On the Gaussianity of Kolmogorov Complexity of Mixing Sequences

Morgane Austern, Arian Maleki

Abstract

Let $K(X_1, ..., X_n)$ and $H(X_n | X_{n-1}, ..., X_1)$ denote the Kolmogorov complexity and Shannon's entropy rate of a stationary and ergodic process $\{X_i\}_{i=-\infty}^{\infty}$. It has been proved that

 $\frac{K(X_1,\ldots,X_n)}{n}-H(X_n|X_{n-1},\ldots,X_1)\to 0,$

almost surely. This paper studies the convergence rate of this asymptotic result. In particular, we show that if the process satisfies certain mixing conditions, then there exists $\sigma < \infty$ such that

 $\sqrt{n}\left(\frac{K(X_{1:n})}{n}-H(X_0|X_1,\ldots,X_{-\infty})\right)\to_d N(0,\sigma^2).$

Furthermore, we show that under slightly stronger mixing conditions one may obtain non-asymptotic concentration bounds for the Kolmogorov complexity.

1 Introduction

1.1 Motivation and objective

Kolmogorov complexity of a binary sequence is defined as the length of the shortest program fed to a universal Turing machine that would print the sequence and halt. More formally, let U denote a Universal Turing machine. Given a program p the sequence printed by U is denoted with U(p).

Definition 1.1. Let \mathcal{P}_X denote the set of all binary programs that can generate a finite length binary sequence X and halt. Then, the Kolmogorov complexity of X is denoted with K(X) and is defined as

$$K(X) \triangleq \inf_{p \in \mathcal{P}_X} \operatorname{length}(p),$$

where length(p) denotes the length of the sequence. Furthermore, the Kolmogorov complexity of any finite-length finite-alphabet sequence is the Kolmogorov complexity of its binary representation.

Apart from its mathematical elegance, Kolmogorov complexity has exhibited promising theoretical results in other areas of research including inductive inference [1], denoising [2], linear regression [3], density estimation [4], etc. However, such theoretical results are overshadowed by the fact that Kolmogorov complexity is not computable (see Theorem 1.5 in [5]).

Both the usefulness of the Kolmogorov complexity and its incomputability has motivated researchers to find approximations of this quantity. One of the main approaches is to restrict the class of sequences to stationary and ergodic sequences, and use the properties of such sequences to find good approximations. The following theorem, due to Levin, clarifies why such assumptions might be useful.

Theorem 1. [5] Let $\{X_i\}_{i=-\infty}^{\infty}$ denote a binary stationary and ergodic process (under left shift), whose law is computable. Let its Shannon conditional entropy rate be $H(X_1|X_0,\ldots,X_{-\infty})$. Then.¹

$$\frac{K(X_{1:n})}{n} \stackrel{a.s.}{\to} H(X_1|X_0,\ldots,X_{-\infty}),$$

where $X_{1:n}$ denotes the vector (X_1, X_2, \ldots, X_n) .

According to this theorem Shannon's entropy can be seen as an approximation of the Kolmogorov's complexity of the process. This result is asymptotic and it is not clear how accurate this approximations is even if n is large. This paper establishes the accuracy of this approximation under certain mixing assumptions on the process, which will be clarified later.

1.2 Related work

Komogorov complexity evolved in the seminal papers of Solomonoff [6, 7], Kolmogorov [8], and Chaitin [9, 10, 11, 12]. Each author developed and used this quantity for different purposes. For instance, inspired by Shannon's theory of information, Kolmogorov developed his notion of complexity to quantify the amount of information that is present in a sequence of bits. Kolmogorov also conjectured that binary sequences that have maximal complexity, e.g. $K(X_1, X_2, \ldots, X_n) \ge n - c$, for some fixed c, are random in an intuitive sense. This conjecture was later established by Martin-Lof. Intuitively speaking he proved that if a sequence satisfies $K(X_1, X_2, \ldots, X_n) \ge n - c$, then any test that can be implemented by a turing machine will accept the randomness of this sequence (it should possibly use a different significance level). The intellectual value of this test of randomness was overshadowed by the incomputability of Kolmogorov complexity.

Many researchers explored new ways to improve the applicability of Kolmogorov complexity. For instance, [13, 12, 14, 15] explored computable approximations of Kolmogorov complexity. Another popular direction of research pursued connections between Kolmogorov's complexity and Shannon entropy [8, 5, 16, 17]. Levin's result, i.e. Theorem 1, is one of the most general connections between Kolmogorov complexity and Shannon entropy. In this paper, we push these connections one step further by providing convergence rate and concentration results for the Kolmogorov complexity.

2 Main result

According to Theorem 1, for every stationary and ergodic sequence, $\{X_i\}_{i=-\infty}^{\infty}$, on probability space $\{\Omega, \mathbb{P}, \mathcal{A}\}$, we have

$$\frac{K(X_{1:n})}{n} \stackrel{a.s.}{\to} H(X_1|X_0,\ldots,X_{-\infty}).$$

As we discussed before, in this paper we would like to characterize the rate of convergence for this asymptotic result. As our first goal, we would like to show that under some general conditions the convergence rate is $\frac{1}{\sqrt{n}}$. More specifically, we would like to show that

$$\sqrt{n}\left(\frac{K(X_{1:n})}{n} - H(X_1|X_0,\dots,X_{-\infty})\right) \tag{1}$$

¹This is Theorem 5.1 of [5]. As mentioned by Levin even ergodicity is not necessary, but then we should be careful in defining the entropy. For more information refer to [5].

converges in distribution to a non-degenerate random variable. Before we discuss our main theorem we would like to show that if we do not impose extra conditions on the process, the convergence rate could be slower than $\frac{1}{\sqrt{n}}$.

Example 2.1. Let ν_1 denote a probability measure on \mathbb{N}^* (the set of positive natural numbers) with probability mass function $\nu_1(t) = \frac{Z}{t^{3/2-\epsilon}}$, where $\epsilon \in (0, \frac{1}{6})$ and Z is the normalizing constant. Let $\{\tau_i\}_{i=1}^{\infty}$ denote iid samples from this distribution. Furthermore, let $\{Y_i\}_{i=1}^{\infty}$ and $\{\tilde{Y}_i\}_{i=1}^{\infty}$ denote two independent sequences of iid $\mathrm{Bern}(\frac{1}{2})$ random variables (independent from $\{\tau_i\}_{i=1}^{\infty}$). Given a natural number a let $\mathbf{0}_a$ denote a vector of size a with all elements being zero. We construct a binary sequence $\{\tilde{X}_i\}_{i=1}^{\infty}$ in the following way:

- (i) Pick τ_1 from our first sequence and set $\tilde{X}_{1:\tau_1} = Y_1 \mathbf{0}_{\tau_1} + (1 Y_1)\tilde{Y}_{1:\tau_1}$.
- (ii) To construct the i^{th} block we repeat what we did above. More specifically we draw τ_i and we set $\tilde{X}_{\tau_1+\ldots+\tau_{i-1}:\tau_1+\ldots+\tau_{i-1}+\tau_i} = Y_i\mathbf{0}_{\tau_i} + (1-Y_i)\tilde{Y}_{\tau_1+\ldots+\tau_{i-1}:\tau_1+\ldots+\tau_{i-1}+\tau_i}$.

To make $\tilde{X} = \{\tilde{X}\}_{i=1}^{\infty}$ stationary, draw $\theta | \tau_1 \sim \text{unif}(0, \tau_1 - 1)$ (uniform on the integers from 0 to $\tau_1 - 1$). Then, we generate the new process as $X = \Theta^{\theta} \tilde{X}$, where Θ is the left shift operator. It is straightforward to see that the process is stationary and ergodic. Hence, $\frac{K(X_{1:n})}{n} \stackrel{a.s.}{\to} H(X_1|X_0,\ldots,X_{-\infty})$. However, the convergence rate is slower than $1/\sqrt{n}$. The proof of our claim can be found in Section 3.6.1.

Note that a major issue in the above example is the fact that the elements of the sequence that are far apart can still have strong dependencies. Hence, intuitively speaking we expect that if the dependency of the process is weaker, then we may be able to obtain the $1/\sqrt{n}$ convergence rate. Mixing conditions are defined to capture the dependancies of stochastic process. We start with mixing conditions that will be used in our paper. Let $\{X_i\}_{i=-\infty}^{\infty}$ denote a stationary and ergodic process, and let $\mathbb{F}_{-\infty}^n = \sigma(X_i, i \leq n)$ denote the σ -filed of events generated by random variables $\ldots, X_{n-2}, X_{n-1}, X_n$ and $\mathbb{F}_n^{\infty} = \sigma(X_i, i \geq n)$ denote the σ -filed of events generated by X_n, X_{n+1}, \ldots

Definition 2.1. α -mixing coefficients of the process $\{X_i\}_{i=-\infty}^{\infty}$ is define as

$$\alpha(n) \triangleq \sup_{\substack{j \\ B \in \mathbb{F}_{-\infty}^{j} \\ B \in \mathbb{F}_{j+n}^{\infty}}} \left| P\left(A \cap B\right) - P(A)P(B) \right|.$$

A process X is α -mixing if $\alpha(n) \to 0$.

 α -mixing condition ensures that the parts of the process that are far apart are almost independent. Hence, we hope that if $\alpha(n)$ decays fast enough, then it will avoid the dependency issue raised in Example 2.1. In some of our results we will need a slightly stronger notion of mixing that we define below.

Definition 2.2. The ϕ -mixing coefficient of the process $\{X_i\}_{i=-\infty}^{\infty}$ is defined as

$$\phi(n) \triangleq \sup_{\substack{j \\ B \in \mathbb{F}_{j+n}^{o}}} \left| P\left(B|A\right) - P(B) \right|,$$

Furthermore, a process is called ϕ -mixing if $\phi(n) \to 0$.

