ and $P = P \times P$, respectively.

Let the set $K$ and the variance $\sigma^2$ be defined as in the previous section.

10
Theorem 4. Let $X$ be a finite-valued stationary irreducible aperiodic Markov chain and suppose that $\sigma^2 > 0$. Then:

(i) FCLT: The sequences

$$\left\{ \frac{r_n(t)}{\sigma \sqrt{n}} ; 0 \leq t \leq 1 \right\}_{n \geq 1} \quad \text{and} \quad \left\{ \frac{w_n(t)}{\sigma \sqrt{n}} ; 0 \leq t \leq 1 \right\}_{n \geq 1}$$

converge in distribution to standard Brownian motion.

(ii) FLIL: With probability one the sequences

$$\left\{ \frac{r_n(t)}{\sigma \sqrt{2n \log^2 n}} ; 0 \leq t \leq 1 \right\}_{n \geq 3} \quad \text{and} \quad \left\{ \frac{w_n(t)}{\sigma \sqrt{2n \log^2 n}} ; 0 \leq t \leq 1 \right\}_{n \geq 3}$$

are relatively compact and the sets of their limit points both coincide with $K$.

Proof. Simply observe that

$$\sup_{0 \leq t \leq 1} \left| \frac{r_n(t)}{\sqrt{n}} - \frac{p_n(t)}{\sqrt{n}} \right| = \frac{1}{\sqrt{n}} \max_{1 \leq i \leq n} | \log[R_i P(X_i)] |.$$ 

Since Markov chains have $\rho(d) = 0$ for all $d \geq 1$, Theorem A, (a) implies that the right-hand-side above converges to zero almost surely as $n \to \infty$. Applying Theorem 2 completes the proof. Similarly for $w_n(t)$ (recall that stationary, irreducible, aperiodic Markov chains are $\psi$-mixing).

As mentioned in the introduction, these results can equivalently be phrased in terms of match lengths. Here we simply state the one-dimensional CLTs for $L_m$ and $M_m$.

Corollary 1. Under the assumptions of Theorem 4, the quantities

$$\frac{L_m - \log m}{\sigma H^{-3/2} \sqrt{\log m}} \quad \text{and} \quad \frac{M_m - \log m}{\sigma H^{-3/2} \sqrt{\log m}}$$

both have an asymptotic standard normal distribution.

5. Extensions. In this section we extend the results of the previous two sections by weakening the assumptions in Theorem 2 and Theorem 4. First we mention that all these results remain true for Markov chains of any finite order. If $X$ is $k$th order Markov, $k > 1$, then instead of looking at $\tilde{X}_n = (X_n, X_{n+1}, \ldots, X_{n+k})$ we can define $\tilde{X}_n = (X_n, X_{n+1}, \ldots, X_{n+k})$ and proceed exactly as before to obtain Theorem 2. This together with Theorem 1 imply will Theorem 4.

Below we consider waiting times between independent realizations produced by different processes, and we discuss the extent to which the Markovian assumption can be relaxed.

Waiting Times Between Different Processes.

Let $X$, $Y$ be two independent stationary processes distributed according to the measures $P$ and $Q$, respectively, with values in the finite state-space $S$. We consider the waiting time $W_n(x^n_1, y)$
until the opening string $x_1^n$ in the realization $x$ of the $X$-process first appears in an independent realization $y$ produced by the $Y$-process. All asymptotic results about $W_n$ generalize 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. Throughout this subsection we will assume that:

1. $X, Y$ are independent stationary processes distributed according to $P, Q$, respectively;
2. $P_n \ll Q_n$, for all $n$; and
3. $Y$ is $\psi$-mixing.

Let $\mathbb{P}$ denote the product measure $P \times Q$. The analog of Theorem 1 reads:

**Theorem 5.** For any sequence $\{c(n)\}$ of non-negative constants such that $\sum n2^{-c(n)} < \infty$:

$$|\log(W_n Q(x_1^n))| \leq c(n), \text{ eventually } \mathbb{P} - a.s.$$

The proof of Theorem 5 is identical to the proof of the corresponding waiting time results in Theorem 1. We just replace $P$ by $Q$ throughout the arguments that lead to equations (14) and (17) 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.

