Limit Theorems for the Sample Entropy of Hidden Markov Chains

Guangyue Han
University of Hong Kong
Email: ghan@hku.hk

***Abstract*—The Shannon-McMillan-Breiman theorem asserts that the sample entropy of a stationary and ergodic stochastic process converges to the entropy rate of the same process (as the sample size tends to infinity) almost surely. In this paper, we restrict our attention to the convergence behavior of the sample entropy of hidden Markov chains. Under certain positivity assumptions, we prove that a central limit theorem (CLT) with some Berry-Esseen bound for the sample entropy of a hidden Markov chain, and we use this CLT to establish a law of iterated logarithm (LIL) for the sample entropy.**

I. Introduction and Notations

Consider a bi-infinite stationary stochastic process $Y = (Y_i, i \in \mathbb{Z})$ on a finite alphabet $\mathcal{Y} = \{1, 2, \cdots, B\}$. The *entropy rate* of Y is defined to be

$$H(Y) = \lim_{n\to\infty} H(Y_1^n)/n,$$

where

$$H(Y_1^n) = -\sum_{y_1^n} p(y_1^n)\log p(y_1^n),$$

here $y_1^n := (y_1, y_2, \cdots, y_n)$ denotes an instance of $Y_1^n := (Y_1, Y_2, \cdots, Y_n)$ and $p(y_1^n)$ denotes the probability mass at y_1^n. It is well known that $H(Y)$ can also be written as

$$H(Y) = \lim_{n\to\infty} H(Y_n|Y_1^{n-1}),$$

where

$$H(Y_n|Y_1^n) = -\sum_{y_1^n} p(y_1^{n-1})\log p(y_n|y_1^{n-1}),$$

here $p(y_n|y_1^{n-1})$ denotes the conditional probability mass at y_n given y_1^{n-1}.

We call $-\log P(Y_1^n)/n$ the n-th order *sample entropy* of Y. If Y is also ergodic, the celebrated Shannon-McMillan-Breiman theorem asserts that the n-th order sample entropy of Y converges to $H(Y)$ as $n \to \infty$ almost surely. The Shannon-McMillan-Breiman theorem can be viewed as an analog of the law of large numbers, a fundamental limit theorem in probability theory. So, it is natural to ask if analogs of other limit theorems in probability theory, such as the central limit theorem (CLT) and the law of iterated logarithm (LIL), also hold for the sample entropy. It turns out that CLT and LIL do not hold when we assume Y is as general a process as stationary and ergodic; so, in this paper, we restrict our attention to hidden Markov chains (some special stochastic processes which will be defined later).

From now on, assume that Y is a stationary finite-state Markov chain with transition probability matrix Δ with entries

$$\Delta(i,j) = P(Y_1 = j|Y_0 = i), \qquad 1 \le i, j \le B.$$

A *hidden Markov chain* Z is a process of the form $Z = \Phi(Y)$, where Φ is a deterministic function defined on $\mathcal{Y}$ with values from a finite alphabet $\mathcal{Z}$. Often a hidden Markov chain is alternatively defined as a Markov chain observed when passing though a discrete-time memoryless noisy channel. It is well known that the two definitions are equivalent. For the Markov chain Y, $H(Y)$ has a simple analytic form:

$$H(Y) = -\sum_{i,j} P(Y_0 = i)\Delta(i,j)\log\Delta(i,j).$$

For the hidden Markov chain Z, Blackwell [6] gave an integral formula for $H(Z)$, however using a measure that is typically too complicated for effective computation of $H(Z)$. So far, there is no simple and explicit formula for $H(Z)$. So, many approaches have been adopted to compute and estimate $H(Z)$ instead: Blackwell's measure has been used to bound the entropy rate [22], a variation on the classical Birch bounds [5] can be found in [9] and a new numerical approximation of $H(Z)$ has been proposed in [20]. Generalizing Blackwell's idea, an integral formula for the derivatives of $H(Z)$ has been derived in [27]. In another direction, [1], [18], [22], [31], [32], [21], [12], [14], [15], [16], [24], [27] have studied the variation of the entropy rate as parameters of the underlying Markov chain vary.

