COLLOQUIUM MATHEMATICUM

VOL. 84/85 2000 PART 1

AVERAGE CONVERGENCE RATE OF THE FIRST RETURN TIME

BY

GEON HO CHOE AND DONG HAN KIM (TAEJON)

Abstract. The convergence rate of the expectation of the logarithm of the first return time R_n, after being properly normalized, is investigated for ergodic Markov chains. I. Kontoyiannis showed that for any $\beta > 0$ we have $\log[R_n(x)P_n(x)] = o(n^\beta)$ a.s. for aperiodic cases and A. J. Wyner proved that for any $\varepsilon > 0$ we have $-(1+\varepsilon)\log n \le \log[R_n(x)P_n(x)] \le \log\log n$ eventually, a.s., where $P_n(x)$ is the probability of the initial n-block in x. In this paper we prove that $E[\log R_{(L,S)} - (L-1)h]$ converges to a constant depending only on the process where $R_{(L,S)}$ is the modified first return time with block length L and gap size S. In the last section a formula is proposed for measuring entropy sharply; it may detect periodicity of the process.

1. Introduction. Convergence of the logarithm of first return time normalized by the block length has recently been investigated in relation to data compression methods such as Ziv–Lempel algorithms [17], [18]. For each sample sequence $x = (\xi_1, \xi_2, \ldots)$ from an ergodic stationary information source, define $P_n(x)$ to be the probability of the initial n-block in x, i.e., $P_n(x) = \Pr(x_1 \ldots x_n)$. The classical Shannon–Breiman–McMillan Theorem states that $-(\log P_n)/n$ converges to measure-theoretic entropy h in L^1 and almost surely. Define

$$R_n(x) = \min\{j \ge 1 : \xi_1 \ldots \xi_n = \xi_{j+1} \ldots \xi_{j+n}\}.$$

Kac's Lemma [2] states that $E[R_n \mid a_1 \ldots a_n] = 1/\Pr(a_1 \ldots a_n)$.

LEMMA 1.1. *$E[R_n] = E[1/P_n] =$ the number of n-blocks with positive probability.*

Proof. Let $\mathcal{P}_n$ be the partition according to the first n-blocks. Note that

$$E[R_n] = \sum_{B \in \mathcal{P}_n,\, \Pr(B)>0} E[R_n \mid B] \Pr(B) = \sum_{B \in \mathcal{P}_n,\, \Pr(B)>0} 1,$$

2000 *Mathematics Subject Classification*: Primary 94A17; Secondary 60J10.

Key words and phrases: the first return time, Markov chain, Wyner–Ziv–Ornstein–Weiss theorem, entropy, data compression, period of an irreducible matrix.

Research supported by the Ministry of Information and by Communication and by GARC-SRC-KOSEF.

which equals the number of n-blocks B with positive probability. Similarly,

$$E\left[\frac{1}{P_n}\right] = \sum_{B\in\mathcal{P}_n,\,\Pr(B)>0} E\left[\frac{1}{P_n}\,\middle|\, B\right]\Pr(B) = \sum_{B\in\mathcal{P}_n,\,\Pr(B)>0} \frac{1}{\Pr(B)}\Pr(B),$$

and we obtain the same number. ■

Observe that $E[R_n]$ is an integer. This suggests that $R_n(x)$ is close to $1/P_n(x)$, hence we expect that $(\log R_n)/n$ converges to entropy h in a suitable sense. It may be viewed in the following way: According to the Asymptotic Equipartition Property the number of typical subsets is approximately equal to 2^{nh} in the nth stage. Because of ergodicity a generic orbit would visit almost all the typical subsets, hence we conjecture that the return time $R_n(x)$ for almost every starting point x would be approximately equal to 2^{nh}. This is indeed the case. Conventionally a slightly modified definition for the first return time is used. It was proved that $(\log R_n)/n$ converges to entropy in probability by Wyner and Ziv [16] and almost surely by Ornstein and Weiss [9]. See [13] for a related result. For a comprehensive introduction to the subject consult Shields [11] and the references therein. Recently several interesting results have been obtained regarding convergence rates by other investigators for R_n and related concepts such as the longest match-length, waiting time and redundancy rate, etc. See [1], [4], [5], [7], [12], [15].

In this article we define a modified first return time for estimating entropy for a Markov chain and obtain a very sharp estimate of the convergence rate of its average and propose an algorithm for estimating the entropy for a Markov chain. Since the formula contains a correction terms it approximates the entropy very well. See the last section for simulations.

A *Markov chain* with a stochastic matrix $P = (p_{ij})_{0\le i,j\le k-1}$ is the set of all sample paths on symbols $\{0, 1, \ldots, k-1\}$ with $\Pr(x_{s+1} = j \mid x_s = i) = p_{ij}$. The probability of the cylinder set $[b_1, \ldots, b_n] = [b_1, \ldots, b_n]_{1,\ldots,n}$ or the initial n-block $b_1 \ldots b_n$ is given by $\pi_{b_1} p_{b_1 b_2} \ldots p_{b_{n-1} b_n}$. The entropy of the Markov chain is equal to $h = -\sum_{i,j} \pi_i p_{ij} \log p_{ij}$.

