FLUCTUATIONS OF REPETITION TIMES FOR GIBBSIAN SOURCES

Pierre Collet°, Antonio Galves*, Bernard Schmitt†

Abstract : In this paper we consider the class of stochastic stationary sources induced by one-dimensional Gibbs states, with Hölder continuous potentials. We show that the time elapsed before the source repeats its first n symbols converges in law, when suitably renormalized, to a lognormal distribution.

Key words : Gibbsian sources, entrance time, convergence in law, lognormal, entropy.

AMS (MOS) Subject Classification : 94A17, 60F05, 60F10.

° C.N.R.S., Centre de Physique Théorique, Ecole Polytechnique, 91128 Palaiseau Cedex, France. e-mail: collet@pth.polytechnique.fr

* Instituto de Matemática e Estatística, Universidade de São Paulo, BP 66281, 05315-970, São Paulo, SP, Brasil. e-mail: galves@ime.usp.br

† Université de Bourgogne, Département de Mathématiques, Faculté des Sciences Mirande, BP-400, 21011 Dijon Cedex, France. e-mail: schmittb@satie.u-bourgogne.fr

This work was partially supported by USP-ProInter, USP-Cofecub (45-97), CNPq (grant 301301/79), FAPESP (Projeto Temático 95/0790-1) and Pronex (grant 41.96.0923.00).

1.Introduction.

In this paper, we study the law of R_n which is the first time a stochastic stationary source repeats its first n symbols of output. We show that this time suitably renormalized converges in law to a lognormal distribution. The result holds for a large class of ϕ-mixing sources, namely the class of stationary processes induced by one-dimensional Gibbs states.

The original motivation for studying R_n came from the Zempel-Liv and related compression algorithms. A first result on the asymptotics of R_n was obtained by Wyner and Ziv (1989) who proved that for stochastic ergodic sources

$$\lim_{n\to\infty} \frac{1}{n} \log R_n = h \tag{1.1}$$

in probability, where h denotes the entropy of the process. Later on, Ornstein and Weiss (1993) improved the result by showing almost sure convergence.

In the present paper we study the fluctuations of $\log R_n$ around its limiting behavior for Gibbsian sources. We prove that the random variables $(\log R_n - nh)/\sqrt{n}$ converges in law to a Gaussian random variable when n tends to infinity.

As will be clear from the proof, the log-normal asymptotics comes out from the fact that the law of R_n is a mixture of asymptotically exponential laws with fluctuating rates. The main tools of the proof are a theorem about exponential approximations for entrance times in cylinder sets in ϕ-mixing systems proven in Galves and Schmitt (1997) and a central limit theorem for the fluctuations in the Shannon-McMillan-Breiman theorem. For a recent review on exponential and Poissonian approximations we refer to Aldous (1989). Lognormal asymptotics has a long history starting with Galton (1879) and passing through Kolmogorov (1941). For a survey of this field we refer to Crow and Shimizu (1988).

This paper is organized as follows. In section 2 we give the definitions and state the main theorem. The proof is given in section 3.

2.Definitions and main result.

We consider a stationary ergodic discrete time stochastic process $(X_n)_{n\in\mathbf{Z}}$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with values in a finite set A.

For any sequence $(a_k)_{k\in\mathbf{Z}}$ with elements in A, we will denote the partial sequence $(a_i, a_{i+1}, \cdots, a_j)$ by a_i^j, for $i < j$. In particular, $a_i^{+\infty}$ denotes the sequence $(a_j)_{j>i}$. The same notation will be used for the sequence of random variables. Coherently we shall use the notation $\{X_1^n = a_1^n\}$ to denote the cylinder set

$$\{X_j(\omega) = a_j \ , \ j = 1 \, , \cdots , n\} \ .$$

For a finite sequence a_1^n the entrance time $\tau_{a_1^n}$ is defined by

$$\tau_{a_1^n} = \inf \{ j \geq 2 \,|\, X_j^{j+n} = a_1^n \} \ .$$

We recall that the ergodic theorem implies that $\tau_{a_1^n}$ is almost surely finite, provided that $\mathbb{P}\big(\{X_1^n = a_1^n\}\big) > 0$.