Another interesting approach, which has greatly motivated this work, is to use Monte Carlo simulation to approximate $H(Z)$: Recently, based on the Shannon-McMillan-Breiman theorem, efficient Monte Carlo methods for approximating $H(Z)$ were proposed independently by Arnold and Loeliger [2], Pfister, Soriaga and Siegel [25], Sharma and Singh [28]. The limiting behavior of the sample entropy of a hidden Markov chain, which governs the convergence behavior of such algorithms, is then of great interest. In this direction, a CLT [26] for the sample entropy is derived as a corollary of a CLT for the top Lyapunov exponent of a product of random matrices; a functional CLT is also established in [17]. In essence, both of the two CLTs are proved using effective Martingale approximations of the sample entropy.

In this paper, adapting some standard techniques for proving limit theorems for mixing sequences, we further characterize the limiting behavior of the sample entropy of Z under

978-1-4577-0595-3/11/$26.00 ©2011 IEEE

certain positivity assumptions. Formally, define $X = (X_i, i = 1, 2, \cdots)$ as

$$X_i = -\log P(Z_i|Z_1^{i-1}) - H(Z_i|Z_1^{i-1}),$$

and

$$S_n = \sum_{i=1}^{n} X_i, \qquad \sigma_n^2 = \mathrm{Var}(\mathrm{S_n}).$$

Unless specified otherwise, we assume, throughout the paper, that

(I) Δ is a strictly positive matrix; and
(II) $\sigma > 0$, where $\sigma^2 = \lim_{n\to\infty} \sigma_n^2/n$ (the existence of the limit under Condition (I) will be established in Lemma 2.5 and Remark 2.6).

Remark 1.1: It can be shown that given Condition (I) is satisfied, Condition (II) is equivalent to either one of the following

1) $\lim_{n\to\infty} E[S_n^2] = \infty$.
2) $\limsup_{n\to\infty} E[S_n^2] = \infty$.

We will prove the following central limit theorem with a Berry-Esseen bound (see [4], [10]; such bound, which is absent in [26], [17], characterizes rate of convergence of the CLT).

Theorem 1.2: Under Conditions (I) and (II), for any $\varepsilon > 0$, there exists $C > 0$ such that for any n

$$\sup_x |P(S_n/\sigma_n < x) - \int_{-\infty}^{x} (2\pi)^{-1/2} \exp(-y^2/2) dy| \le C n^{-1/11+\varepsilon}.$$

We will use the above CLT to prove the following law of iterated logarithm.

Theorem 1.3: Under Conditions (I) and (II), we have

$$\limsup_{n\to\infty} \frac{S_n}{(2n\sigma^2 \log\log n\sigma^2)^{1/2}} = 1 \qquad a.s.$$

Most of the proofs in this paper are omitted due to space limit; we refer to [13] for a complete version of this paper.

II. Key Lemmas

This section includes several key lemmas, among which Lemmas 2.1, 2.2, 2.3 require Condition (I) only. Here, we remark that in this paper C is used to denote a constant, which may not be the same on each appearance.

The following lemma can is well-known (see, e.g., [12] for a rigorous proof).

Lemma 2.1: There exist $C > 0$ and $0 < \rho < 1$ such that for any two hidden Markov sequences $z_{-m}^0, \hat{z}_{-\hat{m}}^0$ with $z_{-n}^0 = \hat{z}_{-n}^0$ (here $m, \hat{m} \ge n \ge 0$), we have

$$|p(z_0|z_{-m}^{-1}) - p(\hat{z}_0|\hat{z}_{-\hat{m}}^{-1})| \le C\rho^n.$$

Consequently, there exists $C > 0$ and $0 < \rho < 1$ such that for any $n, l \ge 0$,

$$|\log p(z_0|z_{-n-l}^{-1}) - \log p(z_0|z_{-n}^{-1})| \le C\rho^n,$$
$$|H(Z_0|Z_{-n-l}^{-1}) - H(Z_0|Z_{-n}^{-1})| \le C\rho^n.$$

For a stationary stochastic process $T = T_{-\infty}^{\infty}$, let $\mathcal{B}(T_i^j)$ denote the σ-field generated by $T_k, k = i, i+1, \cdots, j$. Define

$$\psi(n) = \sup_{U\in\mathcal{B}(T_{\infty}^{-n}), V\in\mathcal{B}(T_0^{\infty}), P(U)>0, P(V)>0} \frac{|P(V|U) - P(V)|}{P(V)}.$$

T is said to be a *ψ-mixing sequence* if $\psi(n) \to 0$ as $n \to \infty$. It is well known [7] that a finite-state irreducible and aperiodic Markov chain is a ψ-mxing sequence, and the corresponding $\psi(n)$ exponentially decays as $n \to \infty$. The following lemma asserts that under Condition (I), Z is a ψ-mixing sequence and the corresponding $\psi(n)$ exponentially decays as $n \to \infty$. An excellent survey on various mixing sequences can found in [7]; for a comprehensive exposition to the vast literature on this subject, we refer to [8].

