Limit Theorems for the Sample Entropy of Hidden Markov Chains

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

February 16, 2011

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 almost surely. In this paper, we focus our attention on the convergence behavior of the sample entropy of a hidden Markov chain. Under certain positivity assumption, we prove that a central limit theorem (CLT) with some Berry-Esseen bound (such bound characterizes rate of convergence of CLT) 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.

1 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. Such theorems do not appear to 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 process 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 function defined on $\mathcal{Y}$ with values from a finite alphabet $\mathcal{Z} = \{1, 2, \cdots, A\}$. 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 [9] showed that $H(Z)$ can be written as an integral of an explicit function on a simplex with respect to Blackwell's Measure Q. However, the measure Q seems to be rather 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 [32], a variation on the classical Birch bounds [8] can be found in [13] and a new numerical approximation of $H(Z)$ has been proposed in [17]. Generalizing Blackwell's idea, an integral formula for the derivatives of $H(Z)$ has been derived in [37]. In another direction, [4, 26, 32, 42, 43, 31, 18, 20, 21, 22, 23, 24, 1, 37] 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 Mente 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 [35], Sharma and Singh [40]. 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 [36] 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 [25]. In essence, both of the two CLTs are proved using effective Martingale approximations of the sample entropy (see [19] for this standard technique).

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 certain positivity assumptions. In Section 3, we establish a CLT with some Berry-Esseen bound (such bound [5, 15] characterizes rate of convergence of CLT) for the sample entropy, and we use this CLT can to establish a LIL in Section 4.

Formally, for $i = 1, 2, \cdots$, define X_i as the "centered" version of $-\log P(Z_1|Z_1^{i-1})$, that is,

$$X_i = -\log P(Z_i|Z_1^{i-1}) - E[-\log P(Z_i|Z_1^{i-1})] = -\log P(Z_i|Z_1^{i-1}) - H(Z_i|Z_1^{i-1}).$$

And define

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

obviously S_n is the "centered" version of $-\log P(Z_1^n)$, and $S_n = -\log P(Z_1^n) - H(Z_1^n)$. 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 2.6 and Remark 2.7).

We will prove the following central limit theorem with a Berry-Esseen bound (such bound [5, 15] characterizes rate of convergence of CLT).

Theorem 1.1. *Under Conditions (I) and (II), for any $\varepsilon > 0$, there exists $C > 0$ such that for any n*

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

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

Theorem 1.2. *Under Conditions (I) and (II), we have*

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

2 Key Lemmas

This section includes several key lemmas, among which Lemmas 2.1, 2.2, 2.4 require Condition (I) only.

With the fact that a $n\times n$ positive matrix induces a contraction mapping on the interior of the $(n-1)$-dimensional real simplex under the Hilbert metric [39], the following well-known lemma can be established (see, e.g., [18] 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 $m, \hat{m} \geq n \geq 0$), we have*

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

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

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

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

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

T is said to be a *ψ-mixing sequence* if $\psi(n) \to 0$ as $n \to \infty$. It is well known [10] 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 [10]; for a comprehensive exposition to the vast literature on this subject, we refer to [11].

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

Proof. For each $z \in \mathcal{Z}$, let Δ_z denote the $B \times B$ matrix such that $\Delta_z(i,j) = \Delta(i,j)$ for j with $\Phi(j) = z$, and $\Delta_z(i,j) = 0$ otherwise. Obviously $\sum_{z \in \mathcal{Z}} \Delta_z = \Delta$. One also observes that for any $z_{m_1}^{m_2}$,

$$p(z_{m_1}^{m_2}) = \pi \Delta_{z_{m_1}^{m_2}} \mathbf{1},$$

where π is the stationary vector of Y, $\mathbf{1}$ denotes the all one column vector and $\Delta_{z_{m_1}^{m_2}} = \Delta_{z_{m_1}} \Delta_{z_{m_1+1}} \cdots \Delta_{z_{m_2}}$. It then follows that for any positive n, m, l and any z_0^m, z_{-n-l}^{-n},

$$p(z_0^m | z_{-n-l}^{-n}) = \sum_{z_{-n+1}^{-1}} \frac{\pi \Delta_{z_{-n-l}^m} \mathbf{1}}{\pi \Delta_{z_{-n-l}^{-n}} \mathbf{1}} = \frac{\pi \Delta_{z_{-n-l}^{-n}}}{\pi \Delta_{z_{-n-l}^{-n}} \mathbf{1}} (\sum_{z \in \mathcal{Z}} \Delta_z)^{n-1} \Delta_{z_0^m} \mathbf{1} = \frac{\pi \Delta_{z_{-n-l}^{-n}}}{\pi \Delta_{z_{-n-l}^{-n}} \mathbf{1}} \Delta^{n-1} \Delta_{z_0^m} \mathbf{1}.$$

Let λ_2 denote the second largest (in modulus) eigenvalue of Δ. By the Perron-Frobenius theory (see, e.g., [39]), $|\Delta_2| < 1$; furthermore, for any λ with $|\lambda_2| < \lambda < 1$, there exists $C_1 > 0$ such that for any probability vector x, we have

$$|x\Delta^n - \pi| \leq C_1 \lambda^n.$$

It then follows that

$$p(z_0^m | z_{-n-l}^{-n}) = \pi \Delta_{z_0^m} \mathbf{1} + O(\lambda^n) \Delta_{z_0^m} \mathbf{1} = p(z_0^m) + O(\lambda^n) p(z_0^m).$$

Noting that the constant in $O(\lambda^n)$ is independent of n, m, l and z_0^m, z_{-n-l}^{-n}, we then conclude that for any $U \in \mathcal{B}(Z_{-\infty}^{-n}), V \in \mathcal{B}(Z_0^{\infty})$,

$$P(V|U) = P(V) + O(\lambda^n) P(V),$$

which immediately implies the lemma. $\square$

Remark 2.3. Note that Lemma 2.2 still holds with the same proof if Condition (I) is replaced by "Δ is an irreducible and aperiodic matrix". This fact, however, will not be used in this paper.

In the following, we rewrite $-\log p(z_j|z_i^{j-1}) - H(Z_j|Z_i^{j-1})$ as $z_i^j)$ for notational simplicity.