We now define the sequence of return times (R_n) by

$$R_n = \tau_{X_1^n} \ .$$

¿From now on we shall always assume that

$$\mathbb{P}\left\{X_1^n = a_1^n\right\} > 0 \ , \tag{2.1}$$

for any $n \geq 1$ and any finite sequence $a_1^n \in A^n$.

A stationary source is Gibbsian if there is a Hölder continuous function $\varphi : A^{\mathbf{N}} \to \mathbf{R}$ and two constants P and $K > 1$ such that for any integer $n > 0$, for any finite sequence $a_1^n \in A^n$, we have

$$K^{-1} \leq \frac{\mathbb{P}\big(\left\{X_1^n = a_1^n\right\}\big)}{e^{-nP + \sum_{j=0}^{n-1} \varphi(a_j^{+\infty})}} \leq K \ , \tag{2.2}$$

where $(a_j^{+\infty})$ is defined by completing the finite sequence a_j^n in an arbitrary way.

We have assumed here for simplicity (through condition (2.1)), that the subshift of finite time associated to the Gibbs measure is the shift on the full alphabet A. We leave to the interested reader the trivial extension of our result to general subshifts of finite type. We refer the reader to Bowen (1975) for more details and properties of Gibbsian sources. We observe that this condition holds for irreducible and aperiodic Markov chains of any finite order, and more generally for chains with complete connections with exponential decay of correlations (see Lalley (1986)).

For the convenience of the reader we recall some important facts about Gibbsian sources which are used below.

Gibbsian processes are exponentially ϕ-mixing, namely there is a sequence $(\phi(l))$ of positive numbers, decreasing exponentially fast to zero such that for all integers n and l larger than or equal to one we have

$$\sup_{A \in \mathcal{F}_0^n \, , \, B \in \mathcal{F}_{n+l}^{\infty}} \frac{|\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)|}{\mathbb{P}(A)\mathbb{P}(B)} \leq \phi(l) \ , \tag{2.3}$$

where $\mathcal{F}_r^s$ denotes the σ-algebra generated by X_r^s.

We observe that in formula (2.2) we can add a constant to φ and subtract it from P without changing the law of the process. In particular we can always assume without loss of generality that

$$\mathbb{E}(\varphi(X_0^{+\infty})) = 0 \ . \tag{2.4}$$

¿From now on we shall assume that (2.4) holds. In this case the constant P in formula (2.2) is equal to the entropy h, and is strictly positive.

In what follows we will have to deal with the sequence of mean zero random variables S_n, $n \geq 1$ defined by

$$S_n = \sum_{j=0}^{n-1} \varphi(X_j^{+\infty}) \ . \tag{2.5}$$

The exponential ϕ-mixing property of Gibbsian sources implies that the variance of $S_n/\sqrt{n}$ converges to

$$\sigma^2 = \mathbb{E}(\varphi(X_0^{+\infty})^2) + 2\sum_{j=1}^{\infty} \mathbb{E}(\varphi(X_0^{\infty})\varphi(X_j^{+\infty})) \ , \tag{2.6}$$

when n tends to infinity.

We now state our main result.

Theorem. *Assume that the stationary process (X_n) taking values in the finite set A is Gibbsian. If $\sigma^2 > 0$, the random variable*

$$\mathcal{L}\left(\frac{\log R_n - nh}{\sigma\sqrt{n}}\right) \Rightarrow \ N(0,1) \ ,$$

as n goes to infinity.

In the statement of the theorem, $\mathcal{L}$ denotes the law of the random variable, and $\Rightarrow$ means weak convergence.

3. Proof of the theorem.

The proof has three steps. In the first one we show that the law of the return time can be approximated by a mixture of laws of entrance times into fixed cylinder sets. In the second step we give an exponential approximation for each of the laws of entrance times in the cylinder sets. These exponential approximations have the probability of the cylinder as respective parameters. The third step uses a central limit theorem to deal with the lognormal fluctuations of the measures of the cylinders around a typical value expressed in terms of the entropy according to the Shannon-McMillan-Breiman theorem.

Observe that in the first two steps we only use the hypothesis that the process is exponentially ϕ-mixing (*i.e.* satisfies condition (2.3)). The stronger Gibbs condition (2.2) is only used in the third step.

The first step is handled in the following two lemmas.