Lemma 2.2: Z is a ψ-mixing sequence and there exist $C > 0$ and $0 < \lambda < 1$ such that for any positive n, $\psi(n) \le C\lambda^n$.

The following lemma shows that for a fixed $j > 0$, $E[X_iX_{i+j}]$ exponentially converges as $i \to \infty$; and for any $i < j$, $E[X_iX_j]$ exponentially decays in $j - i$.

Lemma 2.3: 1) There exist $C > 0$ and $0 < \rho < 1$ such that for all $i, j \ge 0$,

$$|E[X_{i+1}X_{i+1+j}] - E[X_iX_{i+j}]| \le C\rho^i.$$

2) There exist $C > 0$ and $0 < \theta < 1$ such that for any positive $i < j$,

$$|E[X_iX_j]| \le C\theta^{j-i}.$$

Remark 2.4: By Part 1) of Lemma 2.3, for any fixed j, the sequence $E[X_iX_{i+j}]$, $i = 1, 2, \cdots$, is a Cauchy sequence that exponentially converges. For any fixed j, let $a_j = \lim_{i\to\infty} E[X_iX_{i+j}]$. Then by Part 2), $|a_j|$ exponentially decays as $j \to \infty$; consequently, we deduce (for later use) that $a_0 + 2\sum_{j=1}^{\infty} a_j$ converges.

Lemma 2.5: For any $0 < \alpha < 1$, there exists $C > 0$ such that for any m and n,

$$\left| \frac{E[(S_{n+m} - S_m)^2]}{n} - (a_0 + 2\sum_{j=1}^{\infty} a_j) \right| \le Cn^{-\alpha}$$

here, recall that, as defined in Remark 2.4, $a_j = \lim_{i\to\infty} E[X_iX_{i+j}]$.

Remark 2.6: Choosing m in Lemma 2.5 to be 0, we deduce that $\lim_{n\to\infty} \sigma_n^2/n$ exists and is equal to $\sigma^2 = a_0 + 2\sum_{j=1}^{\infty} a_j$.

Lemma 2.7: There exists $C > 0$ such that for all m and n

$$E[|S_{n+m} - S_m|^3] \le Cn^{3/2}.$$

III. Central Limit Theorem

Fix an arbitrarily small $\varepsilon_0 > 0$, and let $p = p(n) = \lfloor n^{3/11+\varepsilon_0} \rfloor, q = q(n) = \lfloor n^{\varepsilon_0} \rfloor$. Choose $k = k(n)$ such that $kp + (k-1)q \le n < (k+1)p + kq$; one easily checks that $k = O(n^{8/11-\varepsilon_0})$. Then, for $1 \le i \le k$, define

$$\zeta_i = X_{(i-1)(p+q)+1} + \cdots + X_{ip+(i-1)q}.$$

For $1 \le i \le k-1$, define

$$\eta_i = X_{ip+(i-1)q+1} + \cdots + X_{ip+iq},$$

and define

$$\eta_k = \begin{cases} X_{kp+(k-1)q+1} + \cdots + X_n & \text{if } kp + (k-1)q + 1 \le n \\ 0 & \text{otherwise.} \end{cases}$$

Now S_n can be rewritten as a sum of ζ-"blocks" and η-"blocks":

$$S_n = S'_n + S''_n := \sum_{i=1}^{k} \zeta_i + \sum_{i=1}^{k} \eta_i.$$

The above so called "Bernstein blocking method" is a standard technique to the proof of limit theorems for a variety of mixing sequences. Roughly speaking, the partial sum S_n is partitioned into "long" blocks $\zeta_1, \zeta_2, \cdots, \zeta_k$ and "short" blocks $\eta_1, \eta_2, \cdots, \eta_k$. Under certain mixing conditions, all long blocks are "weakly dependent" on each other, while all short blocks are "negligible" in some sense.