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

Lemma 2.4. *1. There exist $C > 0$ and $0 < \rho < 1$ such that for all $i, j \geq 0$,*

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

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

$$|E[X_i X_j]| \leq C\theta^{j-i}.$$

Proof. 1. Simple computations lead to

$$\begin{aligned}
E[X_{i+1}X_{i+1+j}] - E[X_i X_{i+j}] &= \sum_{z_1^{i+1+j}} p(z_1^{i+1+j}) f(z_1^{i+1+j}) f(z_1^{i+1}) - \sum_{z_1^{i+j}} p(z_1^{i+j}) f(z_1^{i+j}) f(z_1^i) \\
&= \sum_{z_{-i-j}^0} p(z_{-i-j}^0) f(z_{-i-j}^0) f(z_{-i-j}^{-j}) - \sum_{z_{-i-j+1}^0} p(z_{-i-j+1}^0) f(z_{-i-j+1}^0) f(z_{-i-j+1}^{-j}) \\
&= \sum_{z_{-i-j}^0} p(z_{-i-j}^0)(f(z_{-i-j}^0) f(z_{-i-j}^{-j}) - f(z_{-i-j+1}^0) f(z_{-i-j+1}^{-j})) \\
&= \sum_{z_{-i-j}^0} p(z_{-i-j}^0) f(z_{-i-j}^0)(f(z_{-i-j}^{-j}) - f(z_{-i-j+1}^{-j})) \\
&\quad + \sum_{z_{-i-j}^0} p(z_{-i-j}^0)(f(z_{-i-j}^0) - f(z_{-i-j+1}^0)) f(z_{-i-j+1}^{-j}). \qquad (1)
\end{aligned}$$

Since Δ is a strictly positive matrix, $\log p(z_0|z_{-i}^{-1})$ and $H(Z_0|Z_{-i}^0)$ are all bounded from above and below uniformly in i. It then follows from this fact and Lemma 2.1 that there exist $C > 0$ and $0 < \rho < 1$ such that

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

Part 1 of the lemma then immediately follows.

2. Let $l = \lfloor i + j \rfloor / 2$. By Lemma 2.1 and Lemma 2.2, there exist $0 < \rho, \lambda < 1$ such that

$$\begin{aligned}
E[X_i X_j] &= \sum_{z_1^j} p(z_1^j) f(z_1^i) f(z_1^j) \\
&= \sum_{z_1^j} p(z_1^j) f(z_1^i)(f(z_l^j) + O(\rho^{j-l})) \\
&= \sum_{z_1^i, z_l^j} p(z_1^i) f(z_1^i) p(z_l^j|z_1^i) f(z_l^j) + O(\rho^{j-l}) \\
&= \sum_{z_1^i, z_l^j} p(z_1^i) f(z_1^i)(p(z_l^j) + O(\lambda^{l-i}) p(z_l^j)) f(z_l^j) + O(\rho^{j-l}) \\
&= \sum_{z_1^i, z_l^j} p(z_1^i) f(z_1^i) p(z_l^j) f(z_l^j) + \sum_{z_1^i, z_l^j} p(z_1^i) f(z_1^i) O(\lambda^{l-i}) p(z_l^j) f(z_l^j) + O(\rho^{j-l}) \\
&= 0 + O(\lambda^{l-i}) + O(\rho^{j-l}).
\end{aligned}$$

Notice that the constants in $O(\lambda^{l-i})$, $O(\rho^{j-l})$ above does not depend on z_1^j. Part 2 then immediately follows . $\square$

Remark 2.5. By Part 1 of Lemma 2.4, 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 2.6. *For any $0 < \alpha < 1$, there exists $C > 0$ such that for any m and n,*

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

here, recall that, as defined in Remark 2.5, $a_j = \lim_{i\to\infty} E[X_iX_{i+j}]$.

Proof. Letting $\beta = n^{-\alpha}$ for a fixed $0 < \alpha < 1$, we then have

$$\frac{E[(S_{n+m} - S_m)^2]}{n} = \frac{E[(\sum_{i=m+1}^{n+m} X_i)^2]}{n} = \frac{\sum_{m+1\leq i,\ i+j\leq n+m}(\sum_{j=0} +2\sum_{0<j\leq\beta n} +2\sum_{j>\beta n})E[X_iX_{i+j}]}{n}.$$

By Part 1 of Lemma 2.4 and Remark 2.5, for any $j > 0$, $E[X_iX_{i+j}] - a_j = O(\rho^i)$ for some $0 < \rho < 1$. It then follows that for $0 \leq j \leq \beta n$,

$$\sum_{m+1\leq i,\ i+j\leq n+m} E[X_iX_{i+j}] = (n-j)a_j + O(1);$$

here the constant in $O(1)$ does not depend on j. Also, by Part 2 of Lemma 2.4 and Remark 2.5, there exists $0 < \theta < 1$ such that for all $j > \beta n$, $E[X_iX_{i+j}] = O(\theta^{\beta n})$, and thus $a_j = O(\theta^{\beta n})$. Continuing the computation, we have

$$\frac{E[(S_{n+m} - S_m)^2]}{n} = \frac{(na_0 + O(1)) + (2(n-1)a_1 + O(1)) + \cdots + (2(n-\beta n)a_{\beta n} + O(1))}{n} + \frac{O(n^2\theta^{\beta n})}{n}$$

$$= a_0 + 2a_1 + \cdots + 2a_{\beta n} - 2\frac{a_1 + 2a_2 + \cdots + \beta n a_{\beta n}}{n} + \beta O(1) + O(n\theta^{\beta n}).$$

The lemma then immediately follows. $\square$

Remark 2.7. Choosing m in Lemma 2.6 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 2.8. *There exists $C > 0$ such that for all m and n*

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

Proof. By Lemmas 2.1, 2.6 and the stationarity of Z, we observe that for any m,

$$E[|S_{n+m} - S_m|^3] = E[|\sum_{i=m+1}^{n+m} f(Z_1^i)|^3] = E[|\sum_{i=m+1}^{n+m} (f(Z_{m+1}^i) + O(\rho^{i-m-1}))|^3]$$

$$= E[|S_n|^3] + O(E[|S_n|^2]) + O(E[|S_n|]) + O(1) = E[|S_n|^3] + O(n), \tag{2}$$

where the constant in $O(n)$ does not depend on m. So, to prove the lemma, it suffices to prove that there exists $C > 0$ such that for all n,

$$E[|S_n|^3] \leq Cn^{3/2}. \tag{3}$$

First, we observe that

$$E[|S_{2n}|^3] = E[|S_n+(S_{2n}-S_n)|^3] \leq E[|S_n|^3]+E[|S_{2n}-S_n|^3]+3E[|S_n|^2|S_{2n}-S_n|]+3E[|S_n||S_{2n}-S_n|^2].$$