Remark. It is straightforward to see that $\forall n, \phi(n) \geq \alpha(n)$. Hence, if a process is ϕ -mixing, then it will be α -mixing.

In addition to mixing, our proof requires another condition that is described below:

Definition 2.3. Let $\{X_i\}_{-\infty}^{\infty}$ denote a stationary and ergodic process. Consider $\delta > 0$ and define $\nu_{\delta}(n) \triangleq \mathbb{E}(|\log(P(X_0|X_{-1},\ldots,X_{-\infty})) - \log(P(X_0|X_{-1},\ldots,X_{-n}))|^{\frac{2+\delta}{1+\delta}}).$

Note that in this paper all the logarithms are in base 2. Since $\nu_{\delta}(n)$ is not a standard notion in probability theory, we explain some of its interesting features below:

- 1. The definition of $\nu_{\delta}(n)$ is close to the definition of the Kullback-Leiber divergence between $P(X_0|X_{-1},\ldots,X_{-\infty})$ and $P(X_0|X_{-1},\ldots,X_{-n})$. Hence, it measures the discrepancy of a process from a Markov process.
- 2. For a b-Markov source $\{X_i\}_{i=-\infty}^{\infty}$ we have $\mathcal{L}(X_1|X_{0:-b+1})=\mathcal{L}(X_1|X_{0:-\infty})^2$. Hence $\nu_{\delta}(n)=0$ 0 for every $n \ge b$.
- 3. Sequences generated by a hidden Markov model also have very fast decaying $\nu_{\delta}(n)$. The following lemma justifies our claim:

Lemma 2. Consider a hidden Markov model with $q: \mathcal{X} \times \mathcal{X} \to (0, \infty)$ denoting the transition kernel of the underlying Markov process and $g(\cdot \mid x)$ denoting the distribution of the observed variables for a given value x of the hidden variable. Also, suppose that the process satisfies the following conditions:

(i)
$$\epsilon \triangleq \frac{\text{essinf } q(x,x')}{\text{esssup } q(x,x')} \in (0,1).$$

(i)
$$\epsilon \triangleq \frac{\text{essinf } q(x,x')}{\text{esssup } q(x,x')} \in (0,1).$$

(ii) $1 < \eta \triangleq \sup_{y} \frac{\text{esssup}_{x} g(y|x)}{\text{essinf}_{x} g(y|x)} < \infty.$

Then, if Y_1, Y_2, \ldots is the sequence of observations generated by this process there exists a value of $\tau \in (0,1)$ only depending on ϵ , δ , and η , such that

$$\nu_{\delta}(n) = \mathbb{E}\left(\left|\log\left(\frac{P(Y_0|Y_{-1:-n})}{P(Y_0|Y_{-1:-\infty})}\right)\right|^{\frac{2+\delta}{1+\delta}}\right) \leqslant C\tau^n,$$

where C is a constant that depends only on ϵ and η .

The proof of this lemma is presented in Section 3.3.

In addition to the above mixing conditions, we require a notion of stability for the likelihood of a process for our finite sample concentration results. To understand this notion we should first define the Hamming distance between two vectors.

Definition 2.4. The Hamming distance between two sequences $x_{1:n} \in \mathbb{R}^n$ and $y_{1:n} \in \mathbb{R}^n$ is defined as

$$d_n(x_{1:n}, y_{1:n}) \triangleq \sum_{i=1}^n \mathbb{I}_{x_i \neq y_i},$$

where \mathbb{I} denotes the indicator function.

Where $\mathcal{L}(X_1|X_{0:-b+1})$ is the conditional distribution knowing $X_{0:-b+1}$ of X_1 , and $\mathcal{L}(X_1|X_{0:-\infty})$ is the conditional distribution knowing $X_{0:-\infty}$ of X_1

This notion enables us to define the notion of M-stability.

Definition 2.5. The M-stability coefficient of a finite state m-Markov process $\{X_i\}_{i=-\infty}^{\infty}$, with $X_i \in A$, is defined as

$$M \triangleq \sup_{n} \sup_{(X_{1:n}, X'_{1:n}) \in A^{n2}, \text{ s.t. } d_n(X_{1:n}, X'_{1:n}) \leq 1} |\log(P(X_{1:n}) - \log(P(X'_{1:n}))|.$$

We will say that $\{X_i\}_{i=-\infty}^{\infty}$ is M-stable if its M-stability coefficient is finite.

Remark. Consider a finite-state m-Markov chain $\{X_i\}_{i=-\infty}^{\infty}$. If $\rho \triangleq \min_{x_{1:-m} \in A^{m+2}} P(x_1|x_{-m:0}) > 0$, then the M-stability coefficient satisfies

$$M \leqslant (m+1)\log\left(\frac{1}{\rho}\right).$$

The proof of this claim is presented in Section 3.4.

The notion of M-stability will be used to obtain finite-sample concentration results. This notion can be seen in relation with the vast majority of concentration inequalities, such as Azuma, Hoefding, and McMiarmid that require boundedness conditions.

Now using the notions we developed above we state our first main result that confirms the asymptotic Gaussianity of the Kolmogorov complexity of ergodic sequences.

Theorem 3. Let $\{X_i\}_{-\infty}^{\infty}$ denote a stationary and ergodic process. We assume that $X_1 \in A$, where $A = \{a_1, ..., a_l\}$ with $l < \infty$. Furthermore, we suppose that

 C_1 . The Kolmogorov complexity of all a_js is finite, i.e., $\max_{i\in\{1,\ldots,l\}}K(a_i)<\infty$.

 C_2 . We assume that there are fixed numbers $K, \beta > 1, C > 1$, and $\delta \in (0,1]$, such that

$$-\alpha(n) \leqslant K n^{-\beta \frac{(2+\delta)(1+\delta)}{\delta^2}}.$$
$$-\nu_{\delta}(n)^{\frac{1+\delta}{2+\delta}} = O\left(2^{-Cn\log(l)}\right).$$

If we define

$$\sigma^{2} \triangleq \text{var}(\log(P(X_{0}|X_{-1},\ldots,X_{-\infty}))) + 2\sum_{k} \text{cov}(\log(P(X_{0}|X_{-1},\ldots,X_{-\infty})), \log(P(X_{k}|X_{k-1},\ldots,X_{-\infty}))),$$

then $\sigma^2 < \infty$, and

$$\sqrt{n}\left(\frac{K(X_{1:n})}{n} - H(X_0|X_{-1},\dots,X_{-\infty})\right) \to_d N(0,\sigma^2),$$

where the notation \rightarrow_d is used for the convergence in distribution.

The proof of this theorem is presented in Section 3.5. Note that Theorem 3 implies Theorem 1. However, this result provides the rate of convergence as well. Both Theorem 1 and Theorem 3 are concerned with the asymptotic behavior of the Kolmogorov complexity, and do not provide any information on the finite sample behavior of this quantity. The following corollary simplifies the statement of this theorem for an independent and identically distributed sequence.

Corollary 3.1. Let $\{X_i\}_{-\infty}^{\infty}$ denote an independent and identically distributed process. We assume that $X_1 \in A$, where $A = \{a_1, ..., a_l\}$ with $l < \infty$. Furthermore, we assume that $\max_{i \in \{1, ..., l\}} K(a_i) < \infty$. Then,

$$\sqrt{n}\left(\frac{K(X_{1:n})}{n} - H(X_0)\right) \to_d N(0, \sigma^2),$$

where $\sigma^2 = \text{var}(\log(P(X_0)))$.

Our next goal is to derive probabilistic upper bounds on the discrepancy of the Kolmogorov complexity and Shannon entropy in finite sample sizes. Our next theorem shows that such bounds can be obtained with slightly stronger mixing conditions than those in Theorem 3. For an integer number n define

$$\log^*(n) = \begin{cases} 0 & n \le 1\\ 1 + \log^*(\log(n)) & n > 1. \end{cases}$$
 (2)

Theorem 4. Let $\{X_i\}_{-\infty}^{\infty}$ denote a stationary m-Markov process. We assume that $X_1 \in A$, where $A = \{a_1, ..., a_l\}$ with $l < \infty$. Furthermore, we assume that

- 1. The Kolmogorov complexity of all a_js is finite, i.e., $\max_{i\in\{1,\ldots,l\}} K(a_i) < \infty$.
- 2. The M-stability coefficient of the process, M, is finite.
- 3. The ϕ -mixing coefficients of the process satisfy $\Delta \triangleq 1 + 24 \sum_{k=0}^{\infty} \phi(k) < \infty$.

Let $\eta \in (0,0.5)$ be a fixed number. We will have C' a constant the depends only on the universal mahcine, and define $\gamma_n \triangleq \frac{C_1(n)}{n} + \frac{m}{n}H(X_1|X_{0:-m+1}) + n^{-\frac{1}{2}-\eta} = O(n^{-\frac{1}{2}-\eta})$, where $C_1(n) \triangleq C' + \log^*(m) + l \max_{j \leq l} K(a_j) + l^{(m+1)} \log^* n - m \log^* l$. Moreover have $\gamma'(n) \triangleq C' + \log^*(m) + l \max_{j \leq l} K(a_j) + l^{m+1} \log^* n + m \log^* l + mH(X_1) = O(\frac{\log^*(n)}{n})$, $K_1 = 2M^2\Delta^2$ and $K_1'(n) \triangleq 2\Delta^2[C' + \log^*(n) + \max_i K(a_i)]^2$. Finally let ζ be a constant less than or equal to $C' + \max_{i \leq l} K(a_i)$

Then for any $t > \gamma'(n)$,

$$P\left(\left|\frac{1}{n}K(X_{1:n}) - H(X_1|X_{0:-m+1})\right| > t\right) \le 2e^{-\frac{n(t-\gamma'(n))^2}{K_1'(n)}} \sim 2e^{-\frac{nt^2}{2\Delta^2 \log^*(n)^2}},\tag{3}$$

Furthermore, for any $t > \gamma_n$ we have

$$\mathbb{P}\left(\left|\frac{K(X_{1:n})}{n} - H(X_1|X_{0:-m+1})\right| \ge t\right) \le 2e^{-\frac{n(t-\gamma_n)^2}{K_1}} + n\zeta 2^{-n^{\frac{1}{2}-\eta}},\tag{4}$$

Theorem 4 can be formulated in the following slightly different way:

Corollary 4.1. Let $\{X_i\}_{-\infty}^{\infty}$ be a m-markov process that satisfies all the conditions of Theorem 4. Fix K_1 to be the value defined in Theorem 4. Then, for every $\epsilon > 0$, $\exists N$ such that $\forall n \geq N$

$$\frac{P(\sqrt{n}|\frac{K(X_{1:n})}{n} - H(X_1|X_{0:-m+1})| \ge t)}{2e^{-\frac{t^2}{K_1}}} \le 1 + \epsilon$$

Proof. A straightforward application of Theorem 4.