With lemmas in Section II established, the remainder of the proof of Theorem 1.2 becomes more or less standard, which can be roughly outlined as follows:

1) We first show $E[\exp(itS'_n/\sigma_n)]$ and $\prod_{j=1}^{k} E[\exp(it\zeta_j/\sigma_n)$ are "close" (see Lemma 3.1).
2) Standard analysis shows that $\prod_{j=1}^{k} E[\exp(it\zeta_j/\sigma_n)$ and $\exp(-t^2/2)$ are "close" (see Lemma 3.2).
3) Then by Esseen's Lemma [10], $P(S'_n/\sigma_n < x)$ and $\int_{-\infty}^{x}(2\pi)^{-1/2}\exp(-y^2/2)dy$ are "close" (see Lemma 3.4).
4) Finally, since S''_n are "negligible", we conclude, in the proof of Theorem 1.2, that $P(S_n/\sigma_n < x)$ and $P(S'_n/\sigma_n < x)$ are "close", and thus $P(S_n/\sigma_n < x)$ and $\int_{-\infty}^{x}(2\pi)^{-1/2}\exp(-y^2/2)dy$ are "close".

Lemma 3.1: There exists $C > 0$ and $0 < \rho_1 < 1$ such that for all n and $|t| \le n^{1/11}$,

$$|E[\exp(itS'_n/\sigma_n)] - \prod_{j=1}^{k} E[\exp(it\zeta_j/\sigma_n)]| \le C\rho_1^{q(n)}.$$

Lemma 3.2: There exists $C > 0$ such that for all n and $|t| \le n^{1/11}$,

$$|\prod_{j=1}^{k} E[\exp(it\zeta_j/\sigma_n)] - \exp(-t^2/2)| \le Cn^{-1/11+\varepsilon_0/2}.$$

The following lemma is a version of Esseen's lemma, which gives upper bounds on the difference between two distribution functions using the difference between the two corresponding characteristic functions. We refer to page 314 of [29] for a standard proof.

Lemma 3.3: Let F and G be distribution functions with characteristic functions ϕ_F and ϕ_G, respectively. Suppose that F and G each has mean 0, and G has a derivative g such that $|g| \le M$. Then

$$\sup_x |F(x) - G(x)| \le \frac{1}{\pi}\int_{-T}^{T}\left|\frac{\phi_F(t) - \phi_G(t)}{t}\right| dt + \frac{24M}{\pi T}$$

for every $T > 0$.

Lemma 3.4: For any $\varepsilon > 0$, there exists $C > 0$ such that for all n

$$\sup_x \left|P(S'_n/\sigma_n < x) - \int_{-\infty}^{x}(2\pi)^{-1/2}\exp(-y^2/2)dy\right| \le Cn^{-1/11+\varepsilon}.$$

We are now ready to prove Theorem 1.2.

Proof of Theorem 1.2:

Let A denote the event "$|S''_n|/\sigma_n \le n^{-1/11}$". Then

$$\begin{aligned}
&|P(S'_n/\sigma_n \le x) - P(S_n/\sigma_n \le x)| \\
&\le |P(S'_n/\sigma_n \le x, A) - P(S_n/\sigma_n \le x, A)| \\
&+|P(S'_n/\sigma_n \le x, A^c) - P(S_n/\sigma_n \le x, A^c)| \\
&\le |P(S'_n/\sigma_n \le x, A) - P(S_n/\sigma_n \le x, A)| + P(|S''_n|/\sigma_n > n^{-1/11}).
\end{aligned}$$

Applying Lemma 3.4, we have, for any $\varepsilon > 0$, there exists $C_1 > 0$ such that for any n

$$\begin{aligned}
&|P(S'_n/\sigma_n \le x, A) - P(S_n/\sigma_n \le x, A)| \\
&\le \max\{P(S'_n/\sigma_n \le x + n^{-1/11}, A) - P(S'_n/\sigma_n \le x, A), \\
&\quad P(S'_n/\sigma_n \le x, A) - P(S'_n/\sigma_n \le x - n^{-1/11}, A)\} \\
&\le C_1 n^{-1/11+\varepsilon} + \int_{-n^{-1/11}}^{n^{-1/11}}(2\pi)^{-1/2}\exp(-y^2/2)dy \\
&= O(n^{-1/11+\varepsilon}) + O(n^{-1/11}) = O(n^{-1/11+\varepsilon}).
\end{aligned}$$

Applying Lemma 2.5 and Lemma 2.2, we deduce that for some $0 < \theta < 1$,

$$\begin{aligned}
\frac{E[(S''_n)^2]}{\sigma_n^2} &= \frac{\sum_{i=1}^{k} E[\eta_i^2] + \sum_{i<j} E[\eta_i\eta_j]}{\sigma_n^2} \\
&= \frac{k(n)q(n)\sigma^2(1+o(1)) + O(n^2\theta^{q(n)})}{\sigma_n^2} = O(n^{-3/11}).
\end{aligned}$$