It immediatedly follows from (2) that

$$E[|S_{2n} - S_n|^3] = E[|S_n|^3] + O(n).$$

And by Lemmas 2.2, 2.1 and the Holder inequality, there exist $0 < \rho < 1$ such that

$$\begin{aligned}
E[|S_n|^2|S_{2n} - S_n|] &= E[|S_n|^2|\sum_{i=n+1}^{2n} f(Z_1^i)|] \\
&= E[|S_n|^2|\sum_{i=n+1}^{2n} (f(Z_{n+1}^i) + O(\rho^{i-n-1}))|] \\
&= E[|S_n|^2|\sum_{i=n+1}^{2n} f(Z_{n+1}^i)|] + O(1)E[|S_n|^2] \\
&= \sum_{z_1^n, z_{n+1}^{2n}} (p(z_1^n)p(z_{n+1}^{2n}|z_1^n)|S_n|^2|\sum_{i=n+1}^{2n} f(Z_{n+1}^i)|) + O(1)E[|S_n|^2] \\
&= O(1)\sum_{z_1^n, z_{n+1}^{2n}} (p(z_1^n)p(z_{n+1}^{2n})|S_n|^2|\sum_{i=n+1}^{2n} f(Z_{n+1}^i)|) + O(1)E[|S_n|^2] \\
&= O(1)E[|S_n|^2]E[|\sum_{i=n+1}^{2n} f(Z_{n+1}^i)|] + O(1)E[|S_n|^2] \\
&= O(1)E[|S_n|^2]E[|S_{2n} - S_n|] + O(1)E[|S_n|^2] \\
&= O(1)E[|S_n|^2]E[|S_{2n} - S_n|^2]^{1/2} + O(1)E[|S_n|^2].
\end{aligned}$$

We then deduce that

$$E[|S_n|^2|S_{2n} - S_n|] = O(n^{3/2}).$$

With a similar argument, we also deduce that

$$E[|S_n||S_{2n} - S_n|^2] = O(n^{3/2}).$$

It then follows that there exists $C_1 > 0$ such that for all n

$$E[|S_{2n}|^3] \leq 2E[|S_n|^3] + C_1 n^{3/2}. \tag{4}$$

An iterative application of (4) with $n = 2^{r-1}, 2^{r-2}, \cdots, 2^0$ gives us

$$E[|S_{2^r}|^3] \le 2E[|S_{2^{r-1}}|^3] + C_1 2^{3(r-1)/2} \le 2^2 E[|S_{2^{r-2}}|^3] + C_1(2 \cdot 2^{3(r-2)/2} + 2^{3(r-1)/2})$$

$$\le \cdots \le 2^r E[|S_1|^3] + C_1(2^{r-1} + 2^{r-2}2^{3/2} + \cdots + 2 \cdot 2^{3(r-2)/2} + 2^{3(r-1)/2}) \le 2^r E[|S_1|^3] + C_1 \frac{2^{3(r-1)/2}}{1 - 2 \cdot 2^{-3/2}}.$$

It then follows that there exists $C_2 > 0$ such that for all r

$$E[|S_{2^r}|^3] \le C_2 (2^r)^{3/2}.$$

and by (2), choosing C_2 to be larger if necessary, we have for all r and all m

$$E[|S_{m+2^r} - S_m|^3] \le C_2 (2^r)^{3/2}. \tag{5}$$

Now, consider the general case when $2^r \le n < 2^{r+1}$. Expand

$$n = v_0 2^r + v_1 2^{r-1} + \cdots + v_r$$

where $v_0 = 1$, and $v_j = 0$ or 1 for $j = 1, 2, \cdots, r$. And define

$$w_i = v_{r-i} 2^i + v_{r-i+1} 2^{i-1} + \cdots + v_r.$$

Applying the Minkowski inequality and (5), we deduce that

$$E[|S_n|^3] \le E[|\sum_{i=0}^{r} (S_{w_i} - S_{w_{i-1}})|^3] \le (\sum_{i=0}^{r} E^{1/3}[|S_{w_i} - S_{w_{i-1}}|^3])^3$$

$$\le (\sum_{i=0}^{r} (C_2 2^{i/2}))^3 \le C_2^3 (2^{r/2} \frac{1}{1 - 1/2})^3 = 8C_2^3 2^{3r/2} \le 8C_2^3 n^{3/2}.$$

Inequality (3) is then immediately established if we choose $C = 8C_2^3$, □

3 Central Limit Theorem

Recall that

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

and

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

Fix $\varepsilon_0 > 0$ (we will choose ε_0 to be small later), 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" [6] 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 2 established, the remainder of the proof of Theorem 1.1 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 the standard Esseen's Lemma, $P(S_n'/\sigma_n < x)$ 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 1.1, 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 $|t| \le n^{1/11}$,*

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

Proof. Let $l = (k-1)p + (k-2)q + q/2$. By Lemma 2.1 and Lemma 2.2, there exist

$0 < \lambda, \rho < 1$ such that

$$\begin{aligned}
E[\exp(it\sum_{j=1}^{k}\zeta_j/\sigma_n)] &= E[\exp(it\sum_{j=1}^{k-1}\zeta_j/\sigma_n)\exp(it\zeta_k/\sigma_n)] \\
&= E[\exp(it\sum_{j=1}^{k-1}\zeta_j/\sigma_n)\exp(it\sum_{i=(k-1)(p+q)+1}^{kp+(k-1)q} f(z_1^i)/\sigma_n)] \\
&= E[\exp(it\sum_{j=1}^{k-1}\zeta_j/\sigma_n)\exp(it\sum_{i=(k-1)(p+q)+1}^{kp+(k-1)q} f(z_l^i)/\sigma_n)] + O(n^{1/11}\rho^{q(n)/2}/\sigma_n) \\
&= E[\exp(it\sum_{j=1}^{k-1}\zeta_j/\sigma_n)]E[\exp(it\sum_{i=(k-1)(p+q)+1}^{kp+(k-1)q} f(z_l^i)/\sigma_n)] + O(\lambda^{q(n)/2}) + O(n^{1/11}\rho^{q(n)/2}/\sigma_n) \\
&= E[\exp(it\sum_{j=1}^{k-1}\zeta_j/\sigma_n)]E[\exp(it\zeta_k/\sigma_n)] + O(\lambda^{q(n)/2}) + O(n^{1/11}\rho^{q(n)/2}/\sigma_n),
\end{aligned}$$