3 Proof

3.1 Background on Kolmogorov complexity

There are two simple results on the Komogorov complexity that we employ in our proofs. We mention these two as simple lemmas that we can refer to later in the proofs of our main results. For the proof of these results a reader may refer to [18], Chapter 14 (Example 14.2.7 and Theorem 14.2.4)

Lemma 5. Let n denote an integer number. Then we have the following upper-bound on the Kolmogorov complexity of n:

$$K(n) \leq \log^*(n) + c$$

where

$$\log^*(x) = \begin{cases} 0 & x \le 1\\ 1 + \log^*(\log(x)) & x > 1, \end{cases}$$
 (5)

and where c is a constant that depends only on the universal machine.

It is straightforward to show that $\forall n \ge 1, \log^*(n) < 2\log(n) + 2$. Another result that will be used about the Komogorov complexity in our paper is the following:

Lemma 6. Let
$$\{0,1\}^{\infty} \triangleq \bigcup_{i=1}^{\infty} \{0,1\}^{i}$$
. If $C_{v} \triangleq \{x \in \{0,1\}^{\infty} \mid K(x) < v\}$, then $|C_{v}| \leq 2^{v}$.

3.2 Background information on mixing sequences

In our proofs we will also use some well-known results on the central limit theorem for the empirical average of weakly dependent sequences. We summarize these results in this section.

Theorem 7. [19] Let $\{X_i\}_{i=1}^{\infty}$ denote a stationary process with $\mathbb{E}(X_1) = 0$ and $\mathbb{E}(|X_1|^{2+\delta}) < \infty$ for some $\delta \in (0,1]$. Let $n \in \mathbb{N}$ and define

$$\sigma_n^2 \triangleq \operatorname{var}(\sum_{i=1}^n X_i).$$

Suppose that $\frac{\sigma_n^2}{n} \to \sigma^2$, where $\sigma^2 \in (0, \infty)$. Let F_n denote the cdf (cumulative distribution function) of $\frac{\sum_{i=1}^n X_i}{\sigma_n}$. If the α -mixing coefficients satisfy

$$\alpha \triangleq \sum_{i=1}^{\infty} (\alpha(i))^{\frac{\delta}{2+\delta}} < \infty,$$

and there exist k > 1 and m such that the following conditions hold:

$$(C.1) \ k \geqslant \frac{\log(n)}{2\log(16)},$$

$$(C.2) k^{\frac{3}{2}} 4^k (\alpha(m+1))^{\frac{1}{2+\delta}} \leq 1,$$

$$(C.3) 2km + 1 < n.$$

then there is a constant C, that does not depend on the process or n, such that for any n that satisfies $\frac{\sigma_n^2}{n} \geqslant \frac{1}{4}\sigma^2$, we have

$$\sup_{t} |F_{n}(t) - \Phi(t)| \leq C \left[x^{2+\delta} \frac{(m+1)^{\delta+1}}{B_{n}^{\delta}} + x^{3} \frac{(m+1)^{2}}{B_{n}} + x^{2} ((m+1)^{\frac{1}{2}} + \alpha^{\frac{1}{2}}) \frac{m+1}{B_{n}} + x^{2} (1+\alpha)^{\frac{1}{2}} B_{n} (\alpha(m+1))^{\frac{\delta}{2(2+\delta)}} + x ((m+1)^{\frac{1}{2}} + \alpha^{\frac{1}{2}}) (\alpha(m+1))^{\frac{\delta}{2+\delta}} \right],$$

where $x \triangleq \frac{2\mathbb{E}(|X_1|^{2+\delta})^{\frac{1}{2+\delta}}}{\sigma}$ and $B_n \triangleq \frac{2\sigma_n}{\sigma}$.

In the proof of Theorem 3 we will approximate the Kolmogorov complexity using triangular arrays. We would like to show that the distribution of $S_{n,n}$, the sum of the first n elements of the n-th row of a triangular array, converges to a normal distribution. To obtain that we will use the following corollary of Theorem 7.

Corollary 7.1. Let $\{X_i^k\}_{i,k=1}^{\infty}$ be a double-index process. Furthermore, let $\alpha^k(n)$ denote the α -mixing coefficients of $\{X_i^k\}_{i=1}^{\infty}$. Assume that $\{X_i^k\}_{i=1}^{\infty}$ is a stationary process, with $\mathbb{E}(X_1^k) = 0$. Suppose there exists a value of $\delta \in (0,1]$ such that $\forall \zeta > 0, \mathbb{E}(|X_1^k|^{2+\delta}) = o(k^{\zeta})$, and that $\mathbb{E}(|X_1^k|^{2+\delta}) < \infty$, for all k. Let $n \in \mathbb{N}$ and define $\sigma_n^2 \triangleq \text{var}(\sum_{i=1}^n X_i^n)$. Suppose that $\frac{\sigma_n^2}{n} \to \sigma^2$, where $\sigma^2 \in (0,\infty)$. Let F_n denote the cdf (cumulative distribution function) of $\frac{\sum_{i=1}^n X_i^n}{\sigma_n}$. Suppose that there exist $\epsilon > 0$ and $\beta > 1$ such that

- $\epsilon < \min(\frac{\delta}{(\beta+1)(\delta+1)}, 1 \delta \frac{\beta-1}{\beta+1})$,
- $\forall (n,j), \ \alpha^j(n) \leqslant \min(C'(n-j^{\epsilon})^{-\beta\frac{(2+\delta)(1+\delta)}{\delta^2}}, 2), \ where \ C' \ is \ a \ fixed \ number.$

Then.

$$\sup_{t} |F_n(t) - \Phi(t)| = O_n[\mathbb{E}(|X_1^n|^{2+\delta})n^{-\frac{\delta(\beta-1)}{2(\beta+1)}}] = o_n(n^{-\eta}), \text{ for all } \eta < \frac{\delta(\beta-1)}{2(\beta+1)}$$

Proof. We would like to use Theorem 7. Consider the n^{th} sequence X_1^n, X_2^n, \ldots Note that the α -mixing coefficient of this sequence $\alpha^n(k) \leq \min((k-n^{\epsilon})^{-\beta\frac{(2+\delta)(1+\delta)}{\delta^2}}, 1)$. Without loss of generality and for notational simplicity, we assume C' = 1.

Hence, it is straightforward to see that $\sum_{k=1}^{\infty} \alpha^n(k) \leq n^{\epsilon} + d$, where d is a fixed number. To derive this inequality we have used the upper-bound $\alpha^n(k) \leq 1$, for $k \leq n^{\epsilon}$. Furthermore, for each n we choose (m, k) in Theorem 7 in the following way:

$$m_n = n^{\frac{\delta}{(\beta+1)(\delta+1)}}, \quad \text{and} \quad k_n = \frac{1}{4}\log(n).$$

It is straightforward to show that, for n sufficiently large, Conditions C.1, C.2, C.3, required in Theorem 7, hold. Furthermore, it is straightforward to check that $\sigma_n^2 \geqslant \frac{n\sigma^2}{4}$ as required in Theorem 7. Hence, we obtain

$$\sup_{t} |F_{n}(t) - \Phi(t)| \leq C \left[x_{n}^{2+\delta} \frac{(m_{m}+1)^{\delta+1}}{B_{n}^{\delta}} + x_{n}^{3} \frac{(m_{n}+1)^{2}}{B_{n}} + x_{n}^{3} \frac{(m_{n}+1)^{2}}{B_{n}} + x_{n}^{2} ((m_{n}+1)^{\frac{1}{2}} + \alpha^{\frac{1}{2}}) \frac{m_{n}+1}{B_{n}} + x_{n}^{2} (1+\alpha)^{\frac{1}{2}} B_{n} (\alpha(m_{n}+1))^{\frac{\delta}{2(2+\delta)}} + x_{n} ((m_{n}+1)^{\frac{1}{2}} + \alpha^{\frac{1}{2}}) (\alpha(m_{n}+1))^{\frac{\delta}{2+\delta}} \right],$$

where $x_n \triangleq \frac{2\mathbb{E}(|X_1^n|^{2+\delta})^{\frac{1}{2+\delta}}}{\sigma}$. It is straightforward to check that the dominant term is $x_n^{2+\delta} \frac{(m_n+1)^{\delta+1}}{B_n^{\delta}} = O_n[\mathbb{E}(|X_1^n|^{2+\delta})n^{-\frac{\delta(\beta-1)}{2(\beta+1)}}]$. Hence, the proof is complete.

The following lemma enables us to connect the correlation of two random variables that are respectively F_0 and F_n measurable to the mixing coefficients.

Lemma 8. [20] Let the random variables ξ , η be measurable with respect to $\mathbb{F}^t_{-\infty}$ and $\mathbb{F}^{\infty}_{t+\tau}$ respectively. Suppose that there is $\delta > 0$, such that

$$\mathbb{E}(|\xi|^{2+\delta}) < c_1 < \infty$$
 and $\mathbb{E}(|\eta|^{2+\delta}) < c_2 < \infty$.