Define the relative entropy rate between $P$ and $Q$ as

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

where the limit exists by the assumption that $P_n \ll Q_n$. If we let $\epsilon > 0$, arbitrary, let $c(n) = \epsilon n$ in Theorem 5 and apply the SMBT we get:

**Corollary 2.** If $X$ is ergodic and $Y$ is a Markov chain

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

We can also generalize Theorems 2 and 4 to the case $P \neq Q$. Assume that $X$ and $Y$ are both Markov chains and define a sequence of continuous-path processes $\{q_n(t) ; 0 \leq t \leq 1\}$, $n \geq 1$ by setting $q_n(0) = 0$ and linearly interpolating the function $(-\log Q(X_1^n) - nt[H(P) + D(P||Q)])$ at $t = i/n, \ i = 0, 1, \ldots, n, \ 0 \leq t \leq 1$. Let $\sigma^2 = \lim_n \text{Var}(\log Q(X_1^n))$ and proceed exactly as in section 3 to get that the conclusions of Theorem 2 remain valid with $q_n(t)$ in place of $p_n(t)$. Combining this with Theorem 5 as in the proof of Theorem 4 leads to:

**Corollary 3.** Let $X$ and $Y$ be Markov chains, let $\tilde{w}_n(t)$ be the the continuous path obtained by letting $\tilde{w}_n(0) = 0$ and linearly interpolating the function $(\log W_{nt} - nt[H(P) + D(P||Q)])$ at $t = i/n, \ i = 0, 1, \ldots, n, \ 0 \leq t \leq 1$.

If $\sigma^2 = \lim_n (1/n)\text{Var}(-\log Q(X_1^n)) > 0$, then the sequence

$$\{\tilde{w}_n(t) ; 0 \leq t \leq 1\}_{n \geq 1}$$

satisfies a FCLT and a FLIL with asymptotic variance $\sigma^2$.  

12
In the special case where both $X$ and $Y$ are IID processes Wyner (1993) used the Chen-Stein method for Poisson approximation to prove Corollary 2 and the one-dimensional version of the CLT in Corollary 3.

**Non-Markov Processes.**

Since the results of Theorem 1 concerning $R_n$ are valid not only for Markov chains, but for all stationary ergodic processes such that $\sum_p(d) < \infty$, if we look at the proof of Theorem 4 it becomes clear that the extent of its validity is exactly the extent to which Theorem 2 is valid. Therefore we can replace the Markovian assumption in Theorem 4 by the requirements that $X$ is stationary ergodic, it has $\sum_p(d) < \infty$, and it satisfies the conclusions of Theorem 2. For clarity of exposition we assumed in sections 3 and 4 that $X$ is a Markov chain but this requirement can be replaced by certain strong mixing conditions. The nature of these conditions and the results they imply are discussed below.

With $B_i^j$ as in section 1, we define the **strong mixing coefficients** of $X$ by

$$\alpha(d) = \sup_{A \in B_{d=\infty}^\infty, B \in B_{d}^\infty} |P(B \cap A) - P(B)P(A)|.$$  

$X$ is called **strongly mixing** if $\alpha(d) \to 0$ as $d \to \infty$. Without the Markovian assumption we can no longer approximate $-\log P(X^n_1)$ by a stationary random walk. The equivalent form of (18) in this case would be

$$-\log P(X^n_1) = \sum_{i=1}^{n-1} [-\log P(X_{i+1} | X^i_1)] - \log P(X_1)$$

and the summands do not form a stationary process. Ibragimov (1962) showed that when $\alpha(d)$ and $\rho(d)$ decay fast enough to zero we can approximate $-\log P(X^n_1)$ by the partial sums of a related stationary process

$$\sum_{i=1}^{n-1} [-\log P(X_{i+1} | X^i_{-\infty})],$$

and then apply the CLT for strongly mixing stationary sequences. Ibragimov's result was strengthened to an almost sure invariance principle by Philipp and Stout (1975). They used they same approximating sequence (20) in conjunction with their results on strongly mixing sequences to prove an almost sure invariance principle for $-\log P(X^n_1)$. From Philipp and Stout (1975), Theorem 9.1, we easily deduce that if $X$ is a finite-valued stationary process such that $\alpha(d) = O(d^{-336})$ and $\rho(d) = O(d^{-48})$ then

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

converges, and if $\sigma^2 > 0$ then the sequence

$$\{p_n(t) ; 0 \leq t \leq 1\}_{n \geq 1}$$

satisfies a FCLT and a FLIL with asymptotic variance $\sigma^2$. 

13
We can now combine this result with Theorem 2 as in the proof of Theorem 4 to get the following:

**THEOREM 6.** Let \( X \) be a finite-valued stationary process such that \( \alpha(d) = O(d^{-33\varepsilon}) \) and \( \rho(d) = O(d^{-4\varepsilon}) \). If \( \sigma^2 > 0 \) then the sequence

\[
\{ r_n(t) ; 0 \leq t \leq 1 \}_{n \geq 1}
\]

satisfies a FCLT and a FLIL with asymptotic variance \( \sigma^2 \).