here, again, $-\log p(z_j|z_i^{j-1}) - H(Z_j|Z_i^{j-1})$ is rewritten as $f(z_i^j)$. Noticing that $|E[\exp(it\zeta_j/\sigma_n)]| \leq 1$ and applying an inductive argument, we conclude that

$$E[\exp(itS_n'/\sigma_n)] = E[\exp(it\sum_{j=1}^{k}\zeta_j/\sigma_n)] = \prod_{j=1}^{k} E[\exp(it\zeta_j/\sigma_n)]| + O(\lambda^{q(n)/2}) + O(n^{1/11}\rho^{q(n)/2}/\sigma_n),$$

which immediately implies the lemma. □

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

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

Proof. It is well known (see, e.g., page 343 of [7]) that for any random variable X and any $t \in \mathbb{R}$, we have

$$|E[\exp(itX)] - \sum_{k=0}^{n}(it)^k/k!E[X^k]| \leq E[|tX|^{n+1}/(n+1)!]. \tag{6}$$

Replacing X by ζ_j/σ_n, we deduce that

$$E[\exp(it\zeta_j/\sigma_n)] = 1 - E[\zeta_j^2]t^2/(2\sigma_n^2) + O(E[|\zeta_j|^3]t^3/(6\sigma_n^3)).$$

With Lemmas 2.6 and 2.8, one checks that for any $|t| \leq n^{1/11}$,

$$E[\zeta_j^2]t^2/(2\sigma_n^2), E[|\zeta_j|^3]t^3/(6\sigma_n^3) \to 0 \text{ as } n \to \infty.$$

Using the fact that $\log(1-x) = -x + O(x^2)$ for $|x| < 1$, we deduce that

$$\log E[\exp(it\zeta_j/\sigma_n)] = -E[\zeta_j^2]t^2/(2\sigma_n^2) + O(E[|\zeta_j|^3]t^3/\sigma_n^3) + O(E^2[\zeta_j^2]t^4/\sigma_n^4)$$

$$+O(E^2[|\zeta_j|^3]t^6/\sigma_n^6) + O(E[\zeta_j^2]E[|\zeta_j|^3]t^5/\sigma_n^5).$$

Now,

$$\log \prod_{j=1}^{k} E[\exp(it\zeta_j/\sigma_n)] = \sum_{j=1}^{k} \log E[\exp(it\zeta_j/\sigma_n)]$$

$$= -t^2/2 - t^2(\sum_{j=1}^{k} E[\zeta_j^2]/\sigma_n^2 - 1)/2 + O(\sum_{j=1}^{k} E[|\zeta_j|^3]t^3/\sigma_n^3) + O(\sum_{j=1}^{k} E^2[\zeta_j^2]t^4/\sigma_n^4)$$

$$+ O(\sum_{j=1}^{k} E^2[|\zeta_j|^3]t^6/\sigma_n^6) + O(\sum_{j=1}^{k} E[\zeta_j^2]E[|\zeta_j|^3]t^5/\sigma_n^5). \tag{7}$$

It follows from Lemma 2.6 that for any $\alpha > 0$

$$E[\zeta_i^2] = p(n)(\sigma^2 + O(p(n)^{-\alpha})), \qquad i = 1, 2, \cdots, k,$$

and

$$\sigma_n^2 = n(\sigma^2 + O(n^{-\alpha})).$$

Now, choosing α sufficiently close to 1, we then have that for $|t| \leq n^{1/11}$

$$t^2(\frac{\sum_{j=1}^{k} E[\zeta_j^2]}{\sigma_n^2} - 1) = t^2(\frac{k(n)p(n)(\sigma^2 + O(p(n)^{-\alpha}))}{n(\sigma^2 + O(n^{-\alpha}))} - 1)$$

$$= t^2(\frac{O(n^{-\alpha(3/11+\varepsilon_0)}) + \sigma^2 O(n^{-3/11}) - O(n^{-\alpha}) - O(n^{-(\alpha+3/11)})}{(\sigma^2 + O(n^{-\alpha}))(1 + O(n^{-3/11}))}) = O(n^{-1/11}),$$

where we used $k = O(n^{8/11-\varepsilon_0})$. One also easily checks that

$$\sum_{j=1}^{k} E[|\zeta_j|^3]t^3/\sigma_n^3 = O(n^{1/2(3/11+\varepsilon_0)+3/11-1/2}) = O(n^{-1/11+\varepsilon_0/2}), \tag{8}$$

and

$$\sum_{j=1}^{k} E^2[\zeta_j^2]t^4/\sigma_n^4, \sum_{j=1}^{k} E^2[|\zeta_j|^3]t^6/\sigma_n^6, \sum_{j=1}^{k} E[\zeta_j^2]E[|\zeta_j|^3]t^5/\sigma_n^5 = O(n^{-1/11}). \tag{9}$$

So we deduce that

$$\log \prod_{j=1}^{k} E[\exp(it\zeta_j/\sigma_n)] + t^2/2 = O(n^{-1/11+\varepsilon_0/2})$$

uniformly for $|t| \leq n^{1/11}$. It then follows from $|e^x - 1| \leq |x|e^{|x|}$ for all x that

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

uniformly for $|t| \leq n^{1/11}$, which immediately implies the lemma. $\square$

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 [41] 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 $|g| \leq M$. Then*

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

for every $T > 0$.