Then,

$$|\mathbb{E}(\xi\eta) - \mathbb{E}(\xi)\mathbb{E}(\eta)| \le \alpha(\tau)^{1-\frac{2}{2+\delta}} (4 + 3(c_1^{\beta}c_2^{1-\beta} + c_1^{1-\beta}c_2^{\beta})),$$

where $\beta \triangleq \frac{1}{2+\delta}$.

Most concentration inequalities on random processes assume independence. However here we do not want to make such assumptions. In the proof of Theorem 4 we will use the following result by Kontorovich and Ramanan that generalizes the martingale method to dependent variables:

Lemma 9. [21] Suppose that Ω is a countable space, and let $\{X_i\}_{i=-\infty}^{\infty}$ be a stationary process with $X_i \in \Omega$. Furthermore, let $g: \Omega^n \to \mathbb{R}$ be a 1-Lipschitz function with respect to the Hamming metric on Ω^n . Define $\Phi'_{i,j} \triangleq \sup_{x_{0:i},y_{0:i}} \|P(X_{j:n} \in \cdot | X_{0:i} = x_{0:i}) - P(X_{j:n} \in \cdot | X_{0:i} = y_{0:i})\|_{TV}$. Let H_n be an $n \times n$ matrix defined in the following way:

$$H_{n,\{i,j\}} = \begin{cases} 1, & \text{if } i = j \\ \Phi'_{i,j}, & \text{if } i < j \\ 0 & \text{otherwise} \end{cases}.$$

Then, for all t > 0 we have

•
$$P(g(X_{1:n}) - \mathbb{E}(g(X_{1:n})) \ge t) \le e^{-\frac{t^2}{2n\Delta_n^2}}$$

•
$$P(g(X_{1:n}) - \mathbb{E}(g(X_{1:n})) \le -t) \le e^{-\frac{t^2}{2n\Delta_n^2}}$$
,

where $\Delta_n \triangleq \|H_n\|_{\infty} = \max_{i \leq n} (1 + \Phi_{i,i+1} + \dots + \Phi_{i,n}).$

Remark. Note that the lemma proposed in [21] has a two-sided bound. Here we use a one-sided version. Furthermore note that the conditions on $\Phi'_{i,j}$ is a bit stronger than the one proposed in [21], but for simplicity we use this condition.

3.3 Proof of Lemma 2

According to Proposition 1 in [22], there exits $\tau \in (0,1)$ such that

$$\mathbb{E}(\|P(Y_0 \in \cdot | Y_{-1 \cdot -m}) - P(Y_0 \in \cdot | Y_{-1 \cdot -m})\|_{TV}) \leqslant C\tau^m.$$

Also, note the following three facts that are straightforward to prove:

- (i) $h(x) \triangleq \frac{x|\log(x)|^{\frac{2+\delta}{1+\delta}}}{|x-1|}$ is an increasing function of $x \in (1, \infty)$.
- (ii) $h(x) \le 2 \text{ for } x \in (0, 1].$

$$\text{(ii)} \ \ \frac{dP(Y_0|Y_{-1:-\infty})}{dP(Y_0|Y_{-1:-m})} = \frac{\int_{x_0} dP(x_0|Y_{-1:-\infty})g(Y_0|x_0)dx_0}{\int_{x_0} dP(x_0|Y_{-1:-m})g(Y_0|x_0)dx_0} \leqslant \frac{\mathrm{esssup}_{x_0}g(Y_0|x_0)}{\mathrm{essinf}_{x_0}g(Y_0|x_0)} \leqslant \eta.$$

By employing these facts we obtain

$$\begin{split} &\mathbb{E}(|\log(\frac{P(Y_0|Y_{-1:-\infty})}{P(Y_0|Y_{-1:-m})})|^{\frac{2+\delta}{1+\delta}})\\ &= \mathbb{E}(\int |\log(\frac{dP(Y_0|Y_{-1:-\infty})}{dP(Y_0|Y_{-1:-m})})|^{\frac{2+\delta}{1+\delta}}dP(Y_0|Y_{-1:-\infty}))\\ &= \mathbb{E}(\int |\log(\frac{dP(Y_0|Y_{-1:-\infty})}{dP(Y_0|Y_{-1:-m})})|^{\frac{2+\delta}{1+\delta}}\frac{dP(Y_0|Y_{-1:-\infty})}{dP(Y_0|Y_{-1:-m})}dP(Y_0|Y_{-1:-m}))\\ &\leqslant \max(2, \frac{\eta|\log(\eta)|^{\frac{2+\delta}{1+\delta}}}{|\eta-1|})\mathbb{E}(\int |1-\frac{dP(Y_0|Y_{-1:-\infty})}{dP(Y_0|Y_{-1:-m})}|dP(Y_0|Y_{-1:-m}))\\ &\leqslant 2\max(2, \frac{\eta|\log(\eta)|^{\frac{2+\delta}{1+\delta}}}{|\eta-1|})\mathbb{E}(\|P(Y_0\in\cdot|Y_{-1:-m})-P(Y_0\in\cdot|Y_{-1:-\infty})\|_{TV})\\ &\leqslant C'\tau^m. \end{split}$$

Note that similar ideas have been used in [23].

3.4 Proof of Remark 2

For $n \in \mathbb{N}$, consider the two vectors $x, x' \in A^n$ such that $d_n(x, x') \leq 1$. If $d_n(x, x') = 0$, then we can easily see that $|\log(P(x) - \log(P(x'))| = 0$. Hence, we assume that $d_n(x, x') = 1$. Suppose that $x_i \neq x_i'$. If $i \in [2, |n - m - 1|]$, then

$$\begin{aligned} &|\log(P(x_{1:n}) - \log(P(x'_{1:n}))| \\ &\leqslant |\log(P(x_{1:i-1}) - \log(P(x'_{1:i-1}))| + \sum_{j=i}^{m+1+i} |\log(P(x_j|x_{j-1:j-m})) - \log(P(x'_j|x'_{j-1:j-m}))| \\ &+ |\log(P(x_{m+1+i:n}|x_{i+1:m+i})) - \log(P(x'_{m+1+i:n}|x'_{i+1:m+i}))| \\ &= \sum_{j=i}^{m+1+i} |\log(P(x_j|x_{j-1:j-m})) - \log(P(x'_j|x'_{j-1:j-m}))| \leqslant -(m+1)\log(\rho). \end{aligned}$$

This comes from the following facts: (i) For every $j < i, x_{1:j} = x'_{1:j}$. Hence, $|\log(P(x_{1:i-1}) - \log(P(x'_{1:i-1}))| = 0$, (ii) For every $j > m+1+i, x_{j:n} = x'_{j:n}$. Hence, $|\log(P(x_{m+1+i:n}|x_{i+1:m+i})) - \log(P(x'_{m+1+i:n}|x'_{i+1:m+i}))| = 0$. (iii) Finally, $\forall i \in [|i, m+1+i|] |\log(P(x_j|x_{j-1:j-m})) - \log(P(x'_j|x'_{j-1:j-m}))| \le -\log(\rho)$. The proof for $i \notin [2, |n-m-1|]$ is similar and is hence skipped.

3.5 Proof of Theorem 3

3.5.1 Lower bound

Proof. Before we discuss the details of the proof, we give a brief overview of the proof strategy to help the reader navigate through the proof more easily. Consider the sequence X_1, X_2, \ldots, X_n with $X_i \in A$, for all i. We assume that |A| = l. In this section, we first present a simple program that a universal computer can use to generate this sequence.

that a universal computer can use to generate this sequence. Define $m_n \triangleq \frac{\frac{1}{2} - \epsilon}{\log(l)} \log(n)$, where $\frac{1}{2} - \frac{1}{2C} > \epsilon > 0$. Note that C is the same constant as the one used in Condition 3 in the statement of the theorem. The program first tells the universal

computer the first m_n bits in the sequence. Then, counts the number of times each (m_n+1) -tuple is present in the remaining sequence and reports it.³ In other words, if we define

$$f_j^{m_n,n} \triangleq \frac{\sum_{k=m_n+1}^n \mathbb{I}_{X_{k-m_n:k} = a_j^{m_n}}}{n - m_n},\tag{6}$$

where $a_j^{m_n}$ is the j^{th} element (in a specific order that is described to the universal computer) of A^{m_n} , then the numbers $f_j^{m_n,n}$ are described to the universal computer. Let $\mathbf{f}^{m_n,n}$ denote the vector of all the empirical counts, i.e.,

$$\mathbf{f}^{m_n,n} \triangleq (f_1^{m_n,n}, f_2^{m,n}, \dots, f_{l^{m_n+1}}^{m_n,n}).$$

Define an operator $O_f: A^n \to [0,1]^{l^{m_n+1}}$ that takes X_1, X_2, \dots, X_n as input and returns $\mathbf{f}^{m_n,n}$ as its output. Then, define the type of a sequence $X_{1:n}$ as the following set:

$$\mathcal{T}_{X_{1:n}} \triangleq \{Z_{1:n} : O_f(X_{1:n}) = O_f(Z_{1:n}) \text{ and } Z_{1:m_n} = X_{1:m_n}\}.$$

Given the information known to the universal computer so far, it has already access to $\mathcal{T}_{X_{1:n}}$. The only remaining piece of information that the universal computer should have to reconstruct the entire sequence is the index of the sequence X_1, X_2, \ldots, X_n among all the sequences in its type. Let's count the number of bits we have used so far to describe the sequence.

Our description requires bits to specify the following quantities: (i) m_n , (ii)each a_j , (iii) the first m_n bits, (iv)the frequency of observing each possible block of length $(m_n + 1)$ in $X_{1:n}$, (v) a systematic way to build all the sequences of length n in $\mathcal{T}_{X_{1:n}}$, (vi) the index of $X_{1:n}$ in $\mathcal{T}_{X_{1:n}}$.