Lemma 1. *There exists a constant $\gamma > 0$ such that for all $n \geq 1$ the following inequality holds*

$$\sup_{a_1^n \in A^n} \mathbb{P}\{X_1^n = a_1^n\} \leq e^{-\gamma n} \ .$$

Proof. This follows almost immediately from the exponential ϕ-mixing condition. We refer the reader to Lemma 1 in Galves and Schmitt (1997) for the details.

Lemma 2. *Assume the source (X_n) is a stationary stochastic process exponentially ϕ-mixing. Then for any sequence (t_n) such that*

$$\lim_{n\to\infty} \frac{t_n}{n} = +\infty$$

we have

$$\lim_{n\to\infty} \left| \mathbb{P}\{R_n > t_n\} - \sum_{a_1^n \in A^n} \mathbb{P}\{X_1^n = a_1^n\}\mathbb{P}\{\tau_{a_1^n} > t_n\} \right| = 0 \ .$$

Proof. By definition for any integer $t > 0$ we have

$$\mathbb{P}\{R_n > t\} = \sum_{a_1^n \in A^n} \mathbb{P}\{X_1^n = a_1^n, \tau_{a_1^n} > t\} . \tag{3.1}$$

We now observe that for any r with $n < r < t$ we have

$$\begin{aligned} &\left|\mathbb{P}\{X_1^n = a_1^n, \tau_{a_1^n} > t\} - \mathbb{P}\{X_1^n = a_1^n\}\mathbb{P}\{\tau_{a_1^n} > t\}\right| \le \\ &\left|\mathbb{P}\{X_1^n = a_1^n, \tau_{a_1^n} > t\} - \mathbb{P}\{X_1^n = a_1^n, X_s^{s+n-1} \neq a_1^n \ , \ r < s \le t\}\right| + \\ &\left|\mathbb{P}\{X_1^n = a_1^n, X_s^{s+n-1} \neq a_1^n \ , \ r < s \le t\} - \mathbb{P}\{X_1^n = a_1^n\}\mathbb{P}\{X_s^{s+n-1} \neq a_1^n \ , \ r < s \le t\}\right| + \\ &\mathbb{P}\{X_1^n = a_1^n\} \left|\mathbb{P}\{X_s^{s+n-1} \neq a_1^n \ , \ r < s \le t\} - \mathbb{P}\{\tau_{a_1^n} > t\}\right| \ . \end{aligned} \tag{3.2}$$

The third term in the right hand side of (3.2) is trivially bounded above by

$$\mathbb{P}\{X_1^n = a_1^n\} \ \mathbb{P}\{X_s^{s+n-1} = a_1^n \ , \ 1 \le s \le r\} \le r \ \mathbb{P}\{X_1^n = a_1^n\}^2 \ . \tag{3.3}$$

By the ϕ-mixing property we obtain that the second term in the right hand side of (3.2) is bounded above by

$$\phi(r - n + 1)\mathbb{P}\{X_1^n = a_1^n\} \ . \tag{3.4}$$

To handle the first term in the right hand side of (3.2) we first observe that it is bounded above by

$$\sum_{u=2}^{r} \mathbb{P}\{X_1^n = a_1^n, X_u^{u+n-1} = a_1^n\} \ . \tag{3.5}$$

Let us decompose the sum in (3.5) in three parts, $2 \le u \le n/3$, $n/3 < u \le n$ and $n < u \le r$ where $n/3$ is a shorthand notation for its integer part. An upper bound for the third piece is provided by the exponential ϕ-mixing property together with lemma 1, namely

$$\sum_{u=n+1}^{r} \mathbb{P}\{X_1^n = a_1^n, X_u^{u+n-1} = a_1^n\} \le C_1\mathbb{P}\{X_1^n = a_1^n\}e^{-\gamma n} \ , \tag{3.6}$$

where

$$C_1 = 1 + \sum_{l=1}^{+\infty} \phi(l) \ .$$

The second piece is bounded above by

$$\sum_{k=n/3}^{n} \mathbb{P}\{X_1^n = a_1^n, X_{n+1}^{n+k-1} = a_{n-k+2}^n\} \ , \tag{3.7}$$

By the ϕ-mixing property and lemma 1, (3.7) is bounded above by

$$\frac{(1+\phi(1))e^{-\gamma(n/3-2)}}{1-e^{-\gamma}}\mathbb{P}\{X_1^n = a_1^n\} . \tag{3.8}$$