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

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

Proof. Let $\phi_0(t)$ denote the characteristic function of a standard normal random variable, that is, $\phi_0(t) = \exp(-t^2/2)$. Let $\phi_1(t)$ denote the characteristic function of S'_n/σ_n, that is, $\phi_1(t) = E[\exp(itS'_n/\sigma_n)]$. Let $\phi_2(t)$ denote the function $\prod_{j=1}^{k} E[\exp(it\zeta_j/\sigma_n)]$. Applying Lemma 3.3 with $T = n^{1/11}$, we have

$$\begin{aligned}
\sup_x \left| P(S'_n/\sigma_n < x) \right. &- \left. \int_{-\infty}^{x} (2\pi)^{-1/2} \exp(-y^2/2) dy \right| = O\left(\int_{-n^{1/11}}^{n^{1/11}} \left| \frac{\phi_1(t) - \phi_0(t)}{t} \right| dt + n^{-1/11} \right) \\
&= O\left(\int_{-n^{1/11}}^{n^{1/11}} \left| \frac{\phi_1(t) - \phi_2(t)}{t} \right| dt + \int_{-n^{1/11}}^{n^{1/11}} \left| \frac{\phi_2(t) - \phi_0(t)}{t} \right| dt + n^{-1/11} \right) \\
&= O\left(\int_{|t|<n^{-1/2}} \left| \frac{\phi_1(t) - \phi_2(t)}{t} \right| dt + \int_{n^{-1/2}<|t|<n^{1/11}} \left| \frac{\phi_1(t) - \phi_2(t)}{t} \right| dt \right. \\
&\quad \left. + \int_{|t|<n^{-1/2}} \left| \frac{\phi_2(t) - \phi_0(t)}{t} \right| dt + \int_{n^{-1/2}<|t|<n^{1/11}} \left| \frac{\phi_2(t) - \phi_0(t)}{t} \right| dt + n^{-1/11} \right).
\end{aligned}$$

Note that by Lemmas 2.4, 2.6, we deduce that for any $\alpha > 0$ and some $0 < \theta < 1$,

$$\frac{E[S_n'^2]}{\sigma_n^2} = \frac{\sum_{i=1}^{k} E[\zeta_i^2] + 2\sum_{i<j} E[\zeta_i \zeta_j]}{\sigma_n^2} = \frac{k(n)p(n)(\sigma^2 + O(p(n)^{-\alpha})) + O(n^2 \theta^{q(n)})}{n(\sigma^2 + O(n^{-\alpha}))},$$

which implies that

$$\frac{E[S_n'^2]}{\sigma_n^2} = O(1), \tag{10}$$

uniformly over all n. It then follows that for all $|t| \leq 1$,

$$\phi_1(t) = E[\exp(itS'_n/\sigma_n)] = 1 + t^2 O(E[S_n'^2]/\sigma_n^2) = 1 + O(t^2), \tag{11}$$

where we applied (6). Also, it follows from (7),(10), (8) and (9) that for all $|t| \leq 1$,

$$\log \phi_2(t) = O(t^2),$$

uniformly over all n, and thus for all $|t| \leq 1$,

$$\phi_2(t) = 1 + O(t^2), \tag{12}$$

uniformly over all n. Obviously, we also have for all $|t| \leq 1$,

$$\phi_0(t) = 1 + O(t^2). \tag{13}$$

It then follows from (11), (12) and (13) that there exists $C_1 > 0$ such that for all $|t| \leq 1$ and all n,

$$|\phi_1(t) - \phi_2(t)| \leq C_1 t^2, \qquad |\phi_2(t) - \phi_0(t)| \leq C_1 t^2,$$

which implies that

$$\int_{|t|<n^{-1/2}} \left| \frac{\phi_1(t) - \phi_2(t)}{t} \right| dt = O(n^{-1}), \int_{|t|<n^{-1/2}} \left| \frac{\phi_2(t) - \phi_0(t)}{t} \right| dt = O(n^{-1}). \tag{14}$$

It follows from Lemma 3.2 that

$$\int_{n^{-1/2}<|t|<n^{1/11}} \left| \frac{\phi_2(t) - \phi_0(t)}{t} \right| dt = O(n^{-1/11+\varepsilon_0/2} \log n), \tag{15}$$

and from Lemma 3.1 that

$$\int_{n^{-1/2}<|t|<n^{1/11}} \left| \frac{\phi_1(t) - \phi_2(t)}{t} \right| dt = O(\rho_1^{q(n)} \log n) \tag{16}$$

for some $0 < \rho_1 < 1$. The lemma then follows from (14), (15), (16) and the fact that ε_0 can be chosen to be arbitrarily small. $\square$

We are now ready to prove Theorem 1.1.

Proof of Theorem 1.1. Let F denote the event "$|S_n''|/\sigma_n \leq n^{-1/11}$". Then

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

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

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

Applying Lemma 2.6 and Lemma 2.2, we deduce that for some $0 < \theta < 1$,

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

Also, by the Markov inequality, we have

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

The theorem then immediately follows. $\square$

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

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

4 Law of Iterated Logarithm

From the central limit theorem with a Berry-Esseen bound (Theorem 1.1), we only need to follow a standard "track" to establish the law of iterated logarithm. 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 [41]):

1. As an immediately corollary of Theorem 1.1, the following Lemma 4.1 gives bounds on the tail probability of S_n.
2. We then slightly modified Reznik's maximal inequality to to obtain our maximal inequality in Lemma 4.2;
3. Finally, we are ready for the proof of Theorem 1.2, where some necessary modifications are incorporated into the original Reznik's proof to deal with the complications resulted from the fact that X is not stationary.

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.

Proof. By Theorem 1.1, we have for any $\varepsilon > 0$

$$P(S_n/\sigma_n > (1+\delta)(2\log\log \sigma_n^2)^{1/2}) - \int_{(1+\delta)(2\log\log \sigma_n^2)^{1/2}}^{\infty} (2\pi)^{-1/2} \exp(-y^2/2) dy = O(n^{-1/11+\varepsilon}).$$

It can be verified that for any $x > 0$

$$\int_x^{\infty} (2\pi)^{-1/2} \exp(-y^2/2) dy \leq \exp(-x^2/2),$$

which implies that

$$\int_{(1+\delta)(2\log\log \sigma_n^2)^{1/2}}^{\infty} (2\pi)^{-1/2} \exp(-y^2/2) dy \leq (\log \sigma_n^2)^{-(1+\delta)^2}.$$

It can also be verified that for any $\alpha > 0$,

$$\exp(-(1+\alpha)x^2/2) \leq \int_x^{\infty} (2\pi)^{-1/2} \exp(-y^2/2) dy$$

for x large enough, which implies that for any $\alpha > 0$,

$$(\log \sigma_n^2)^{-(1+\delta)^2(1+\alpha)} \leq \int_{(1+\delta)(2\log\log \sigma_n^2)^{1/2}}^{\infty} (2\pi)^{-1/2} \exp(-y^2/2)dy,$$

for n large enough. The lemma then follows from $\sigma_n^2 = n\sigma^2(1+o(1))$, by Remark 2.7. $\square$

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

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

for sufficiently large n.