Also, by the Markov inequality, we have

$$P(|S''_n|/\sigma_n > n^{-1/11}) \le \frac{E[(S''_n)^2]}{\sigma_n^2 n^{-2/11}} = O(n^{-1/11}).$$

The theorem then immediately follows. ∎

Remark 3.5: If Condition (II) fails, i.e., $\lim_{n\to\infty}\sigma_n^2/n = 0$, then a CLT of degenerated form holds for $(X_i, i \in \mathbb{N})$; more precisely, the distribution of $(X_1 + X_2 + \cdots + X_n)/\sqrt{n}$ converges to that of a centered normal distribution with variance 0, i.e., a point mass at 0, as $n \to \infty$. This is can be readily checked since for any $\varepsilon > 0$, by the Markov inequality, we have

$$P(|(X_1 + X_2 + \cdots + X_n)|/\sqrt{n} \ge \varepsilon|) \le \sigma_n^2/(n\varepsilon^2) \to 0 \text{ as } n \to \infty.$$

IV. Law of Iterated Logarithm

From the central limit theorem with a Berry-Esseen bound (Theorem 1.2), we only need to follow a standard "track" to establish the law of iterated logarithm (Theorem 1.3). In particular, we closely follow the proof of Reznik's law of the iterated logarithm for a stationary ϕ-mixing sequence (see page 307 of [29]):

1) As an immediately corollary of Theorem 1.2, the following Lemma 4.1 gives bounds on the tail probability of S_n.
2) We then slightly modify Reznik's maximal inequality to to obtain our maximal inequality in Lemma 4.2;
3) Finally, we are ready for the proof of Theorem 1.3, where some necessary modifications are incorporated

into the original Reznik's proof to deal with the complications stemming from the fact that X is not a stationary mixing sequence.

Lemma 4.1: For any $|\delta| < 1$ and $\alpha > 0$, we have

$$(\log \sigma_n^2)^{-(1+\delta)^2(1+\alpha)} < P(S_n > (1+\delta)(2\sigma_n^2 \log\log \sigma_n^2)^{1/2}) < (\log \sigma_n^2)^{-(1+\delta)^2(1-\alpha)}$$

for n sufficiently large.

Lemma 4.2: For any $x > 0$, $0 < \alpha < 1/2$ and $C > 0$, we have

$$P(\max_{j \le n} S_j > x) \le 2P(S_n > x - 2\sigma_n) + Cn^{-\alpha},$$

for sufficiently large n.

We are now ready to prove Theorem 1.3.

Proof of Theorem 1.3: We first show that

$$\limsup_{n\to\infty} \frac{S_n}{(2n\sigma^2 \log\log n\sigma^2)^{1/2}} \le 1 \qquad a.s.; \tag{1}$$

equivalently, we show that for any $\delta > 0$,

$$P(\frac{S_n}{(2\sigma_n^2 \log\log \sigma_n^2)^{1/2}} > 1+\delta \ \ i.o.) = 0, \tag{2}$$

here we remind the reader that by Remark 2.6, $\sigma_n^2 = n(\sigma^2 + o(1))$ and "*i.o.*" means "infinitely often".

Fox fixed $M > 1$, define $n_j = M^j$, $j = 1, 2, \cdots$. One then checks that

$$\text{“}\frac{S_n}{(2\sigma_n^2 \log\log \sigma_n^2)^{1/2}} > 1+\delta \ \ i.o.\text{”}$$

$$\subset \text{“} \max_{n \le n_{j+1}} \frac{S_n}{(2\sigma_{n_j}^2 \log\log \sigma_{n_j}^2)^{1/2}} > 1+\delta \ \ i.o.\text{”}.$$

So, to prove (2), it suffices (by the Borel-Cantelli Lemma) to show that

$$\sum_{j=1}^{\infty} P(\max_{n \le n_{j+1}} \frac{S_n}{(2\sigma_{n_j}^2 \log\log \sigma_{n_j}^2)^{1/2}} > 1+\delta) < \infty; \tag{3}$$

in order to prove (3), by Lemma 4.2 and the fact that

$$\sum_{j=1}^{\infty} n_{j+1}^{-\alpha_2} = \sum_{j=1}^{\infty} M^{-(j+1)\alpha_2} < \infty,$$

it suffices to prove that

$$\sum_{j=1}^{\infty} P(S_{n_{j+1}} > (1+\delta)(2\sigma_{n_j}^2 \log\log \sigma_{n_j}^2)^{1/2} - 2\sigma_{n_{j+1}}) < \infty. \tag{4}$$