- (i) $K(m_n) \leq \log^*(m_n) + c$.
- (ii) To describe each a_j at most $l \max_{j \leq l} K(a_j)$ are required.
- (iii) To describe the first m_n symbols we require $m_n(\log^*(l) + c)$.
- (iv) To describe the frequency of each block we require $l^{m_n+1} \log^*(n)$ bits. The reason is clear, there are l^{m_n+1} different l-ary blocks of length m_n+1 . Each of them can have at most n elements in them.
- (iv) So far the universal computer has detected $\mathcal{T}_{X_{1:n}}$. Now we should describe which element of $\mathcal{T}_{X_{1:n}}$ $X_{1:n}$ is. As the first step we write a constant size program so that the universal computer realizes what ordering of sequences we are using. The next step is to specify the index of our sequence in this list. To evaluate the number of bits required for describing the index we count the number of elements in $\mathcal{T}_{X_{1:n}}$.

Define \tilde{P}^{m_n} as a new measure on X_1, X_2, \ldots, X_n that has the following properties:

1. \tilde{P}^{m_n} has the m_n -Markov property, i.e.,

$$\tilde{P}^{m_n}(X_1,\ldots,X_n) = \tilde{P}^{m_n}(X_1,X_2,\ldots,X_{m_n}) \prod_{j=m_n+1}^n \tilde{P}^{m_n}(X_j \mid X_{j-1},\ldots,X_{j-m_n}).$$

2. The $m_n + 1$ th-dimension transition probabilities are the same as those of the original distribution P, i.e.,

$$\tilde{P}^{m_n}(X_j \mid X_{j-1}, \dots, X_{j-m_n}) = P(X_j \mid X_{j-1}, \dots, X_{j-m_n}).$$

³For instance, if $m_n = 1$, then for the sequence 01001 the couple (0, 1) is present twice, the couple (1, 0) once and (0, 0) once.

For notational simplicity we consider the notation

$$Q_j^{m_n} \triangleq \tilde{P}^{m_n}(X_{m_n} = a_{j,m_n}^{m_n} | X_0 = a_{j,0}^{m_n}, \dots, X_{m_n-1} = a_{j,m_n-1}^{m_n}), \tag{7}$$

where $(a_{j,0}^{m_n}, \ldots, a_{j,m_n}^{m_n})$ is the j^{th} element of A^{m_n+1} . With this new notation we count the number of elements in $\mathcal{T}_{X_1:n}$. Note that the first m_n symbols are already known. Let's call them $x_1, x_2, \ldots, x_{m_n}$. Since,

$$\sum_{X_{1:n} \in \mathcal{T}_{X_{1:n}}} \tilde{P}^{m_n} (X_{m_n+1}, \dots, X_n \mid X_1 = x_1, X_2 = x_2, \dots, X_{m_n} = x_{m_n}) \le 1,$$

we have

$$\sum_{X_{1:n}\in\mathcal{T}_{X_{1:n}}}\prod_{j=m_n+1}^n \tilde{P}^{m_n}(X_j|X_{j-1},\ldots,X_{j-m_n}) = |\mathcal{T}_{X_{1:n}}|\prod_{j=1}^{\ell^{m_n+1}} (Q_j^{m_n})^{(n-m_n)f_j^{m_n,n}}$$

Hence,

$$|\mathcal{T}_{X_{1:n}}| < 2^{-(n-m_n)\sum_{j=1}^{l(m_n+1)} f_j^{m_n,n} \log Q_j^{m_n}$$

This implies that to code the index of an element of $\mathcal{T}_{X_{1:n}}$, we require less than $-(n-m_n)\sum_{j=1}^{l^{m_n+1}}f_j^{m_n,n}\log Q_j^{m_n}$ bits. Combining all the above pieces we obtain the following upper bound for the length of our program:

$$K(X_{1:n}) \leq C' + \log^*(m_n) + l \max_{j \leq l} K(a_j) + l^{(m_n+1)} \log^* n + m_n \log^* l - (n - m_n) \sum_{j=1}^{l^{m_n+1}} f_j^{m_n, n} \log Q_j^{m_n}$$
(8)

Our goal is to show that $\frac{K(X_1,\dots,X_n)}{\sqrt{n}} - \sqrt{n}H(X_1|X_{0:-\infty})$ converges in distribution to a normal random variable. Note that the first five terms in (8) are deterministic and when divided by \sqrt{n} , they converge to zero. Hence, we focus on the only remaining term, i.e., $(n-m_n)\sum_{j=1}^{l^{m_n+1}} f_j^{m,n} \log Q_j^{m_n}$. We have

$$\sqrt{n} \left(\frac{1}{n}(n-m_n) \sum_{j=1}^{l^{m_n+1}} f_j^{m_n,n} \log Q_j^{m_n} + H(X_0|X_{-1},\dots,X_{-\infty})\right)
= \sqrt{n} \left(\frac{n-m_n}{n} \sum_{j=1}^{l^{m_n+1}} f_j^{m_n,n} \log Q_j^{m_n} + H(X_0|X_{-1},\dots,X_{-m_n})\right)
+ \sqrt{n} \left(H(X_0|X_{-1},\dots,X_{-\infty}) - H(X_0|X_{-1},\dots,X_{-m_n})\right).$$
(9)

Our first claim is that

$$\sqrt{n}(H(X_0|X_{-1},\dots,X_{-\infty}) - H(X_0|X_{-1},\dots,X_{-m_n})) \to 0,$$
(10)

as $n \to 0$. To see why this holds, note that

$$\sqrt{n}|H(X_{0}|X_{-1},\ldots,X_{-\infty}) - H(X_{0}|X_{-1},\ldots,X_{-m_{n}}))|
\leq \sqrt{n}\mathbb{E}|\log P(X_{0} \mid X_{-1},\ldots,X_{\infty}) - \log P(X_{0} \mid X_{-1},\ldots,X_{-m_{n}})|
\stackrel{(a)}{\leq} \sqrt{n}(\mathbb{E}|\log P(X_{0} \mid X_{-1},\ldots,X_{\infty}) - \log P(X_{0} \mid X_{-1},\ldots,X_{-m_{n}})|^{\frac{2+\delta}{1+\delta}})^{\frac{1+\delta}{2+\delta}}
= \sqrt{n}(\nu^{\delta}(m_{n}))^{\frac{1+\delta}{2+\delta}} \to 0,$$

as $n \to \infty$. Here we should remind the reader that we have picked $m_n = \frac{\frac{1}{2} - \epsilon}{\log l} \log n$ with ϵ satisfying $\frac{1}{2} - \frac{1}{2C} > \epsilon > 0$. Note that to obtain (a) we have used Holder inequality and the last step is derived form condition 2 of the theorem regarding the decay of ν_{δ} . Combining (9) and (10) we conclude that the only remaining step is to show that $\sqrt{n}(\frac{n-m_n}{n}\sum_{j=1}^{l^{m_n+1}}f_j^{m_n,n}\log Q_j^{m_n} + H(X_0|X_{-1},\ldots,X_{-m_n}))$ is Gaussian. Toward this goal we first define

$$Y_j^{m_n} \triangleq \log P(X_j | X_{j-1}, X_{j-2}, \dots, X_{j-m_n}).$$

Note that $\sum_{j=1}^{l^{m_n+1}} f_j^{m_n,n} \log Q_j^{m_n} = \frac{1}{n-m_n} \sum_{j=m_n+1}^n Y_j^{m_n}$. Define

$$S_n^{m_n} \triangleq \sum_{i=m_n+1}^n Y_i^{m_n}.$$

To prove the Gaussianity of $S_n^{m_n}$ we employ Corollary 7.1. First let us check the conditions of this theorem for $Y_j^{m_n}$:

1. Boundedness of $\mathbb{E}|Y_j^{m_n}|^{2+\delta}$: First note that

$$\mathbb{E}|Y_{j}^{m_{n}}|^{2+\delta} = \mathbb{E}\sum_{x_{j}\in A} P(x_{j}|X_{j-1}, X_{j-2}, \dots, X_{j-m_{n}})|\log P(x_{j}|X_{j-1}, X_{j-2}, \dots, X_{j-m_{n}})|^{2+\delta}$$

$$= \mathbb{E}\sum_{x_{j}\in A} g(P(x_{j}|X_{j-1}, X_{j-2}, \dots, X_{j-m_{n}})), \tag{11}$$

where the function g is defined in the following way: $g:[0,1] \to \mathbb{R}$ and $g(t) = t|\log(t)|^{2+\delta}$ for $t \neq 0$, and also g(0) = 0. It is straightforward to check the following properties of g:

- (i) g(t) is continuous at zero.
- (ii) There exists $C_{\delta} \in (0,1)$ such that $g'(C_{\delta}) = 0$
- (iii) g'(t) > 0 for $t < C_{\delta}$
- (iv) $g'(t) \leq 0$ for $t > C_{\delta}$.

This automatically implies that $g(t) \leq g(C_{\delta})$ for all $t \in [0, 1]$. Combing this fact with (11) implies

$$\mathbb{E}|Y_j^{m_n}|^{2+\delta} = \mathbb{E}\sum_{X_j \in A} g(P(X_j|X_{j-1}, X_{j-2}, \dots, X_{j-m_n}) \le lg(C_\delta).$$
 (12)

Note that the upper bound does not depend on either m_n , n or j.

2. The mixing coefficient α : First let $\alpha^{Y^{m_n}}(i)$ denote the α -mixing coefficient for the Y^{m_n} sequence, and let $\alpha(i)$ denote the α mixing coefficient for the original process X_1, \ldots, X_n . It is straightforward to check that for every $i > m_n$

$$\alpha^{Y^{m_n}}(i) \leq \alpha(i - m_n) \leq \begin{cases} K(i - m_n)^{-\beta \frac{(2+\delta)(1+\delta)}{\delta^2}}, & i > m_n \\ 1, & \text{otherwise.} \end{cases}$$

where the last step is due to Condition 2 in the statement of the theorem. As a reminder we have $m_n = O(\log(n))$.