Throughout the paper we assume that P is *irreducible*, i.e., for every (i, j) there exists $n = n(i, j) > 0$ such that $(P^n)_{i,j} > 0$. For $0 \le i \le k-1$, the *period* of a state i, denoted by $\mathrm{Per}(i)$, is the greatest common divisor of those integers $n \ge 1$ for which $(P^n)_{ii} > 0$. The period of P is the greatest common divisor of the numbers $\mathrm{Per}(i)$ that are finite. If P is irreducible, then all the states have the same period, so the period of P is the period of any of its states. A matrix is *aperiodic* if it has period 1. If P is aperiodic, then there exists $n > 0$ satisfying $(P^n)_{i,j} > 0$ for every (i, j). For the details, consult p. 125 in [6]. Let $\pi = (\pi_0, \pi_1, \ldots, \pi_{k-1})$, $\sum_i \pi_i = 1$, $\pi_i > 0$, be the unique left eigenvector corresponding to the simple eigenvalue 1, which is called the *Perron–Frobenius eigenvector*. Put

$$H_0 = -\sum_{i=0}^{k-1} \pi_i \log \pi_i.$$

If P is aperiodic, then all the eigenvalues other than 1 have modulus less than 1. The irreducibility of P implies the ergodicity of the Markov chain, and the aperiodicity gives the mixing property.

DEFINITION 1.2. Given an integer $S \geq 0$ and a block size L, the *modified first return time* $R_{(L,S)}$ is defined by

$$R_{(L,S)}(x) = \min\{j \geq 1 : \xi_1 \dots \xi_L = \xi_{j(L+S)+1} \dots \xi_{j(L+S)+L}\}.$$

DEFINITION 1.3. For $0 < r < 1$, define

$$v(r) \equiv r \sum_{i=1}^{\infty} (1-r)^{i-1} \log i.$$

Put $r = 2^{-L}$. Then the expectation of $\log R_{(L,S)}$ equals $v(r)$ in the case of the Bernoulli $(1/2, 1/2)$-shift. Note that

$$\begin{aligned}\lim_{r\to 0+} [v(r) + \log r] &= \lim_{s \to 1-} [v(1-s) + \log(1-s)] \\ &= \sum_{i=1}^{\infty} \left(\ln \frac{i+1}{i} - \frac{1}{i} \right) \Big/ \ln 2 = -\gamma/\ln 2 = -0.832746\dots,\end{aligned}$$

where $\gamma = \lim_{n\to\infty} (\sum_{i=1}^{n} (1/i) - \ln n)$ is Euler's constant. Hence the expectation of $\log R_{(L,0)}$ is approximately equal to $L - \gamma/\ln 2$ for large L.

In this paper we investigate the speed of convergence of the average of $\log R_{(L,S)}$ to entropy after being properly normalized. The case of Bernoulli processes was solved by Maurer [8]. His algorithm corresponds to $R_{(L,S)}$ for $S = 0$. He showed that the speed is asymptotically proportional to $1/L$ and conjectured that a similar result would hold for Markov chains. In Section 2 we prove the conjecture for Markov chains using the modified algorithm given in Definition 1.2. The dependence on the past memory decreases exponentially, hence the odd-numbered blocks become almost independent of each other as the gap between the neighboring blocks increases.

In his Ph.D. thesis [14] A. J. Wyner discovered that for a stationary aperiodic Markov chain with entropy h we have a second order limit law:

$$\lim_{n\to\infty} \Pr\left(\frac{\log R_n - nh}{\sigma\sqrt{n}} \leq \alpha \right) = \Phi(\alpha)$$

where

$$\Phi(\alpha) = \int_{-\infty}^{\alpha} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2} \right) dx \quad \text{and} \quad \sigma^2 = \lim_{n\to\infty} \frac{\operatorname{Var}(-\log P_n(x))}{n}.$$

I. Kontoyiannis ([4], Corollary 1) showed that for any $\beta > 0$,

$$\log[R_n(x)P_n(x)] = o(n^\beta)$$

almost surely for ergodic Markov chains where $P_n(x)$ is the probability of the initial n-block in x. Later A. J. Wyner ([15], Corollary B5) proved that for any $\varepsilon > 0$,

$$-(1+\varepsilon)\log n \le \log[R_n(x)P_n(x)] \le \log\log n$$

eventually, almost surely for ergodic Markov chains. Note that

$$\begin{aligned} E[\log \Pr(x_1 \ldots x_L)] &= \sum \pi_{a_1} p_{a_1 a_2} \ldots p_{a_{n-1} a_n} \log(\pi_{a_1} p_{a_1 a_2} \ldots p_{a_{n-1} a_n}) \\ &= \sum_i \pi_i \log \pi_i + (L-1) \sum_{i,j} \pi_i p_{ij} \log p_{ij} \\ &= -H_0 - (L-1)h, \end{aligned}$$