Proof. For $j = 1, 2, \cdots, n$, let F_j be the event "$S_1, S_2, \cdots, S_{j-1} \leq x, S_j > x$", that is, F_j is the event "j is the smallest index such that $S_j > x$". Then, for any $x > 0$, $0 < \alpha < 1/2$ and $C > 0$, we have

$$P(\max_{j\leq n} S_j > x) = \sum_{j=1}^{n} P(F_j) = (\sum_{P(F_j) > Cn^{-(1+\alpha)}} + \sum_{P(F_j)\leq Cn^{-(1+\alpha)}})P(F_j) \leq \sum_{P(F_j) > Cn^{-(1+\alpha)}} P(F_j) + Cn^{-\alpha}.$$

Since

$$P(S_n > x - 2\sigma_n) \geq \sum_{j=1}^{n} P(|S_n - S_j| \leq 2\sigma_n, F_j) = \sum_{j=1}^{n} P(|S_n - S_j| \leq 2\sigma_n | F_j)P(F_j),$$

we only need to prove that for n sufficiently large, $P(|S_n - S_j| > 2\sigma_n|F_j) \leq 1/2$ for any F_j with $P(F_j) > Cn^{-(1+\alpha)}$.

Now fix a positive integer n_0 (we will choose n_0 large enough later). For the case when $n - j \leq n_0$, applying the Markov inequality, we have

$$P(|S_n - S_j| > 2\sigma_n | F_j) \leq P(|S_n - S_j| > 2\sigma_n)/p(F_j) \leq E[|\sum_{i=j+1}^{n} X_i|^3](8P(F_j)\sigma_n^3)^{-1} = O(n^{\alpha-1/2}),$$

where we used $P(F_j) > Cn^{-(1+\alpha)}$. For the case when $n - j > n_0$, we have

$$P(|S_n - S_j| > 2\sigma_n | F_j) \leq P(|\sum_{i=j+1}^{j+n_0} X_i| > \sigma_n/2 | F_j) + P(|\sum_{i=j+1+n_0}^{n} X_i| > 3\sigma_n/2 | F_j).$$

Again, using $P(F_j) > Cn^{-(1+\alpha)}$ and the Markov inequality, we have

$$P(|\sum_{i=j+1}^{j+n_0} X_i| > \sigma_n/2 | F_j) \leq P(|\sum_{i=j+1}^{j+n_0} X_i| > \sigma_n/2)/P(F_j) \leq 8E[|\sum_{i=j+1}^{j+n_0} X_i|^3](P(F_j)\sigma_n^3)^{-1} = O(n^{\alpha-1/2}).$$

Also, it follows from Lemma 2.2 and Lemma 2.6 that

$$P(|\sum_{i=j+1+n_0}^{n} X_i| > 3\sigma_n/2 | F_j) \leq \psi(n_0) + P(|\sum_{i=j+1+n_0}^{n} X_i| > 3\sigma_n/2)$$

$$\leq \psi(n_0) + 4E[|\sum_{i=j+1+n_0}^{n} X_i|^2]/9\sigma_n^2 = \psi(n_0) + 4/9 + o(1).$$

Apparently, for both cases, choosing n_0 sufficiently large so that $\psi(n_0)$ is small enough, and then choosing n sufficiently large, we deduce that for any F_j with $P(F_j) > Cn^{-(1+\alpha)}$,

$$P(|S_n - S_j| > 2\sigma_n | F_j) \leq 1/2.$$

The lemma then immediately follows. $\square$

We are now ready to prove Theorem 1.2.

Proof of Theorem 1.2. We first show that

$$\limsup_{n\to\infty} \frac{S_n}{(2n\sigma^2 \log\log n\sigma^2)^{1/2}} \leq 1 \qquad a.s.; \tag{17}$$

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

$$P(\frac{S_n}{(2\sigma_n^2 \log\log \sigma_n^2)^{1/2}} > 1 + \delta \;\; i.o.) = 0, \tag{18}$$

here we remind the reader that by Remark 2.7, $\sigma_n^2 = n(\sigma^2 + o(1))$ and "*i.o.*" means "infinitely often".

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

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

So, to prove (18), it suffices (by the Borel-Cantelli Lemma) to show that

$$\sum_{j=1}^{\infty} P(\max_{n \leq n_{j+1}} \frac{S_n}{(2\sigma_{n_j}^2 \log\log \sigma_{n_j}^2)^{1/2}} > 1 + \delta) < \infty. \tag{19}$$

Now, by Lemma 4.2,

$$\sum_{j=1}^{\infty} P(\max_{n \leq n_{j+1}} \frac{S_n}{(2\sigma_{n_j}^2 \log\log \sigma_{n_j}^2)^{1/2}} > 1+\delta) \leq \sum_{j=1}^{\infty} P(S_{n_{j+1}} > (1+\delta)(2\sigma_{n_j}^2 \log\log \sigma_{n_j}^2)^{1/2} - 2\sigma_{n_{j+1}}). \tag{20}$$

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

$$(1 + \delta)(2\sigma_{n_j}^2 \log\log \sigma_{n_j}^2)^{1/2} - 2\sigma_{n_{j+1}} > (1 + \delta_1)(2\sigma_{n_j}^2 \log\log \sigma_{n_j}^2)^{1/2}. \tag{21}$$

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

$$\sum_{j=1}^{\infty} P(S_{n_{j+1}} > (1+\delta_1)(2\sigma_{n_j}^2 \log\log \sigma_{n_j}^2)^{1/2}) \leq \sum_{j=1}^{\infty} (\log \sigma_{n_j}^2)^{-(1+\delta_1)^2(1-\alpha)} = \sum_{j=1}^{\infty} O(j^{-(1+\delta_1)^2(1-\alpha)}) < \infty. \tag{22}$$