3. For notational simplicity in the rest of the proof we use the notation $\sum_{j=1}^{n-m_n} Y_j^{m_n}$ instead of $\sum_{j=m_n+1}^n Y_j^{m_n}$. Define $\tilde{\sigma}_n^2 = \text{var}(Y_{1:n-m_n}^{m_n})$. We will later prove that $\frac{\tilde{\sigma}_n^2}{n} \to \sigma^2$, where $\sigma^2 \triangleq \text{var}(\log(P(X_0|X_{-1},\ldots,X_{-\infty}))) + 2\sum_k \text{cov}(\log(P(X_0|X_{-1},\ldots,X_{-\infty}))), \log(P(X_k|X_{k-1},\ldots,X_{-\infty})))$.

First we can see that $\sigma^2 < \infty$. In that goal define

$$W_i \triangleq \log(P(X_i|X_{i/2:i-1})).$$

We have

$$\begin{split} & \sum_{k} \text{cov}(\log(P(X_{0}|X_{-1},\ldots,X_{-\infty})), \log(P(X_{k}|X_{k-1},\ldots,X_{-\infty}))) \\ & = \sum_{k} \text{cov}(\log(P(X_{0}|X_{-1},\ldots,X_{-\infty})), W_{k}) + \sum_{k} \text{cov}(\log(P(X_{0}|X_{-1},\ldots,X_{-\infty})), \log(P(X_{k}|X_{k-1},\ldots,X_{-\infty})) - W_{k}) \\ & \stackrel{(a)}{\leqslant} \sum_{k} \alpha(\frac{k}{2})^{\frac{\delta}{\delta+2}} (4 + 6lg(C_{\delta})) + \sum_{k} (\nu_{\delta}(k/2))^{\frac{1+\delta}{2+\delta}} \\ & \leqslant K(4 + 6lg(C_{\delta})) \sum_{k} n^{-\frac{\beta(1+\delta)}{\delta}} + \sum_{k} 2^{-C \log(\ell) \frac{k}{2}} < \infty. \end{split}$$

To obtain the first term in Inequality (a) we employed Lemma 8. To obtain the second term after Inequality (a) we used Holder's inequality and Definition 2.3. The last inequality is the result of Condition 2 in the statement of our theorem.

We can now prove that $\frac{\tilde{\sigma}_n^2}{n} \to \sigma^2$. We have

$$\frac{\operatorname{var}(\sum_{j=1}^{n-m_n} Y_j^{m_n})}{n-m_n} = \operatorname{var}(Y_1^{m_n}) + \frac{2}{n} \sum_{i=1}^n \sum_{k=i+1}^n \operatorname{cov}(Y_i^{m_n}, Y_k^{m_n})$$

$$= \operatorname{var}(Y_1^{m_n}) + \frac{2}{n} \sum_{i=1}^n \sum_{k=2}^i \operatorname{cov}(Y_1^{m_n}, Y_k^{m_n}), \tag{13}$$

where to obtain the last equality we used the stationarity of the process $Y_1^{m_n}, Y_2^{m_n}, \ldots$ Our goal is to show that this quantity converges to σ^2 . We simplify the expression of (13) in the following two steps:

1. Simplifying $var(Y_1^{m_n})$: First note that

$$\begin{aligned} & |\mathbb{E}(\log(P(X_1|X_{0:-m_n+1}))) - \mathbb{E}(\log(P(X_1|X_{0:-\infty})))| \\ & \leq (\mathbb{E}|(\log(P(X_1|X_{0:-m_n+1}))) - \log(P(X_1|X_{0:-\infty}))|^{\frac{2+\delta}{1+\delta}})^{\frac{1+\delta}{2+\delta}} = (\nu_{\delta}(m_n))^{\frac{1+\delta}{2+\delta}} \to 0. \end{aligned} \tag{14}$$

To obtain the last inequality we used Holder's and to obtain the last convergence we used Condition 2 in the statement of the theorem. Furthermore, note that

$$|\mathbb{E}(\log^{2}(P(X_{1}|X_{0:-m_{n}}))) - \mathbb{E}(\log^{2}(P(X_{1}|X_{0:-\infty})))|$$

$$\leq (\mathbb{E}|(\log(P(X_{1}|X_{0:-m_{n}+1}))) - \log(P(X_{1}|X_{0:-\infty}))|^{\frac{2+\delta}{1+\delta}})^{\frac{1+\delta}{2+\delta}}$$

$$\times (\mathbb{E}|(\log(P(X_{1}|X_{0:-m_{n}+1}))) + \log(P(X_{1}|X_{0:-\infty}))|^{2+\delta})^{\frac{1}{2+\delta}} \to 0.$$
(15)

To prove the last convergence we should note that the first term goes to zero according to Condition 2 in the statement of the theorem. Furthermore, similar to the proof of (12) we can show that the last expectation is bounded. Hence, it is straightforward to combine the above two equations and obtain

$$\operatorname{var}(Y_1^{m_n}) = \operatorname{var}(\log(P(X_1|X_{0:-m_n+1}))) \to \operatorname{var}(\log(P(X_1|X_{0:-\infty}))). \tag{16}$$

2. Our second step is to discuss the covariance terms in (13). Define

$$s_{i,n} \triangleq \sum_{k=2}^{i} \operatorname{cov}(Y_1^{m_n}, Y_k^{m_n}),$$

$$s \triangleq \sum_{j} \operatorname{cov}(\log(P(X_1|X_{-1:-\infty})), \log(P(X_j|X_{j-1:-\infty})))).$$

Note that our goal is to bound

$$\frac{1}{n} \left| \sum_{i=1}^{n} (s_{i,n} - s) \right| \le \frac{1}{n} \sum_{i=1}^{2m_n} |s_{i,n} - s| + \frac{1}{n} \sum_{i=2m_n+1}^{n} |s_{i,n} - s|.$$
 (17)

We will prove later that $\sup_i |s_{i,n} - s|$ is bounded. Hence, since $m_n/n \to 0$, we conclude that the first term goes to zero. Hence, we focus on the second term. Define $Z_j \triangleq \log(P(X_j|X_{j-1},\infty))$. Then we have

$$\frac{1}{n} \sum_{i=2m_n+1}^{n} |s_{i,n} - s| \leq \frac{1}{n} \sum_{i=2m_n+1}^{n} \sum_{j=2}^{2m_n} |\operatorname{cov}(Y_1^{m_n}, Y_j^{m_n}) - \operatorname{cov}(Z_1, Z_j)|
+ \frac{1}{n} \sum_{i=2m_n+1}^{n} \sum_{j=2m_n+1}^{i} |\operatorname{cov}(Y_1^{m_n}, Y_j^{m_n}) - \operatorname{cov}(Z_1, Z_j)|
+ \frac{1}{n} \sum_{i=2m_n+1}^{n} \sum_{j=i}^{\infty} |\operatorname{cov}(Z_1, Z_j)|.$$
(18)

We will show that the each of the three terms on the right converge to zero. Before we proceed further, note that

$$\mathbb{E}(|Y_1^{m_n} - Z_1|^{\frac{2+\delta}{1+\delta}}) = \mathbb{E}|\log(P(X_1|X_{0:-m_n+1})) - \log(P(X_1|X_{0:-\infty}))|^{\frac{2+\delta}{1+\delta}} = \nu_{\delta}(m_n). \quad (19)$$

Furthermore, similar to the proof of (12) it is straightforward to show that

$$\mathbb{E}(|Z_j|) \leqslant (\mathbb{E}|Z_j|^{2+\delta})^{\frac{1}{2+\delta}} < M,$$

$$\mathbb{E}|Y_j^{m_n}| \leqslant (\mathbb{E}|Y_j^{m_n}|^{2+\delta})^{\frac{1}{2+\delta}} < M,$$
(20)

where $M^{2+\delta} = l \sup_{t \in [0,1]} |g_{2+\delta}(t)|$ with $g_{2+\delta}(t) = t |\log(t)|^{2+\delta}$. Now we turn our attention to bounding the terms in (18).

$$\begin{aligned} |\text{cov}(Y_{1}^{m_{n}}, Y_{j}^{m_{n}}) - \text{cov}(Z_{1}, Z_{j})| &\leq |\text{cov}(Y_{1}^{m_{n}} - Z_{1}, Z_{j})| + |\text{cov}(Y_{1}^{m_{n}}, Z_{j} - Y_{j}^{m_{n}})| \\ &\leq \mathbb{E}|(Y_{1}^{m_{n}} - Z_{1})Z_{j}| + |\mathbb{E}(Y_{1}^{m_{n}} - Z_{1})\mathbb{E}Z_{j}| + \mathbb{E}|Y_{1}^{m_{n}}(Z_{j} - Y_{j}^{m_{n}})| + |\mathbb{E}(Y_{1}^{m_{n}})\mathbb{E}(Z_{j} - Y_{j}^{m_{n}})| \\ &\leq (\mathbb{E}|Y_{1}^{m_{n}} - Z_{1}|^{\frac{2+\delta}{1+\delta}})^{\frac{1+\delta}{1+\delta}} (\mathbb{E}|Z_{j}|^{2+\delta})^{\frac{1}{2+\delta}} + (\mathbb{E}|Y_{j}^{m_{n}} - Z_{j}|^{\frac{2+\delta}{1+\delta}})^{\frac{1+\delta}{2+\delta}} \mathbb{E}|Z_{j}| \\ &+ (\mathbb{E}|Y_{j}^{m_{n}} - Z_{j}|^{\frac{2+\delta}{1+\delta}})^{\frac{1+\delta}{2+\delta}} (\mathbb{E}|Y_{1}^{m_{n}}|^{2+\delta})^{\frac{1}{2+\delta}} + (\mathbb{E}|Y_{j}^{m_{n}} - Z_{j}|^{\frac{2+\delta}{1+\delta}})^{\frac{1+\delta}{2+\delta}} \mathbb{E}|Y_{1}^{m_{n}}| \\ &\leq 4M(\nu_{\delta}(m_{n}))^{\frac{1+\delta}{2+\delta}}. \end{aligned} \tag{21}$$