Note that there exists $0 < \delta_1 < \delta$ such that for j sufficiently large,

$$(1+\delta)(2\sigma_{n_j}^2 \log\log \sigma_{n_j}^2)^{1/2} - 2\sigma_{n_{j+1}} > (1+\delta_1)(2\sigma_{n_j}^2 \log\log \sigma_{n_j}^2)^{1/2}. \tag{5}$$

Applying Lemma 4.1 with α chosen such that $(1+\delta_1)^2(1-\alpha) > 1$, we deduce that

$$\sum_{j=1}^{\infty} P(S_{n_{j+1}} > (1+\delta_1)(2\sigma_{n_j}^2 \log\log \sigma_{n_j}^2)^{1/2}) \le \sum_{j=1}^{\infty} (\log \sigma_{n_j}^2)^{-(1+\delta_1)^2(1-\alpha)} = \sum_{j=1}^{\infty} O(j^{-(1+\delta_1)^2(1-\alpha)}) < \infty. \tag{6}$$

Immediately, (4) and then (3) and then (2) follow. Here, we remark that the same argument as above with X_i replaced by $-X_i$ leads to

$$\liminf_{n\to\infty} \frac{S_n}{(2n\sigma^2 \log\log n\sigma^2)^{1/2}} \ge -1 \qquad a.s. \tag{7}$$

For the other direction, we next show that

$$\limsup_{n\to\infty} \frac{S_n}{(2n\sigma^2 \log\log n\sigma^2)^{1/2}} \ge 1 \qquad a.s.;$$

equivalently, we show that for any $\delta > 0$,

$$P(\frac{S_n}{(2\sigma_n^2 \log\log \sigma_n^2)^{1/2}} > 1-\delta \ \ i.o.) = 1. \tag{8}$$

For fixed $N > 1$ and $\delta > 0$, let $C_n(\delta)$ be the event

$$\text{“}S_{N^n} - S_{N^{n-1}+N^{n/2}} > (1-\delta)g(N^n - N^{n-1} - N^{n/2})\text{”},$$

where $g(n) = (2n\sigma^2 \log\log n\sigma^2)^{1/2}$. With Lemmas 2.1 and 4.1, one checks that there exists $0 < \delta_2 < \delta$ such that for a given $\alpha > 0$

$$P(C_n(\delta)) \ge P(S_{N^n - N^{n-1} - N^{n/2}} > (1-\delta_2)g(N^n - N^{n-1} - N^{n/2}))$$

$$\ge \log(N^n - N^{n-1} - N^{n/2})^{-(1-\delta_2)^2(1+\alpha)}/2 \tag{9}$$

for sufficiently large n. From now on, we choose $\alpha > 0$ such that $(1-\delta_2)^2(1+\alpha) < 1$. If n and N are large enough, we have

$$N^n - N^{n-1} - N^{n/2} \ge N^n/2,$$

which, together with (9), implies that for any $\delta > 0$

$$\sum_{n=1}^{\infty} P(C_n(\delta)) = \infty. \tag{10}$$

Similarly, let $\hat{C}_n(\delta)$ be the event

$$\text{“} \sum_{i=N^{n-1}+N^{n/2}+1}^{N^n} - \log p(Z_i | Z_{N^{n-1}+N^{n/4}}^{i-1}) - H(Z_i | Z_{N^{n-1}+N^{n/4}}^{i-1}) > (1-\delta)g(N^n - N^{n-1} - N^{n/2})\text{”}.$$