where the first sum is taken over all L-blocks $a_1 \ldots a_L$. Hence from the above we have

$$-(1+\varepsilon)\log n \le E[\log R_n] - (n-1)h - H_0 \le \log\log n$$

approximately for large n and we expect that the corresponding result would hold for $R_{(L,S)}$. On the other hand, Kac's Lemma implies that

$$\begin{aligned} E[R_n P_n] &= \sum_{B \in \mathcal{P}_n} E[R_n P_n \,|\, B] \Pr(B) = \sum_{B \in \mathcal{P}_n} E[R_n \,|\, B] P_n(B) \Pr(B) \\ &= \sum_{B \in \mathcal{P}_n} \frac{1}{\Pr(B)} P_n(B) \Pr(B) = \sum_{B \in \mathcal{P}_n} P_n(B) = 1, \end{aligned}$$

hence

$$\log E[R_n P_n] = 0.$$

Therefore we have

$$-(1+\varepsilon)\log n \le E[\log R_n] - (n-1)h - H_0 \le 0$$

for large n. This answers Maurer's question for Markov chains. In fact we prove a sharp estimate of the convergence rate for expectation:

THEOREM. (i) *If P is aperiodic, then*

$$\lim_{L,S \to \infty} E[\log R_{(L,S)} - (L-1)h] - H_0 = -\frac{\gamma}{\ln 2}.$$

(ii) *If P has period $m > 1$, then choose any $m' \ge 1$ such that $m' \,|\, m$. Let (L_k, S_k), $k = 1, 2, 3, \ldots$, be a sequence of pairs of positive integers such that $m' = \gcd(L_k + S_k, m)$ and $L_k, S_k \to \infty$ as $k \to \infty$. Then*

$$\lim_{k \to \infty} E[\log R_{(L_k, S_k)} - (L_k - 1)h] - H_0 = -\frac{\gamma}{\ln 2} - \log m'.$$

Preliminary computer simulations indicate that the divisibility condition may be indispensable in this formulation. Another possible application of the theorem other than estimating entropy is that we may tell whether a given ergodic Markov chain is mixing or not by checking the presence of the term $\log m'$ in the collected data from the source when we have *a priori* knowledge of entropy. Since m is an integer we would find it easily by taking the integer that best represents the experimental data.

2. Proof of the Theorem. The following facts will be needed:

FACT 2.1 ([3], p. 71). *Assume that P is aperiodic. There exist constants c and $0 < d < 1$ such that for any nonnegative vector $\vec{v}$ with $\|\vec{v}\|_1 = 1$ we have*

$$\|\vec{v}P^n - \vec{\pi}\|_\infty \le cd^n.$$

FACT 2.2 ([6]). *Assume that P has period $m > 1$. The set of symbols $\{0, 1, \dots, k-1\}$ is decomposed into a disjoint union $\{0, 1, \dots, k-1\} = \bigcup_{t=0}^{m-1} K_t$ such that if $i \in K_t$ then $(\vec{e_i}P)_j = 0$ if $j \notin K_{t+1}$, where $K_m \equiv K_0$. After a reordering of coordinates, P^m has square matrices $P^{(t)}$ of size $|K_t| \times |K_t|$ on its diagonal. In other words, $(P^{(t)})_{ij} = (P^m)_{ij}$ for $i, j \in K_t$, after a renaming of indices. Each $P^{(t)}$ is irreducible and aperiodic. We let $\vec{\pi}^{(t)}$ denote the Perron–Frobenius eigenvector of $P^{(t)}$. Note that $\vec{\pi} = (1/m)(\vec{\pi}^{(0)}, \dots, \vec{\pi}^{(m-1)})$. By Fact 1 there exist constants $c^{(t)}$ and $d^{(t)}$ such that for any nonnegative $|K_t|$-dimensional vector $\vec{v}^{(t)}$ with $\|\vec{v}^{(t)}\|_1 = 1$,*

$$\|\vec{v}^{(t)}(P^{(t)})^n - \vec{\pi^{(t)}}\|_\infty \le c^{(t)}(d^{(t)})^n.$$

Proof of the Theorem. (i) *Aperiodic case.* Put $C_P = c/\min_i \pi_i$ and suppose S is large enough so that $d^{S+1}C_P < 1$ where c and d are the constants obtained in Fact 2.1. Let T denote the left shift on the Markov chain defined by $T(x_1, x_2, \dots) = (x_2, x_3, \dots)$. For arbitrary blocks $a_1 \dots a_L$ and $b_1 \dots b_L$, we have

$$\begin{aligned}\Pr([b_1 \dots b_L] \cap T^{-(L+S)}[a_1 \dots a_L])& \\ &= \pi_{b_1} p_{b_1 b_2} \dots p_{b_{L-1} b_L} (e_{b_L} P^{S+1})_{a_1} p_{a_1 a_2} \dots p_{a_{L-1} a_L},\end{aligned}$$