Hence, we conclude that

$$\frac{1}{n} \sum_{i=2m_n+1}^n \sum_{j=1}^{2m_n} |\text{cov}(Y_1^{m_n}, Y_j^{m_n}) - \text{cov}(Z_1, Z_j)| \leqslant \frac{n - 2m_n}{n} 2m_n 4M(\nu_{\delta}(m_n))^{\frac{1+\delta}{2+\delta}} \to 0,$$

as $n \to \infty$. Note that the last convergence in the theorem is derived from Condition 2 in the statement of the theorem. Now we find a bound on the second term in (18). Define

$$W_j \triangleq \log(P(X_j|X_{j/2:j-1})).$$

Then, we have

$$\begin{aligned} &|\operatorname{cov}(Y_{1}^{m_{n}}, Y_{j}^{m_{n}}) - \operatorname{cov}(Z_{1}, Z_{j})| \leq |\operatorname{cov}(Y_{1}^{m_{n}} - Z_{1}, Z_{j})| + |\operatorname{cov}(Y_{1}^{m_{n}}, Z_{j} - Y_{j}^{m_{n}})| \\ &\leq |\operatorname{cov}(Y_{1}^{m_{n}} - Z_{1}, Z_{j} - W_{j})| + |\operatorname{cov}(Y_{1}^{m_{n}} - Z_{1}, W_{j})| + |\operatorname{cov}(Y_{1}^{m_{n}}, Z_{j} - Y_{j}^{m_{n}})| \\ &\leq |\operatorname{cov}(Y_{1}^{m_{n}} - Z_{1}, Z_{j} - W_{j})| + |\operatorname{cov}(Y_{1}^{m_{n}} - Z_{1}, W_{j})| + |\operatorname{cov}(Y_{1}^{m_{n}}, Z_{j} - W_{j})| \\ &+ |\operatorname{cov}(Y_{1}^{m_{n}}, W_{j})| + |\operatorname{cov}(Y_{1}^{m_{n}}, Y_{j}^{m_{n}})| \end{aligned} \tag{22}$$

The strategy that we use to bound the terms $|\operatorname{cov}(Y_1^{m_n}-Z_1,Z_j-W_j)|$ and $|\operatorname{cov}(Y_1^{m_n},Z_j-W_j)|$ is the same. Also, the strategy we use to bound $|\operatorname{cov}(Y_1^{m_n}-Z_1,W_j)|$ and $|\operatorname{cov}(Y_1^{m_n},W_j)|$ is the same. Hence, we only derive the bounds for the following three terms: (i) $|\operatorname{cov}(Y_1^{m_n},Z_j-W_j)|$, (ii) $|\operatorname{cov}(Y_1^{m_n},W_j)|$, and (iii) $|\operatorname{cov}(Y_1^{m_n},Y_j^{m_n})|$.

(a) $|\text{cov}(Y_1^{m_n}, Z_j - W_j)|$: By using holder inequality we conclude that

$$|\operatorname{cov}(Y_{1}^{m_{n}}, Z_{j} - W_{j})| \leq \mathbb{E}|Y_{1}^{m_{n}}(Z_{j} - W_{j})| + \mathbb{E}|Y_{1}^{m_{n}}|\mathbb{E}|Z_{j} - W_{j}|$$

$$\leq 2(\mathbb{E}|Z_{j} - W_{j}|^{\frac{2+\delta}{1+\delta}})^{\frac{1+\delta}{2+\delta}}\mathbb{E}(|Y_{1}^{m_{n}}|^{2+\delta})^{\frac{1}{2+\delta}}$$

$$\leq 2(\nu_{\delta}(j/2))^{\frac{1+\delta}{2+\delta}}M. \tag{23}$$

(b) $|\text{cov}(Y_1^{m_n}, W_j)|$: Note that W_j is measurable with respect to $\mathcal{F}_{j/2}^j$ and $Y_1^{m_n}$ is measurable with respect to $\mathcal{F}_{-\infty}^1$. Hence, by employing Lemma 7 we conclude that

$$|\operatorname{cov}(Y_1^{m_n}, W_j)| \le \alpha(j/2)^{\frac{\delta}{\delta+2}} (4 + 2\tilde{M}),$$

where $\tilde{M} = lg(C_{\delta})$. Note that to obtain the last inequality we have used (12).

(c) $|\text{cov}(Y_1^{m_n}, Y_i^{m_n})|$: Similar to the argument of the previous case we conclude that

$$|\operatorname{cov}(Y_1^{m_n}, Y_j^{m_n})|| \leq \alpha(j - m_n)^{\frac{\delta}{\delta + 2}} (4 + 2\tilde{M}).$$

Combining (22) and the above three cases, we conclude that

$$\frac{1}{n} \sum_{i=2m_n+1}^{n} \sum_{j=2m_n+1}^{i} |\operatorname{cov}(Y_1^{m_n}, Y_j^{m_n}) - \operatorname{cov}(Z_1, Z_j)| \\
\leq \frac{1}{n} \sum_{i=2m_n+1}^{n} \sum_{j=2m_n+1}^{\infty} 4(\nu_{\delta}(j/2))^{\frac{1+\delta}{2+\delta}} M + 2\alpha(j/2)^{\frac{\delta}{\delta+2}} (4+2M) + \alpha(j-m_n)^{\frac{\delta}{\delta+2}} (4+2M) \\
\leq \sum_{j=2m_n+1}^{\infty} 4(\nu_{\delta}(j/2))^{\frac{1+\delta}{2+\delta}} M + 2\alpha(j/2)^{\frac{\delta}{\delta+2}} (4+2M) + \alpha(j-m_n)^{\frac{\delta}{\delta+2}} (4+2M) \to 0, \quad (24)$$

as $n \to \infty$. The last term of (18) can be bounded in exactly similar fashion, i.e., we use the upper bound $|\text{cov}(Z_1, Z_j)| \le |\text{cov}(Z_1, Z_j - W_j)| + |\text{cov}(Z_1, W_j)|$, and then employ Lemma 7 and the definition of ν_{δ} to bound the error. Since the proof is similar we skip it.

Combining all these steps we conclude that

$$\frac{1}{n} |\sum_{i=1}^{n} (s_{i,n} - s)| \to 0.$$
 (25)

Equations (13), (16), and (25) together prove that

$$\frac{\operatorname{var}(\sum_{j=1}^{n-m_n} Y_j^{m_n})}{n-m_n} \to \sigma^2.$$

Therefore if $\sigma^2 = 0$ we have proved that:

$$\frac{1}{\sqrt{n-m_n}} \sum_{j=1}^{n-m_n} [Y_j^{m_n} - H(X_1|X_{0:-m_n+1})] \xrightarrow{L_2} 0.$$

Hence

$$\frac{1}{\sqrt{n}} \left[C' + \log^*(m_n) + l \max_{j \le l} K(a_j) + l^{(m_n+1)} \log^* n - m_n \log^* l - (n - m_n) \sum_{j=1}^{l^{m_n+1}} f_j^{m_n, n} \log Q_j^{m_n} \right] \stackrel{d}{\to} 0.$$

Now if $\sigma^2 > 0$ we can apply Corollary 7.1, with $F_k^{m_n}$ denoting the CDF of $\frac{\sum_{j=m_n+1}^{m_n+k} Y_j^{m_n} - kH(X_1|X_0,...,X_{-m_n})}{\sqrt{\operatorname{var}(\sum_{j=m_n+1}^{m_n+k} Y_j^{m_n})}}$.

By employing the triangle inequality we have

$$\sup_{t} \left| P\left(\frac{\sqrt{n - m_{n}}}{\sigma} \left(\sum_{j=1}^{l^{(m_{n}+1)}} f_{j}^{m_{n},n} \log Q_{j}^{m_{n}} + H(X_{0}|X_{-1}, \dots, X_{-m_{n}}) \right) \leqslant t \right) - \Phi(t) \right| \\
\leqslant \sup_{t} \left| \Phi(t) - \Phi\left(t \frac{\sqrt{n - m_{n}}\sigma}{\sqrt{\operatorname{var}\left(\sum_{j=m_{n}+1}^{n} Y_{j}^{m_{n}}\right)}} \right) \right| + \sup_{t} |F_{n-m_{n}}^{m_{n}}(t) - \Phi(t)|. \tag{27}$$

According to Corollary 7.1, $\sup_t |F_{n-m_n}^{m_n}(t) - \Phi(t)| = o((n-m_n)^{-\frac{\delta(\beta-1)}{4(\beta+1)}}) = o(n^{-\frac{\delta(\beta-1)}{4(\beta+1)}})$. Moreover we have proved that,

$$\frac{\sqrt{n - m_n \sigma}}{\sqrt{\operatorname{var}(\sum_{j=1}^{n - m_n} Y_j^{m_n})}} \to 1.$$

By employing the mean value theorem we can then show that:

$$\sup_{t} \left| \Phi(t) - \Phi(t \frac{\sqrt{n - m_n} \sigma}{\sqrt{\operatorname{var}(\sum_{j=1}^{n - m_n} Y_j^{m_n})}}) \right| \to 0,$$

as $n \to \infty$. If we use this in (27) we conclude that $\liminf_{n \to \infty} P(\sqrt{n}(\frac{1}{n}K(X_{1:n}) - H(X_0|X_{-1}, \dots, X_{-\infty})) \le t) \ge \Phi(t\sigma)$, which is one side of what we had to prove.