Applying Lemma 2.1, we deduce that that for any $\delta' > 0$, there exists $0 < \delta < \delta'$ such that for sufficiently large n,

$$\hat{C}_n(\delta') \supset C_n(\delta),$$

which, together with (10), implies that for any $\delta' > 0$

$$\sum_{n=1}^{\infty} P(\hat{C}_n(\delta')) = \infty.$$

Again, by Lemma 2.1, for any $\delta > 0$, there exists $0 < \delta'' < \delta$ such that for sufficiently large n,

$$\hat{C}_n(\delta'') \subset C_n(\delta).$$

It then follows from an iterative application of Lemma 2.2 that there exists $0 < \theta < 1$ such that for any n, l,

$$\begin{aligned} P(\cap_{m=n}^{n+l} C_m^c(\delta)) &\leq P(\cap_{m=n}^{n+l} \hat{C}_m^c(\delta'')) \\ &= \prod_{m=n}^{n+l} P(\hat{C}_m^c(\delta'')) + \sum_{m=n}^{n+l} O(\theta^{N^{m/4}}) \\ &= \prod_{m=n}^{n+l} (1 - P(\hat{C}_m(\delta''))) + \sum_{m=n}^{n+l} O(\theta^{N^{m/4}}) \\ &\leq \exp(-\sum_{m=n}^{n+l} P(\hat{C}_m(\delta''))) + \sum_{m=n}^{l} O(\theta^{N^{m/4}}). \end{aligned} \tag{11}$$

So, as $l, n \to \infty$, $P(\cap_{m=n}^{n+l} C_m^c(\delta)) \to 0$, or equivalently, for any $\delta > 0$,

$$P(C_n(\delta) \ \ i.o.) = 1. \tag{12}$$

Let B_n be the event "$S_{N^{n-1}+N^{n/2}} > -2g(N^{n-1} + N^{n/2})$". It then follows from (7) that

$$P(B_n \ \ i.o.) = 1,$$

which, together with (12), implies that for any $\hat{\delta} > 0$

$$P(B_n \cap C_n(\hat{\delta}) \ \ i.o.) = 1. \tag{13}$$

One then checks that for $\delta > 0$, there exists $0 < \hat{\delta} < \delta$ such that for sufficiently large n,

$$\begin{aligned} &\text{“}S_{N^n} > (1-\delta)g(N^n) \ \ i.o.\text{”} \\ \supset \ &\text{“}S_{N^n} > (1-\hat{\delta})g(N^n - N^{n-1} - N^{n/2}) - 2g(N^{n-1} + N^{n/2}) \ \ i.o.\text{”} \\ \supset \ &\text{“}B_n \cap C_n(\hat{\delta}) \ \ i.o.\text{”}. \end{aligned}$$

It then follows from (13) that

$$P(S_{N^n} > (1-\delta)g(N^n) \ \ i.o.) = 1,$$

which immediately implies (8). ∎

REFERENCES

[1] L. Arnold, V. M. Gundlach and L. Demetrius. Evolutionary formalism for products of positive random matrices. *Annals of Applied Probability*, 4:859–901, 1994.

[2] D. Arnold and H. Loeliger. The information rate of binary-input channels with memory. Proc. 2001 IEEE Int. Conf. on Communications, Helsinki, Finland, pp. 2692–2695, June 11-14 2001.

[3] D. M. Arnold, H.-A. Loeliger, P. O. Vontobel, A. Kavcic, W. Zeng, Simulation-Based Computation of Information Rates for Channels With Memory. *IEEE Trans. Information Theory*, **52**, 3498–3508, 2006.

[4] A. Berry. The accuracy of the Gaussian Approximation to the sum of independent variates. *Trans. Amer. Math. Soc.* 49, 122-126., 1941.

[5] J. J. Birch. Approximations for the entropy for functions of Markov chains. *Ann. Math. Statist.*, 33:930–938, 1962.

[6] D. Blackwell. The entropy of functions of finite-state Markov chains. *Trans. First Prague Conf. Information Thoery, Statistical Decision Functions, Random Processes*, pages 13–20, 1957.

[7] R. Bradley. Basic Properties of Strong Mixing Conditions. A Survey and Some Open Questions Probability Surveys, Volume 2 (2005), 107-144.

[8] R. C. Bradley. Introduction to Strong Mixing Conditions. Volumes 1,2 and 3. Kendrick Press, 2007.

[9] S. Egner, V. Balakirsky, L. Tolhuizen, S. Baggen and H. Hollmann. On the entropy rate of a hidden Markov model. In *Proceedings of the 2004 IEEE International Symposium on Information Theory*, page 12, Chicago, U.S.A., 2004.

[10] C. Esseen. On the Lyapunov limit of error in the theory of probability. *Ark. Math. Astr. och Fysik.*, 28A, 1-19, 1942.

[11] R. Gharavi and V. Anantharam. An upper bound for the largest Lyapunov exponent of a Markovian product of nonnegative matrices. *Theoretical Computer Science*, Vol. 332, Nos. 1-3, pp. 543 -557, February 2005.