If, moreover, \( X \) is \( \psi \)-mixing then the same asymptotic results hold for the sequence

\[
\{ w_n(t) ; 0 \leq t \leq 1 \}_{n \geq 1}.
\]

6. **Proof of Theorem 3.** We begin by deriving a generalization of a formula for the variance due to Fréchet (1938). Let \( Z = \{ Z_n ; n \in \mathbb{Z} \} \) be a stationary irreducible aperiodic Markov chain with finite state-space \( T \), stationary distribution \( (q_i)_{i \in T} \), and \( k \)th 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

\[
\Sigma^2 = \lim_{n \to \infty} \frac{1}{n} \text{Var} \left( \sum_{i=1}^{n} \bar{f}(Z_i) \right) = E(\bar{f}(Z_1))^2 + 2 \sum_{k=1}^{\infty} E(\bar{f}(Z_1)\bar{f}(Z_{k+1})) = \sum_{j \in T} \bar{f}(j)^2 q_j + 2 \sum_{k=1}^{\infty} \sum_{i,j \in T} q_i q_{ij}^{(k)} \bar{f}(i) \bar{f}(j). \tag{21}
\]

Letting \( s_{ij} = \sum_{k=1}^{\infty} [q_{ij}^{(k)} - q_j] < \infty \) (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 (21) we get

\[
\Sigma^2 = \sum_{i} q_i \left[ \bar{f}(i) + \theta_i \right]^2 - \sum_{i} q_i \theta_i^2 = \sum_{j} q_j \left[ \sum_{i} q_{ji} (\bar{f}(i) + \theta_i)^2 - \theta_i^2 \right]. \tag{22}
\]

Expanding

\[
\sum_{i} q_{ji} \theta_i = \sum_{i} q_{ji} \sum_{m} s_{im} \bar{f}(m)
= \sum_{m} \bar{f}(m) \sum_{i} q_{ji} \sum_{k \geq 1} (q_{im}^{(k)} - q_m)
= \sum_{m} \bar{f}(m) \sum_{k \geq 1} (q_{jm}^{(k+1)} - q_m)
\]
\[
\begin{align*}
= \sum_m \bar{f}(m) \left[ \sum_{k \geq 1} (q_{jm}^{(k)} - q_m) - (q_{jm} - q_m) \right] \\
= \sum_m s_{jm} \bar{f}(m) - \sum_m q_{jm} \bar{f}(m) \\
= \theta_j - \sum_m q_{jm} \bar{f}(m),
\end{align*}
\]
and so
\[
\sum_i q_{ji}(\bar{f}(i) + \theta_i)^2 = \sum_i q_{ji}[(\bar{f}(i) + \theta_i - \theta_j)^2 + \theta_j^2] = \sum_i q_{ji}[(\bar{f}(i) + \theta_i - \theta_j)^2 + \theta_j^2], \tag{24}
\]
since by (23) the cross terms vanish
\[
\sum_i q_{ji}2\theta_j(\bar{f}(i) + \theta_i - \theta_j) = 2\theta_j \left( \sum_i q_{ji} \bar{f}(i) - \theta_j + \sum_i q_{ji} \theta_i \right) \\
= 2\theta_j \left( \sum_i q_{ji} \bar{f}(i) - \theta_j + \theta_j - \sum_m q_{jm} \bar{f}(m) \right) = 0.
\]
Substituting (24) into (22) and interchanging \(i\) and \(j\) yields
\[
\Sigma^2 = \sum_j q_j \sum_i q_{ji}(\bar{f}(i) + \theta_i - \theta_j)^2, \tag{25}
\]
which is the generalization of Fréchet's formula for the variance.

Now consider the chain \(\tilde{X}\) 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 \(\tilde{X}\) 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 \(i\), applying (25) we get
\[
\sigma^2 = \sum_{(i,j) \in T} p_i p_{ij} \sum_{(k,l) \in T} \delta_{jk} p_{kl} (f(k,l) + \theta_k - \theta_l)^2 \\
= \sum_{(i,j) \in T} p_i p_{ij} \sum_{l \in S : p_{jl} > 0} p_{jl} (f(j,l) + \theta_l - \theta_j)^2.
\]
For any \((j,l) \in T\) we have \(p_{jl} > 0\) so we may take \(j = l\) in the above formula to conclude that whenever \(p_{jl} \neq 0\), we have \(\bar{f}(j,l) = 0\), i.e., \(p_{jl} = 2^{-H}\).

The converse is obvious. \(\square\)

Acknowledgements. We would like to thank Professors Tom Cover and Tze Leung Lai for their helpful comments and suggestions.
REFERENCES


IOANNIS KONTOYIANNIS
INFORMATION SYSTEMS LABORATORY
DURAND BUILDING 141A
STANFORD UNIVERSITY
STANFORD CA94305–4055