where $\vec{e_i}$ is the ith unit row vector. Hence

$$\begin{aligned}|\Pr([b_1 \dots b_L] \cap T^{-(L+S)}[a_1 \dots a_L]) - \Pr(b_1 \dots b_L)\Pr(a_1 \dots a_L)|& \\ &= \pi_{b_1} p_{b_1 b_2} \dots p_{b_{L-1} b_L} |(e_{b_L} P^{S+1})_{a_1} - \pi_{a_1}| p_{a_1 a_2} \dots p_{a_{L-1} b_L} \\ &\le \Pr(b_1 \dots b_L)\Pr(a_1 \dots a_L) d^{S+1} c/\pi_{b_1} \\ &\le \Pr(b_1 \dots b_L)\Pr(a_1 \dots a_L) d^{S+1} \cdot C_P,\end{aligned}$$

and

$$\begin{aligned}&\Pr(a_1\dots a_L)(1-d^{S+1}C_P)\\&\qquad\le \Pr(x_{L+S+1}\dots x_{L+S+L}=a_1\dots a_L\,|\,x_1\dots x_L=b_1\dots b_L)\\&\qquad\le \Pr(a_1\dots a_L)(1+d^{S+1}C_P).\end{aligned}$$

Put $a_1^{(0)}\dots a_L^{(0)}=a_1^{(i)}\dots a_L^{(i)}=a_1\dots a_L$. Then

$$\begin{aligned}&\Pr(R_{(L,S)}=i\mid x_1\dots x_L=a_1\dots a_L)\\&\qquad=\sum_{a_1^{(i-1)}\dots a_L^{(i-1)}\neq a_1\dots a_L}\cdots\sum_{a_1^{(1)}\dots a_L^{(1)}\neq a_1\dots a_L}\prod_{j=1}^{i}A_j\end{aligned}$$

where

$$A_j=\Pr(x_{L+S+1}\dots x_{L+S+L}=a_1^{(j)}\dots a_L^{(j)}\,|\,x_1\dots x_L=a_1^{(j-1)}\dots a_L^{(j-1)})$$

because A_j is equal to the probability of $x_{j(L+S)+1}\dots x_{j(L+S)+L}=a_1^{(j)}\dots a_L^{(j)}$ given the condition $x_{(j-1)(L+S)+1}\dots x_{(j-1)(L+S)+L}=a_1^{(j-1)}\dots a_L^{(j-1)}$. Hence

$$\begin{aligned}\Pr(a_1\dots a_L)(1-d^{S+1}C_P)\cdot U_1&\le \Pr(R_{(L,S)}=i\mid x_1\dots x_L=a_1\dots a_L)\\&\le \Pr(a_1\dots a_L)(1+d^{S+1}C_P)\cdot U_1\end{aligned}$$

where

$$U_1=\sum_{a_1^{(i-1)}\dots a_L^{(i-1)}\neq a_1\dots a_L}\cdots\sum_{a_1^{(1)}\dots a_L^{(1)}\neq a_1\dots a_L}\prod_{j=1}^{i-1}A_j.$$

Since

$$\begin{aligned}&\sum_{a_1^{(i-1)}\dots a_L^{(i-1)}\neq a_1\dots a_L}A_{i-1}\\&\qquad=1-\Pr(x_{L+S+1}\dots x_{L+S+L}=a_1\dots a_L\mid x_1\dots x_L=a_1^{(i-2)}\dots a_L^{(i-2)}),\end{aligned}$$

the sum is bounded by $1-\Pr(a_1\dots a_L)(1+d^{S+1}C_P)$ and $1-\Pr(a_1\dots a_L)\times(1-d^{S+1}C_P)$. Hence

$$\begin{aligned}&\Pr(a_1\dots a_L)(1-d^{S+1}C_P)(1-\Pr(a_1\dots a_L)(1+d^{S+1}C_P))U_2\\&\qquad\le \Pr(R_{(L,S)}=i\,|\,x_1\dots x_L=a_1\dots a_L)\\&\qquad\le \Pr(a_1\dots a_L)(1+d^{S+1}C_P)(1-\Pr(a_1\dots a_L)(1-d^{S+1}C_P))U_2\end{aligned}$$

where

$$U_2=\sum_{a_1^{(i-2)}\dots a_L^{(i-2)}\neq a_1\dots a_L}\cdots\sum_{a_1^{(1)}\dots a_L^{(1)}\neq a_1\dots a_L}\prod_{j=1}^{i-2}A_j.$$

Inductively we have

$$\begin{aligned}
&\Pr(a_1\dots a_L)(1-d^{S+1}C_P)(1-\Pr(a_1\dots a_L)(1+d^{S+1}C_P))^{i-1}\\
&\qquad \le \Pr(R_{(L,S)}=i \mid x_1\dots x_L=a_1\dots a_L)\\
&\qquad \le \Pr(a_1\dots a_L)(1+d^{S+1}C_P)(1-\Pr(a_1\dots a_L)(1-d^{S+1}C_P))^{i-1}.
\end{aligned}$$