[12] G. Han and B. Marcus. Analyticity of entropy rate of hidden Markov chains. *IEEE Transactions on Information Theory*, Volume 52, Issue 12, December, 2006, pages: 5251-5266.

[13] G. Han. Limit Theorems for the Sample Entropy of Hidden Markov Chains. Available at http://arxiv.org/abs/1102.0365

[14] G. Han and B. Marcus. Derivatives of Entropy Rate in Special Familes of Hidden Markov Chains. *IEEE Transactions on Information Theory*, Volume 53, Issue 7, July 2007, Page(s):2642 -2652.

[15] G. Han and B. Marcus. Concavity of Mutual Information Rate for Input-Restricted Finite-State Memoryless Channels at High SNR. *IEEE International Symposium on Information Theory*, Seoul, Korea, June 28-July 3, 2009, Page(s): 1654 - 1658.

[16] G. Han and B. Marcus. Asymptotics of Entropy Rate in Special Families of Hidden Markov Chains. *IEEE Transactions on Information Theory*, Volume 56, Issue 3, March 2010, Page(s):1287-1295.

[17] T. Holliday, A. Goldsmith, and P. Glynn. Capacity of Finite State Channels Based on Lyapunov Exponents of Random Matrices. *IEEE Transactions on Information Theory*, Volume 52, Issue 8, Aug. 2006, Page(s):3509 - 3532.

[18] P. Jacquet, G. Seroussi, and W. Szpankowski. On the Entropy of a Hidden Markov Process (extended abstract). *Data Compression Conference*, 362–371, Snowbird, 2004.

[19] P. Jacquet, G. Seroussi, and W. Szpankowski. Noisy Constrained Capacity. *International Symposium on Information Theory*, 986-990, Nice, France, 2007.

[20] J. Luo and D. Guo. On the entropy rate of hidden Markov processes observed through arbitrary memoryless channels. *IEEE Trans. Inform. Theory*, vol. 55, pp. 1460-1467, April 2009.

[21] C. Nair, E. Ordentlich and T. Weissman. Asymptotic filtering and entropy rate of a hidden Markov process in the rare transitions regime. International Symposium on Information Theory, pp. 1838-1842, Adelaide, Australia, 2005.

[22] E. Ordentlich and T. Weissman. On the optimality of symbol by symbol filtering and denoising. *Information Theory, IEEE Transactions*, Volume 52, Issue 1, Jan. 2006 Page(s):19 - 40.

[23] E. Ordentlich and T. Weissman. New bounds on the entropy rate of hidden Markov process. *IEEE Information Theory Workshop*, San Antonio, Texas, 24-29 Oct. 2004, Page(s):117 - 122.

[24] Y. Peres and A. Quas. Entropy Rate for Hidden Markov Chains with Rare Transitions. To appear in *Entropy of Hidden Markov Chains and Connections to Dynamical Systems*.

[25] H. Pfister, J. Soriaga and P. Siegel. The achievable information rates of finite-state ISI channels. *Proc. IEEE GLOBECOM*, San Antonio, TX, pp. 2992–2996, Nov. 2001.

[26] H. Pfister. On the Capacity of Finite State Channels and the Analysis of Convolutional Accumulate-m Codes. PhD thesis, University of California, San Diego, La Jolla, CA, USA, March 2003.

[27] H. Pfister. The Capacity of Finite-State Channels in the High-Noise Regime. To appear in *Entropy of Hidden Markov Processes and Connections to Dynamical Systems*.

[28] V. Sharma and S. Singh. Entropy and channel capacity in the regenerative setup with applications to Markov channels. *IEEE International Symposium on Information Theory*, Washington, D.C., Page 283, June 24-29 2001.

[29] W. Stout. Almost sure convergence. New York, Academic Press, 1974.

[30] P. Vontobel, A. Kavcic, D. Arnold and Hans-Andrea Loeliger. A Generalization of the Blahut-Arimoto Algorithm to Finite-State Channels. *IEEE Trans. Inform. Theory*, vol. 54, no. 5, pp. 1887–1918, May, 2008.

[31] O. Zuk, I. Kanter and E. Domany. The entropy of a binary hidden Markov process. *J. Stat. Phys.*, 121(3-4): 343-360, 2005.

[32] O. Zuk, E. Domany, I. Kanter, and M. Aizenman. Taylor series expansions for the entropy rate of hidden Markov Processes. ICC 2006, Istanbul.