Immediately, (19) follows from (20), (21) and (22). Here, we remark that the same argument as above with X_i replaced by $-X_i$ leads to

$$\liminf_{n\to\infty} \frac{S_n}{(2n\sigma^2 \log\log n\sigma^2)^{1/2}} \geq -1 \qquad a.s. \tag{23}$$

For the other direction, we next show that

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

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

$$P(\frac{S_n}{(2\sigma_n^2 \log\log \sigma_n^2)^{1/2}} > 1 - \delta \;\; i.o.) = 1. \tag{24}$$

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 $\alpha > 0$

$$P(C_n(\delta)) \geq P(S_{N^n - N^{n-1} - N^{n/2}} > (1-\delta_2) g(N^n - N^{n-1} - N^{n/2})) \geq \log(N^n - N^{n-1} - N^{n/2})^{-(1-\delta_2)^2(1+\alpha)}/2 \tag{25}$$

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

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

which, together with (25), implies that for any $\delta > 0$

$$\sum_{n=1}^{\infty} P(C_n(\delta)) = \infty. \tag{26}$$

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 (26), 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,

$$P(\cap_{m=n}^{n+l} C_m^c(\delta)) \leq P(\cap_{m=n}^{n+l} \hat{C}_m^c(\delta'')) = \prod_{m=n}^{n+l} P(\hat{C}_m^c(\delta'')) + \sum_{m=n}^{n+l} O(\theta^{N^{m/4}})$$

$$= \prod_{m=n}^{n+l} (1 - P(\hat{C}_m(\delta''))) + \sum_{m=n}^{n+l} O(\theta^{N^{m/4}}) \leq \exp(-\sum_{m=n}^{n+l} P(\hat{C}_m(\delta''))) + \sum_{m=n}^{l} O(\theta^{N^{m/4}}).$$

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

$$P(C_n(\delta) \ \ i.o.) = 1. \tag{27}$$

Let B_n be the event "$S_{N^{n-1}+N^{n/2}} > -2g(N^{n-1} + N^{n/2})$". It then follows from (23) that

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

which, together with (27), implies that for any $\hat{\delta} > 0$

$$P(B_n \cap C_n(\hat{\delta}) \ \ i.o.) = 1. \tag{28}$$

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

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

It then follows from (28) that

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

which immediately implies (24). □

5 Alternatives for Condition (II)

This section only assume Condition (I) and gives alternatives for Condition (II) provided Condition (I) is satisfied.

Let $(\Omega, \mathcal{F}, P)$ be the probability space which Z is defined on, and let $H_0 = H(Z_k, k \in \mathbb{Z})$ be the subspace of $\mathcal{L}^2(\mathcal{F})$ spanned by the equivalence classes of the random variables Z_k, $k \in \mathbb{Z}$, with inner product defined as

$$< V, W >= E[VW],$$

for any $V, W \in H_0$.

Theorem 5.1. *If* $\liminf_{n\to\infty} E[S_n^2] < \infty$, *then there exist a sequence of random variables* $(V_i, i \in \mathbb{N})$ *as* $n \to \infty$ *such that* $X_i = V_i - V_{i+1}$ *with* $E[V_i^2] = O(1)$ *uniformly for all* i, *and thus* $\sup_n E[S_n^2] < \infty$.

Proof. Let Q be an infinite subset of $\mathbb{N}$ such that $\sup_{n\in Q} E[S_n^2] < \infty$. Applying Lemma 2.1, we deduce that there exists $C > 0$ such that for all $i \in \mathbb{N}$,

$$\sup_{n\in Q} E[(S_{n+i-1} - S_{i-1})^2] \leq C,$$

where S_0 is interpreted as 0. It follows from Banach-Alaoglu theorem (which implies that every bounded and closed set in a Hilbert space is weakly compact; see Section 3.15 of [38]) that for any $i \in \mathbb{N}$, there exists $V_i \in H_0$ with $E[V_i^2] \leq C$, and Q_i, an infinite subset of Q such that for all $W \in H_0$,

$$\lim_{n\to\infty, n\in Q_i} < W, S_{n+i-1} - S_{i-1} >=< W, V_i >;$$

here, without loss of generality, we can assume that $Q_{i+1} \subset Q_i$ for all i. Then one verifies that for any $W \in H_0$, we have that for any i,

$$< W, X_i - V_i + V_{i+1} >= \lim_{n\to\infty, n\in Q_{i+1}} < W, X_i - (S_{n+i-1} - S_{i-1}) + (S_{n+i} - S_i) >= \lim_{n\to\infty, n\in Q} < W, X_{n+i} >= 0,$$

where we have applied Lemma 2.4 for the last equality. Choosing $W = X_i - V_i + V_{i+1}$, we then obtain that

$$\|X_i - V_i + V_{i+1}\|_2 = 0,$$

which implies that

$$X_i = V_i - V_{i+1}, \text{ a.s.}$$

It then follows that

$$E[S_n^2] = E[(V_1 - V_{n+1})^2] = E[V_1^2] + E[V_{n+1}^2] - 2E[V_1 V_{n+1}],$$

which, together with $E[V_i^2] \leq C$, implies the theorem.

□

A sequence of positive numbers, $(h(i), i \in \mathbb{N})$, is said to be "slowly varying" if for every positive integer m,

$$\lim_{n\to\infty} h(mn)/h(n) = 1,$$

and it is said to be "slowly varying in the strong sense" if

$$\lim_{m\to\infty} \frac{\min_{m\leq n\leq 2m} h(n)}{\max_{m\leq n\leq 2m} h(n)} = 1.$$

Lemma 5.2. *If* $\lim_{n\to\infty} E[S_n^2] = \infty$, *then* $E[S_n^2] = nh(n)$, *where* $(h(i), i \in \mathbb{N})$ *is a sequence of slowly varying positive numbers.*

Proof. We only need to show that for every positive integer l,

$$\lim_{n\to\infty} \sigma_{ln}^2/\sigma_n^2 = l.$$

Following the proof of Theorem 2.1.2 in [30], let

$$\zeta_j = \sum_{s=1}^{n} X_{(j-1)n+(j-1)r+s}, \qquad j = 1, 2, \cdots, l,$$

$$\eta_j = \sum_{s=1}^{r} X_{jn+(j-1)r+s}, \qquad j = 1, 2, \cdots, l-1,$$

$$\eta_l = -\sum_{s=1}^{(l-1)r} X_{nl+s},$$

where $r = \lfloor \log \sigma_n^2 \rfloor$.