Hence

$$\begin{aligned}
&\Pr(a_1\dots a_L)(1-d^{S+1}C_P)\sum_{i=1}^{\infty}(1-\Pr(a_1\dots a_L)(1+d^{S+1}C_P))^{i-1}\log i\\
&\quad \le E[\log R_{(L,S)} \mid x_1\dots x_L=a_1\dots a_L]\\
&\quad \le \Pr(a_1\dots a_L)(1+d^{S+1}C_P)\sum_{i=1}^{\infty}(1-\Pr(a_1\dots a_L)(1-d^{S+1}C_P))^{i-1}\log i.
\end{aligned}$$

Let v be the function in Definition 1.3. The average over all L-blocks $a_1\dots a_L$ is bounded by

$$\begin{aligned}
&E[v(P_L(x)(1+d^{S+1}C_P))]\frac{1-d^{S+1}C_P}{1+d^{S+1}C_P}\\
&\qquad\qquad \le E[\log R_{(L,S)}] \le E[v(P_L(x)(1-d^{S+1}C_P))]\frac{1+d^{S+1}C_P}{1-d^{S+1}C_P}.
\end{aligned}$$

Multiplying by $(1+d^{S+1}C_P)/(1-d^{S+1}C_P)$ and subtracting Lh we have

$$E[v(P_L(x)(1+d^{S+1}C_P))]-Lh \le E[\log R_{(L,S)}]\frac{1+d^{S+1}C_P}{1-d^{S+1}C_P}-Lh$$

or

$$\begin{aligned}
(2.1)\qquad &E[v(P_L(x)(1+d^{S+1}C_P))+\log(P_L(x)(1+d^{S+1}C_P))]\\
&\qquad\qquad -E[\log P_L(x)+L\cdot h]-\log(1+d^{S+1}C_P)\\
&\qquad \le E[\log R_{(L,S)}]\frac{1+d^{S+1}C_P}{1-d^{S+1}C_P}-Lh
\end{aligned}$$

and similarly from the second inequality

$$\begin{aligned}
(2.2)\qquad &E[\log R_{(L,S)}]\frac{1-d^{S+1}C_P}{1+d^{S+1}C_P}-Lh\\
&\qquad \le E[v(P_L(x)(1-d^{S+1}C_P))+\log(P_L(x)(1-d^{S+1}C_P))]\\
&\qquad\quad -E[\log P_L(x)+Lh]-\log(1-d^{S+1}C_P).
\end{aligned}$$

Recall that $v(r)+\log r$ converges to $-\gamma/\ln 2$ as $r\downarrow 0$. For any small $\delta>0$ there exists L_0 such that if $L\ge L_0$ then $P_L(x)(1+d^{S+1}C_P)\le\delta$. Hence we see that the function

$$v(P_L(x)(1+d^{S+1}C_P))+\log(P_L(x)(1+d^{S+1}C_P))$$

is uniformly bounded for such L by taking

$$r=P_L(x)(1+d^{S+1}C_P).$$

The Lebesgue Dominated Convergence Theorem implies that

$$\lim_{L\to\infty} E[v(P_L(x)(1+d^{S+1}C_P)) + \log(P_L(x)(1+d^{S+1}C_P))] = -\gamma/\ln 2.$$

Recall that $E[\log P_L(x) + Lh] = -H_0 + h$.

As L goes to infinity, (2.1) implies

$$-\frac{\gamma}{\ln 2} + H_0 - h - \log(1+d^{S+1}C_P) \le \lim_{L\to\infty}\left(E[\log R_{(L,S)}]\frac{1+d^{S+1}C_P}{1-d^{S+1}C_P} - Lh\right)$$

and similarly (2.2) implies

$$\lim_{L\to\infty}\left(E[\log R_{(L,S)}]\frac{1-d^{S+1}C_P}{1+d^{S+1}C_P} - Lh\right) \le -\frac{\gamma}{\ln 2} + H_0 - h - \log(1-d^{S+1}C_P).$$

Since $d^{S+1}C_P \to 0$ as $S\to\infty$, we have

$$\lim_{L,S\to\infty}(E[\log R_{(L,S)}] - Lh) = -\gamma/\ln 2 + H_0 - h.$$

(ii) *Periodic case.* Take L_k, S_k satisfying the given conditions. We will write L, S for simplicity of notation. Then $\Pr(R_{(L,S)} = i) = 0$ if i is not a multiple of $m/m' \equiv m_0$.