To obtain an upper bound for the first piece we need to look in more detail to the set

$$\{X_1^n = a_1^n, X_{n+1}^{n+k-1} = a_{n-k+2}^n\}$$

for small k. This set is non empty only when a_1^n has a particular structure, namely when it is obtained by repeating several times a smaller string of length k. Therefore, for any k, with $2 \le k \le n/3$

$$\sum_{a_1^n \in A^n} \mathbb{P}\{X_1^n = a_1^n, X_{n+1}^{n+k-1} = a_{n-k+2}^n\} \le$$

$$\sum_{b_1^k \in A^k} \mathbb{P}\{X_1^k = X_{k+1}^{2k} = \cdots = X_{(m-1)k+1}^{mk} = b_1^k\} , \tag{3.9}$$

where m is the integer part of $(n+k-1)/k$. The ϕ-mixing property implies the following upper bound for (3.9)

$$(1+\phi(1))\mathbb{P}\{X_1^k = b_1^k\}\mathbb{P}\{X_{k+1}^{2k} = \cdots = X_{(m-1)k+1}^{mk} = b_1^k\} . \tag{3.10}$$

By Lemma 1, for any n large enough (3.10) can be bounded by

$$(1+\phi(1))\mathbb{P}\{X_1^k = b_1^k\}e^{-\gamma k(m-1)} \le (1+\phi(1))\mathbb{P}\{X_1^k = b_1^k\}e^{-\gamma n/2} .$$

where in the last inequality we use $k(m-1) > n/2$ which follows at once from $k \le n/3$.

We now choose $t = t_n$, $r = \min\{n^2 , \sqrt{n\, t_n}\}$ for n large enough. Summing up all the upper bounds the result follows.

We now turn to the second step of the proof. It is based on the following result which is proven in Galves and Schmitt (1997).

Theorem 3. *Assume the source (X_n) is a stationary stochastic process exponentially ϕ-mixing. There are four positive numbers Λ_1, Λ_2, C and c such that for any finite sequence $a_1^n \in A^n$ there is a number $\gamma_{a_1^n}$ satisfying*

$$\Lambda_1 \le \frac{\gamma_{a_1^n}}{\mathbb{P}\{X_1^n = a_1^n\}} \le \Lambda_2$$

such that

$$\sup_{t\ge 0}\left|\mathbb{P}\{\tau_{a_1^n} > t/\gamma_{a_1^n}\} - e^{-t}\right| \le Ce^{-cn} .$$

To proceed with the proof of the Theorem, we first observe that it can be restated as

$$\lim_{n\to\infty} \mathbb{P}\left\{R_n > e^{nh}e^{u\sigma\sqrt{n}}\right\} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{u} e^{-x^2/2}dx , \tag{3.11}$$

for any real u. We now apply Lemma 2 with

$$t_n = e^{nh} e^{u\sigma\sqrt{n}} ,$$

and Theorem 3 in the left hand side of (3.11) to obtain

$$\mathbb{P}\left\{R_n > e^{nh} e^{\sigma u \sqrt{n}}\right\} = \sum_{a_1^n \in A^n} \mathbb{P}\{X_1^n = a_1^n\}\, e^{-\lambda_{a_1^n} \mathbb{P}\{X_1^n = a_1^n\}\, e^{nh} e^{u\sqrt{n}}} + \eta_n(u) . \qquad (3.12)$$

where

$$\lambda_{a_1^n} = \frac{\gamma_{a_1^n}}{\mathbb{P}\{X_1^n = a_1^n\}} ,$$

and $\eta_n(u)$ goes to zero when n goes to infinity.

We now come to the third and final step of the proof. It is based on the following central limit theorem.

Theorem 4. *Assume that the stationary process (X_n) taking values in the finite set A is Gibbsian. Let φ, S_n and σ^2 be defined as in (2.2), (2.5) and (2.6) respectively. If $\sigma^2 > 0$ and assuming (2.4), the random variable*

$$\mathcal{L}\left(\frac{S_n}{\sigma\sqrt{n}}\right) \Rightarrow N(0,1) ,$$

when n goes to infinity.

For a proof of this theorem and related bibliography, we refer to Coelho and Parry (1990).