Now

$$\sigma_{ln}^2 = E[S_{ln}^2] = \sum_{j=1}^{l} E[\zeta_j^2] + 2\sum_{i\neq j} E[\zeta_i \zeta_j] + \sum_{i,j} E[\zeta_i \eta_j] + \sum_{i,j} E[\eta_i \eta_j].$$

It follows from Lemma 2.1 that for any j,

$$E[\zeta_j^2] = E[\zeta_1^2] + O(1)E[\zeta_1^2]^{1/2} = \sigma_n^2 + O(\sigma_n) \tag{29}$$

uniformly in j. Using an argument similar to the proof for Part 2 of Lemma 2.4, one has that there exists $0 < \theta < 1$ such that for $i \neq j$,

$$|E[\zeta_i \zeta_j]| = O(\theta^{\lfloor \log \sigma_n^2 \rfloor} \sigma_n^2),$$

where we also used (29). Using Schwartz inequality and (29), we also have

$$|E[\zeta_i \eta_j]| \leq E[\zeta_i^2]^{1/2} E[\eta_j^2]^{1/2} = O(\sigma_n \sigma_r) = O(\sigma_n \log \sigma_n),$$

and

$$|E[\eta_i \eta_j]| \leq O(\sigma_r^2) = O((\log \sigma_n)^2).$$

It then follows that for any positive integer l,

$$\sigma_{ln}^2 = l\sigma_n^2 + o(\sigma_n^2),$$

which immediately implies the lemma. □

Lemma 5.3. *If $\lim_{n\to\infty} E[S_n^2] = \infty$, then $E[S_n^2] = nh(n)$, where $(h(i), i \in \mathbb{N})$ is a sequence of slowly varying positive numbers in the strong sense.*

Proof. Note that by Lemma 2.1, we have that for any j,

$$\lim_{n\to\infty} \frac{E[(S_{n+j} - S_j)^2]}{E[S_n^2]} = 1, \tag{30}$$

uniformly in j. The lemma then follows from (30), Lemma 5.2 and an almost the same proof for Theorem 8.13 of [11]. □

The following lemma is well-known; see, e.g., Proposition 0.16 in [11].

Lemma 5.4. *Suppose* $(h(n), n \in \mathbb{N})$ *is a sequence of positive numbers which is slowly varying in the strong sense. Then for every* $\varepsilon > 0$*, one has that* $n^{\varepsilon} h(n) \to \infty$ *as* $n \to \infty$*.*

Lemma 5.5. *If* $\lim_{n\to\infty} E[S_n^2] = \infty$*, then* $\sigma > 0$*.*

Proof. Assume, for contradictions, that $\sigma = 0$. Since $\lim_{n\to\infty} E[S_n^2] = \infty$, we deduce, by Lemma 5.3, that $E[S_n^2]/n$ is slowly varying in the strong sense. Then, by Lemma 5.4, for any $\alpha > 0$, $n^{\alpha} E[S_n^2]/n \to \infty$ as $n \to \infty$. However, by Lemma 2.6, when $\sigma = 0$, $n^{\alpha} E[S_n^2]/n \to 0$ as $n \to \infty$ for any $0 < \alpha < 1$, which is a contradiction. □

The following theorem immediately follows from Theorem 5.1 and Lemma 5.5, which gives alternatives for Condition (II) given Condition (I) is satsified.

Theorem 5.6. *Under Condition (I), the following statements are equivalent*

1. $\sigma > 0$.
2. $\lim_{n\to\infty} E[S_n^2] = \infty$.
3. $\limsup_{n\to\infty} E[S_n^2] = \infty$.

References

[1] 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.*

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

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

[6] S. Bernstein. Sur l'extension du théorème limite du calcul des probabilités aux sommes de quantités dépendantes. *Math. Ann.* 97, 1-59.

[7] P. Billingsley. Probability and measure. New York, J. Wiley & Sons, c1995. 3rd ed.

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

[9] 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.

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

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

[12] H. Chernoff. A measure of asymptotic effciency for tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics.* 23:4, pp. 493-507, 1952.

[13] 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.

[14] Y. Ephraim and N. Merhav. Hidden Markov Processes. *IEEE Trans. Inform. Theory*, Vol. 48, No. 6, Pages: 1518-1569.

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

[16] 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.

[17] 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.

[18] 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.

[19] P. Hall and C. Heyde. *Margingale Limit Theory and its Application.* Academic Press, New York.

[20] 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.

[21] G. Han and B. Marcus. Asymptotics of Input-Constrained Binary Symmetric Channel Capacity. *Annals of Applied Probability*, Vol. 19, No. 3, June, 2009, Page(s):1063-1091.

[22] G. Han and B. Marcus and Y. Peres. A Complex Hilbert Metric and Applications to Domain of Analyticity for Entropy Rate of Hidden Markov Processes. To appear in *Entropy of Hidden Markov Processes and Connections to Dynamical Systems*.

[23] 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.

[24] 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.

[25] 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.

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

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

[28] P. Lezaud. Chernoff-type Bound for Finite Markov Chains. *The Annals of Applied Probability*. Vol. 8, No. 3, 849-867, 1998.

[29] P. Lezaud. Chernoff and Berry-Esseen Inequalities. *ESAIM: Probability and Statistics.* Vol. 5, 183-201, December 2001.

[30] C. Lin and C. Lu. Limit theory for mixing dependent random variables. New York: Science Press ; Dordrecht: Kluwer Academic Publishers, 1996.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.*

[38] W. Rudin. Functional Analysis (2nd ed.). Boston, MA: McGraw-Hill.

[39] E. Seneta. *Springer Series in Statistics.* Non-negative Matrices and Markov Chains. Springer-Verlag, New York Heidelberg Berlin, 1980.

[40] 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.

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

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

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