Recall Fact 2.2. Put $C_P = \max_t c^{(t)}/\min_i \pi_i$ and $d = \max_t d^{(t)}$. Choose S large enough so that $d^\beta C_P < m$ where $\beta = [(m_0(S+L) - L + 1)/m]$, the greatest integer that does not exceed $(m_0(S+L) - L + 1)/m$. Put $m_0(S+L) - L + 1 = \beta m + r$, where $0 \le r < m$. Consider $a_1 \ldots a_L$ and $b_1 \ldots b_L$ with positive probability. Suppose $b_L \in K_t$ for some t; then

$$(e_{b_L}P^r)_i = 0 \quad \text{if } i \notin K_{t+r}.$$

First consider the case when a_i, b_i are contained in the same component for every $i = 1, \ldots, L$. Since $m_0(L+S) \equiv 0 \pmod m$, we have $L - 1 \equiv -r \pmod m$ and $b_1 \in K_{t-(L-1)} = K_{t+r}$ and $a_1 \in K_{t+r}$. Hence by Fact 2.2,

$$\begin{aligned} |(e_{b_L}P^{m_0(S+L)-L+1})_{a_1} - m\vec{\pi}_{a_1}| &= |((e_{b_L}P^r)P^{m\beta})_{a_1} - m\vec{\pi}_{a_1}| \\ &= |(v^{(t+r)}(P^{(t+r)})^\beta)_{a_1} - \vec{\pi}^{(t+r)}_{a_1}| \\ &\le c^{(t+r)}(d^{(t+r)})^\beta, \end{aligned}$$

where $v^{(t+r)}$ is a $|K_{t+r}|$-dimensional vector such that

$$(v^{(t+r)})_i = (e_{b_L}P^r)_i \quad \text{if } i \in K_{t+r}.$$

Then

$$\begin{aligned} &|\Pr([b_1\ldots b_L] \cap T^{-m_0(L+S)}[a_1\ldots a_L]) - m\Pr(b_1\ldots b_L)\Pr(a_1\ldots a_L)| \\ &\quad = \pi_{b_1}p_{b_1b_2}\cdots p_{b_{L-1}b_L}|(e_{b_L}P^{m_0(S+L)-L+1})_{a_1} - m\pi_{a_1}|p_{a_1a_2}\cdots p_{a_{L-1}b_L} \\ &\quad = \pi_{b_1}p_{b_1b_2}\cdots p_{b_{L-1}b_L}|((e_{b_L}P^r)P^{\beta m})_{a_1} - m\pi_{a_1}|p_{a_1a_2}\cdots p_{a_{L-1}b_L} \\ &\quad \le \Pr(b_1\ldots b_L)\Pr(a_1\ldots a_L)c^{(t+r)}(d^{(t+r)})^\beta/\pi_{a_1} \\ &\quad \le \Pr(b_1\ldots b_L)\Pr(a_1\ldots a_L)d^\beta C_P \end{aligned}$$

and

$$\begin{aligned}\Pr(b_1 \dots b_L)\Pr(a_1 \dots a_L)(m - d^\beta C_P) \\ &\le \Pr([b_1 \dots b_L] \cap T^{-m_0(L+S)}[a_1 \dots a_L]) \\ &\le \Pr(b_1 \dots b_L)\Pr(a_1 \dots a_L)(m + d^\beta C_P).\end{aligned}$$

Next, if a_i and b_i are not contained in the same component for some i, then

$$\begin{aligned}&\Pr([b_1 \dots b_L] \cap T^{-m_0(L+S)}[a_1 \dots a_L]) \\ &\quad = \pi_{b_1} p_{b_1 b_2} \dots p_{b_{L-1} b_L} (e_{b_L} P^{m_0(S+L)-L+1})_{a_1} p_{a_1 a_2} \dots p_{a_{L-1} b_L} \\ &\quad \le \pi_{b_1} p_{b_1 b_2} \dots p_{b_{i-1} b_i} (e_{b_i} P^{m_0(S+L)})_{a_i} p_{a_i a_{i+1}} \dots p_{a_{L-1} b_L} = 0\end{aligned}$$

since $(e_i P^m)_j = 0$ if $j \notin K_i$ for all $i \in K_i$. Hence we have

$$\begin{aligned}&\Pr(a_1 \dots a_L)(m - d^\beta C_P)(1 - \Pr(a_1 \dots a_L)(1 + d^\beta C_P))^{i-1} \\ &\quad \le \Pr(R_{(L,S)} = m_0 i \mid x_1 \dots x_L = a_1 \dots a_L) \\ &\quad \le \Pr(a_1 \dots a_L)(m + d^\beta C_P)(1 - \Pr(a_1 \dots a_L)(1 - d^\beta C_P))^{i-1}\end{aligned}$$

and $\Pr(R_{(L,S)} = j \mid x_1 \dots x_L = a_1 \dots a_L) = 0$ for all $j \nmid m_0$. Now we proceed as before. ∎

3. Comparison of $\log R_{(L,S)}$ and $\log R_L$. In this section we compare averages and variances of $\log R_{(L,S)}$ and $\log R_L$. The notations are the same as in Section 2. Sometimes we write $P_L(x)$ to denote $\Pr(x_1 \dots x_L)$. As before we have

$$\begin{aligned}&\Pr(a_1 \dots a_L)(1 - d^{S+1} C_P)(1 - \Pr(a_1 \dots a_L)(1 + d^{S+1} C_P))^{i-1} \\ &\quad \le \Pr(R_{(L,S)} = i \mid x_1 \dots x_L = a_1 \dots a_L) \\ &\quad \le \Pr(a_1 \dots a_L)(1 + d^{S+1} C_P)(1 - \Pr(a_1 \dots a_L)(1 - d^{S+1} C_P))^{i-1}.\end{aligned}$$