Let Y_n denote the random variable

$$Y_n = \sum_{a_1^n \in A^n} \mathbf{1}_{\{X_1^n = a_1^n\}}\, e^{-\lambda_{a_1^n} \mathbb{P}\{X_1^n = a_1^n\} e^{nh} e^{u\sqrt{n}}} .$$

We observe that the leading term in the left hand side of (3.12) is equal to $\mathbb{E}\{Y_n\}$. To complete the proof of the Theorem we will show that

$$\liminf_{n\to\infty} \mathbb{E}\{Y_n\} = \limsup_{n\to\infty} \mathbb{E}\{Y_n\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-u} e^{-x^2/2} dx . \qquad (3.13)$$

We first derive a lower bound on the lim inf. For any fixed $\eta > 0$, Markov's inequality implies

$$\mathbb{E}\{Y_n\} \geq e^{-e^{-\eta\sqrt{n}}}\, \mathbb{P}\big\{ \log Y_n \geq -\, e^{-\eta\sqrt{n}} \big\} .$$

We recall that (2.4) implies that the constant P in (2.2) is equal to the entropy h. Therefore, using (2.2), we see that

$$\log Y_n \geq -\lambda_{X_1^n} K e^{S_n + u\sqrt{n}} . \qquad (3.14)$$

Since $\Lambda_1 \leq \lambda_{a_1^n} \leq \Lambda_2$, we use (3.14) for n large enough to get

$$\mathbb{P}\left\{ \log Y_n \geq -\, e^{-\eta\sqrt{n}} \right\} \geq \mathbb{P}\left\{ \frac{S_n}{\sqrt{n}} < -u - 2\eta \right\} .$$

We now use Theorem 4 to conclude that

$$\liminf_{n\to\infty} \mathbb{E}\{Y_n\} \geq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-u-2\eta} e^{-x^2/2} dx . \tag{3.15}$$

Since this is true for any $\eta > 0$ we have the following lower bound.

$$\liminf_{n\to\infty} \mathbb{E}\{Y_n\} \geq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-u} e^{-x^2/2} dx .$$

We now derive a similar upper bound for the lim sup. We have obviously

$$\mathbb{E}\{Y_n\} \leq e^{-e^{\eta\sqrt{n}}}\ \mathbb{P}\left\{ \log Y_n < -e^{\eta\sqrt{n}} \right\} + \mathbb{P}\left\{ \log Y_n \geq -e^{\eta\sqrt{n}} \right\} ,$$

and using Theorem 4 as above we conclude that

$$\limsup_{n\to\infty} \mathbb{E}\{Y_n\} \leq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-u+2\eta} e^{-x^2/2} dx . \tag{3.16}$$

Since (3.16) is true for any $\eta > 0$ we have

$$\limsup_{n\to\infty} \mathbb{E}\{Y_n\} \leq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-u} e^{-x^2/2} dx .$$

This concludes the proof of the Theorem.

References.

Aldous, D. (1989). *Probability approximations via the Poisson clumping heuristic.* Applied Mathematical Sciences, 77. Springer-Verlag, New York-Berlin.

Bowen, R. (1975). *Equilibrium states and ergodic theory of Anosov systems.* Lecture Notes in Mathematics **470**. Springer Verlag, Berlin Heidelberg New York.

Coelho, Z. and Parry, W. (1990). Central limit asymptotics for shifts of finite type. *Israel J. Math.* **69**, 235-249.

Crow, E. and Shimizu, K. editors (1988). *Lognormal distributions. Theory and applications.* Marcel Dekker, Statistics Textbooks and Monographs **88**, New York Basel.

Galton, F. (1879). The geometric mean in vital and social statistics. *Proc. Roy. Soc.* **29**, 365-367.

Galves, A. and Schmitt, B. (1997). Inequalities for hitting times in mixing dynamical systems. *Random Comput. Dyn.* to appear. Preprint available as http://www.ma.utexas.edu/mp_ arc/c/97/97-128.ps.gz.

Kolmogorov, N. (1941). Uber das logarithmisch Normale verteilungesgestz der Dimensione der Teilchen bei Zerstücklung. *Doklady URSS* **31**, 99-101.

Lalley, S. (1986). Regenerative representation for one-dimensional Gibbs states. *Ann. of Probab.* **14**, 1262-1271.

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

Wyner, A. 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.