Hence

$$\begin{aligned}&\Pr(a_1 \dots a_L)(1 - d^{S+1} C_P) \sum_{i=1}^{\infty} (1 - \Pr(a_1 \dots a_L)(1 + d^{S+1} C_P))^{i-1} i \\ &\quad \le E[R_{(L,S)} \mid x_1 \dots x_L = a_1 \dots a_L] \\ &\quad \le \Pr(a_1 \dots a_L)(1 + d^{S+1} C_P) \sum_{i=1}^{\infty} (1 - \Pr(a_1 \dots a_L)(1 - d^{S+1} C_P))^{i-1} i.\end{aligned}$$

Put

$$w(r) = r \sum_{i=1}^{\infty} (1 - r)^{i-1} i, \quad 0 < r < 1.$$

Then $w(r) = 1/r$. Hence averaging over all L-blocks $a_1 \dots a_L$ we obtain

$$E[w(P_L(x)(1+d^{S+1}C_P))]\frac{1-d^{S+1}C_P}{1+d^{S+1}C_P} \le E[R_{(L,S)}] \le E[w(P_L(x)(1-d^{S+1}C_P))]\frac{1+d^{S+1}C_P}{1-d^{S+1}C_P},$$

hence

$$E\left[\frac{1}{P_L(x)}\right]\frac{1-d^{S+1}C_P}{(1+d^{S+1}C_P)^2} \le E[R_{(L,S)}] \le E\left[\frac{1}{P_L(x)}\right]\frac{1+d^{S+1}C_P}{(1-d^{S+1}C_P)^2}.$$

Recall that $E[1/P_L] = E[R_L]$ by Lemma 1.1. Thus

$$E[R_L]\frac{1-d^{S+1}C_P}{(1+d^{S+1}C_P)^2} \le E[R_{(L,S)}] \le E[R_L]\frac{1+d^{S+1}C_P}{(1-d^{S+1}C_P)^2},$$

and we conclude that for sufficiently large S there is not much difference between $E[R_L]$ and $E[R_{(L,S)}]$.

4. Estimation of entropy

4.1. *Aperiodic case.* Since $E[\log R_{(L,S)} - (L-1)h]$ is close to $-\gamma/\ln 2 + H_0$ for sufficiently large L and S, it is recommended that we should approximate the entropy by the formula

$$h_{(L,D,S)} \equiv \frac{E[\log R_{(L+D,S)}] - E[\log R_{(L,S)}]}{D}$$

for any integer $D > 0$.

Example 4.1. Consider the Markov chain associated with the aperiodic matrix

$$P = \begin{pmatrix} 1/2 & 1/4 & 1/4 \\ 1/4 & 0 & 3/4 \\ 1/2 & 1/2 & 0 \end{pmatrix}$$

and initial vector $\vec{\pi} = (10/23, 6/23, 7/23)$. Note that $h = 1.16815951\dots$ The ergodicity of the Markov chain enables us to apply the Birkhoff Ergodic Theorem and we estimate $E[\log R_{(L,S)}]$ by taking the average of $\log R_{(L,S)}$ over 10,000 sample paths $x, T^L x, \dots, T^{9999L}x$, which are obtained by shifting L times to reduce the correlation among the sample values of $\log R_{(L,S)}$. We used a pseudorandom number generator in Fortran 90 to generate a sequence x. Here the sample size is rather large to demonstrate the accuracy of the theoretical prediction and in practical applications a sample of small size will do. The test result is given in Table 1, where Error and S.E.M. denotes $h_{(L,D,S)} - h$ and the standard error mean of $h_{(L,D,S)}$ respectively.

The error $h_{(L,D,S)} - h$ is similar to or less than the size of S.E.M.

Table 1. Test result for Example 4.1

		$S = 10$			$S = 5$		
L	D	$h_{(L,D,S)}$	Error	S.E.M.	$h_{(L,D,S)}$	Error	S.E.M.
7	1	1.11888	−0.04928	0.03230	1.14740	−0.02076	0.03199
7	2	1.17008	0.00192	0.01618	1.17964	0.01148	0.01633
7	3	1.17913	0.01097	0.01082	1.16996	0.00180	0.01094
7	4	1.16273	−0.00543	0.00827	1.16359	−0.00457	0.00839
7	5	1.16677	−0.00139	0.00670	1.16994	0.00178	0.00676
8	1	1.22128	0.05312	0.03328	1.21188	0.04372	0.03295
8	2	1.20925	0.04109	0.01691	1.18124	0.01308	0.01686
8	3	1.17734	0.00918	0.01138	1.16899	0.00083	0.01126
8	4	1.17875	0.01059	0.00858	1.17558	0.00742	0.00856
9	1	1.19723	0.02907	0.03393	1.15060	−0.01756	0.03374
9	2	1.15538	−0.01278	0.01732	1.14754	−0.02062	0.01723
9	3	1.16457	−0.00359	0.01152	1.16346	−0.00470	0.01149

4.2. *Periodic case.* It is recommended that we should approximate the entropy by the formula

$$h_{(L,D,S)} \equiv \frac{E[\log R_{(L+D,S)}] - E[\log R_{(L,S)}]}{D},$$

for any integer $D > 0$ that is a multiple of the period of the chain.

EXAMPLE 4.2. Consider the Markov chain associated with the periodic matrix

$$P = \begin{pmatrix} 0 & 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & 1/4 & 3/4 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 3/4 & 1/4 & 0 & 0 & 0 \end{pmatrix}$$

and initial vector $\vec{\pi} = (1/4, 1/12, 7/48, 3/16, 1/3)$. Its period is $m = 3$. Note that $h = 0.58803255\ldots$ We test this example by the same method as before. The test result is given in Table 2 and Table 3 for $m \,|\, D$ and $m \nmid D$ respectively.

Table 2. Test result for Example 4.2

		$S = 7$			$S = 5$		
L	D	$h_{(L,D,S)}$	Error	S.E.M.	$h_{(L,D,S)}$	Error	S.E.M.
10	3	0.57920	−0.00883	0.01030	0.58531	−0.00272	0.01027
10	6	0.57540	−0.01263	0.00534	0.58072	−0.00731	0.00530
10	9	0.57996	−0.00807	0.00374	0.57883	−0.00920	0.00368
11	3	0.57013	−0.01791	0.01056	0.56322	−0.02482	0.01066
11	6	0.58144	−0.00659	0.00537	0.57941	−0.00863	0.00545
12	3	0.59832	0.01028	0.01092	0.58483	−0.00320	0.01097
12	6	0.58883	0.00080	0.00557	0.58418	−0.00385	0.00556
13	3	0.57161	−0.01642	0.01103	0.57613	−0.01190	0.01097
13	6	0.58034	−0.00769	0.00568	0.57559	−0.01244	0.00572

Table 3. Test result for Example 4.2 with inadequate D

		$S = 7$			$S = 5$		
L	D	$h_{(L,D,S)}$	Error	S.E.M.	$h_{(L,D,S)}$	Error	S.E.M.
10	1	−0.99516	−1.58319	0.03147	2.18021	1.59218	0.03144
10	2	0.57029	−0.01774	0.01549	1.37486	0.78683	0.01558
11	1	2.13574	1.54771	0.03166	0.56950	−0.01853	0.03169
11	2	1.36637	0.77834	0.01612	−0.21214	−0.80017	0.01632
12	1	0.59701	0.00898	0.03203	−0.99378	−1.58181	0.03209
12	2	−0.21268	−0.80071	0.01617	0.56007	−0.02796	0.01655

REFERENCES

[1] A. Dembo and I. Kontoyiannis, *The asymptotics of waiting times between stationary processes, allowing distortion*, Ann. Appl. Probab. 9 (1999), 413–429.

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

[3] J. G. Kemeny and J. L. Snell, *Finite Markov Chains*, Springer, New York, 1976.

[4] I. Kontoyiannis, *Asymptotic recurrence and waiting times for stationary processes*, J. Theoret. Probab. 11 (1998), 795–811.

[5] I. Kontoyiannis, P. H. Algoet, Yu. M. Suhov and A. J. Wyner, *Nonparametric entropy estimation for stationary processes and random fields, with applications to English text*, IEEE Trans. Inform. Theory 44 (1998), 1319–1327.

[6] D. Lind and B. Marcus, *An Introduction to Symbolic Dynamics and Coding*, Cambridge Univ. Press, 1995.

[7] G. Louchard and W. Szpankowski, *On the average redundancy rate of the Lempel–Ziv code*, IEEE Trans. Inform. Theory 43 (1997), 2–8.

[8] U. Maurer, *A universal statistical test for random bit generators*, J. Cryptology 5 (1992), 89–105.

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

[10] C. Shannon, *The mathematical theory of communication*, Bell Sys. Tech. J. 27 (1948), 379–423 and 623–656.

[11] P. C. Shields, *The Ergodic Theory of Discrete Sample Paths*, Grad. Stud. Math. 13 Amer. Math. Soc., 1996.

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

[13] F. M. J. Willems, *Universal data compression and repetition times*, ibid. 35 (1989), 54–58.

[14] A. J. Wyner, *Strong matching theorems and applications to data compression and statistics*, Ph.D. thesis, Stanford Univ., 1993.

[15] —, *More on recurrence and waiting times*, Ann. Appl. Probab., to appear.

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

[17] J. Ziv and A. Lempel, *A universal algorithm for sequential data compression*, ibid. 23 (1977), 337–343.

[18] —, —, *Compression of individual sequences via variable rate coding*, ibid. 24 (1978), 530–536.

Department of Mathematics
Korea Advanced Institute of Science and Technology
Taejon, South Korea
E-mail: choe@euclid.kaist.ac.kr
surper@math.kaist.ac.kr

Received 11 June 1999;
revised 18 January 2000

(3775)