3.5.2 Upper bound

Proof. Define $\delta_n \triangleq n^{-\frac{2}{3}}$.

$$P\left(\frac{K(X_{1:n})}{n} < -\frac{\log(P(X_{1:n}))}{n} - \delta_n\right)$$

$$\leq P\left(\frac{K(X_{1:n})}{n} < -\frac{\log(P(X_{1:n}))}{n} - \delta_n, \frac{K(X_{1:n})}{n} < x\right) + P\left(\frac{K(X_{1:n})}{n} > x\right).$$
 (28)

Our goal is to show that under a proper choice of x, both probabilities on the right converge to zero as $n \to \infty$. First note that

$$P\left(\frac{K(X_{1:n})}{n} < -\frac{\log(P(X_{1:n}))}{n} - \delta_n, \frac{K(X_{1:n})}{n} < x\right)$$

$$\leq \sum_{i=1}^{nx} \sum_{\substack{v \text{ as } K(v)=i, \\ P(v) < 2^{-(i+n\delta_n)}}} P(X_{1:n} = v) \leq \sum_{i=1}^{nx} \sum_{\substack{v \text{ as } K(v)=i, \\ P(v) < 2^{-(i+n\delta_n)}}} 2^{\log P(v)}$$

$$\leq \sum_{i=1}^{nx} \sum_{\substack{v \text{ as } K(v)=i, \\ P(v) < 2^{-(i+n\delta_n)}}} 2^{-(i+n\delta_n)} \leq \sum_{i=1}^{nx} 2^{i} 2^{-(i+n\delta_n)} \leq nx 2^{-n\delta_n} \to 0. \tag{29}$$

Furthermore, if we choose $x = \frac{3}{2}H(X_0|X_{-1},\ldots,X_{-\infty})$, we have

$$P\left(\frac{K(X_{1:n})}{n} > \frac{3}{2}H(X_0|X_{-1},\dots,X_{-\infty})\right) \to 0.$$
(30)

as $n \to \infty$. Hence, by combining (28), (29), and (30), we have

$$P\left(\frac{K(X_{1:n})}{n} < -\frac{\log(P(X_{1:n}))}{n} - \delta_n\right) \to 0.$$
(31)

On the other hand, $\forall t$,

$$P\left(\sqrt{n}(\frac{1}{n}K(X_{1:n}) - H(X_0|X_{-1}, \dots, X_{-\infty})) \le t\right)$$

$$\leq P\left(\frac{K(X_{1:n})}{n} < -\frac{\log(P(X_{1:n}))}{n} - \delta_n\right) + P\left(\sqrt{n}(-\frac{\log(P(X_{1:n}))}{n} - \delta_n - H(X_0|X_{-1}, \dots, X_{-\infty})) \le t\right).$$

Note two main points about our last expression: (i) According to (31) the first term goes to zero as $n \to \infty$. (ii) We would like to characterize the limiting distribution of $(-\frac{\log(P(X_{1:n}))}{\sqrt{n}} - \sqrt{n}H(X_0|X_{-1},\ldots,X_{-\infty}))$. We rewrite this expression in the following way:

$$-\frac{\log(P(X_{1:n}))}{\sqrt{n}} - \sqrt{n}\delta_n - \sqrt{n}H(X_0|X_{-1}, \dots, X_{-\infty}))$$

$$= -\frac{\log(P(X_{1:n}))}{\sqrt{n}} - \sqrt{n}\delta_n + \frac{\sum_{j=m_n+1}^n \log(P(X_j|X_{j-1:j-m_n}))}{\sqrt{n}}$$

$$-\frac{\sum_{j=m_n+1}^n \log(P(X_j|X_{j-1:j-m_n}))}{\sqrt{n}} + \sqrt{n}H(X_0|X_{-1}, \dots, X_{-\infty})). \tag{32}$$

where $m_n = \frac{\frac{1}{2} - \epsilon}{\log l} \log n$, where $\frac{1}{2} - \frac{1}{2C} > \epsilon > 0$. Note that if we prove

$$-\frac{\log(P(X_{1:n}))}{\sqrt{n}} - \sqrt{n}\delta_n + \frac{\sum_{j=m_n+1}^n \log(P(X_j|X_{j-1:j-m_n}))}{\sqrt{n}} \stackrel{p}{\to} 0, \tag{33}$$

and

$$-\frac{\sum_{j=m_n+1}^n \log(P(X_j|X_{j-1:j-m_n}))}{\sqrt{n}} + \sqrt{n}H(X_0|X_{-1},\dots,X_{-\infty})) \stackrel{d}{\to} N(0,\sigma^2), \tag{34}$$

then by Slutsky's theorem we conclude that

$$P\left(\sqrt{n}\left(-\frac{\log(P(X_{1:n}))}{n} - \delta_n - H(X_0|X_{-1},\dots,X_{-\infty})\right) \leqslant t\right) \to \Phi(\sigma t).$$

Proof of (34) is the same as the proof we presented in the last section. To prove (33) first note that $\sqrt{n}\delta_n \to 0$. Furthermore,

$$\mathbb{E}\left(\left|\frac{\log(P(X_{1:n}))}{\sqrt{n}} - \frac{\sum_{j=2m_n}^{n} \log(P(X_{j}|X_{j-1:j-m_n-1}))}{\sqrt{n}}\right|\right) \\
\leq \mathbb{E}\left(\left|\frac{\log(P(X_{1:m_n}))}{\sqrt{n}} + \frac{\sum_{j=m_n+1}^{n} \log(P(X_{j}|X_{1:j}))}{\sqrt{n}} - \frac{\sum_{j=m_n+1}^{n} \log(P(X_{j}|X_{j-1:j-m_n}))}{\sqrt{n}}\right|\right) \\
\leq -\mathbb{E}\left(\frac{\log(P(X_{1:m_n}))}{\sqrt{n}}\right) + \frac{1}{\sqrt{n}} \sum_{j=m_n+1}^{n} \mathbb{E}(|\log(P(X_{j}|X_{1:j-1})) - \log(P(X_{j}|X_{j-m_n:j-1}))|) \\
\leq -\mathbb{E}\left(\frac{\log(P(X_{1:m_n}))}{\sqrt{n}}\right) + \frac{1}{\sqrt{n}} \sum_{j=m_n+1}^{n} \mathbb{E}(|\log(P(X_{j}|X_{1:j-1})) - \log(P(X_{j}|X_{-\infty:j-1}))|) \\
+ \frac{1}{\sqrt{n}} \sum_{j=m_n+1}^{n} \mathbb{E}(|\log(P(X_{j}|X_{-\infty:j-1})) - \log(P(X_{j}|X_{j-m_n:j-1}))|) \\
\leq -\mathbb{E}\left(\frac{\log(P(X_{1:m_n}))}{\sqrt{n}}\right) + \frac{1}{\sqrt{n}} \sum_{j=m_n+1}^{n} (\nu_{\delta}(j))^{\frac{1+\delta}{2+\delta}} + \frac{n-m_n}{\sqrt{n}} (\nu_{\delta}(m_n))^{\frac{1+\delta}{2+\delta}} \to 0, \quad (35)$$

as $n \to \infty$. Hence, $\forall t \lim \sup_{n \to \infty} P(\sqrt{n}(\frac{1}{n}K(X_{1:n}) - H(X_0|X_{-1}, \dots, X_{-\infty})) \leqslant t) \leqslant \Phi(t\sigma)$.

3.6 Proof of Theorem 4

Before we go to the details of the proof we will review the main ideas. We are going to use the upper and lower bounds on the Kolmogorov complexity, derived in the proof of Theorem 3 to get inequality 4. For each bound we will obtain concentration-inequalities and combine them to obtain a concentration result for the Kolmogorov complexity. We use the concentration inequality presented in Lemma 9. Note that we use the notations defined in (6) and (7). Define

$$g(X_{1:n}) \triangleq (n-m) \sum_{i=1}^{l^{m+1}} f_j^{m,n} \log Q_j^m.$$

We would like to use Lemma 9 to show that $g(X_{1:n})$ concentrates. Toward this goal we need to do the following two steps: (i) Calculate an upper bound for $\Delta_n = ||H_n||_{\infty}$, where H_n is the $n \times n$ matrix with elements

$$\forall (i,j), \ H_{n,\{i,j\}} = \begin{cases} 1, \ \text{if i=j} \\ \Phi_{i,j}', \ \text{if i} < j \\ 0 \ \text{otherwise} \end{cases}.$$

(ii) $g(X_{1:n})$ is a 1-Lipschitz for the Hamming-distance.

To show inequality 3 we would also to use Lemma 9. Toward this goal we also need to do the following two steps: (i) Prove that $X_{1:n} \to \frac{K(X_{1:n})}{n}$ is a Lipschitz function for the Hamming-distance. (ii) Calculate an upper-bound for $\mathbb{E}(\frac{K(X_{1:n})}{n}) - H(X_1|X_{0:-m+1})$. With this summary we now discuss the details of the proof.

First, we bound Δ_n . For every (j, n) define

$$A_{j,n}(x_{0:i}) \triangleq \{x_{j:n} \in A^{n-j+1}, \text{ such that } P(X_{j:n} = x_{j:n} | X_{0:i} = x_{0:i}) - P(X_{j:n} = x_{j:n}) \ge 0\}.$$

Then,

$$\begin{split} \sup_{x_{0:i}} & \| P(X_{j:n} \in \cdot | X_{0:i} = x_{0:i}) - P(X_{j:n} \in \cdot) \|_{TV} \\ \leqslant \sup_{x_{0:i}} & \sum_{x_{j:n} \in A^{n-j}} | P(X_{j:n} = x_{j:n} | X_{0:i} = x_{0:i}) - P(X_{j:n} = x_{j:n}) | \\ \leqslant \sup_{x_{0:i}} & \sum_{x_{j:n} \in A_{n,j}(x_{0:i})} P(X_{j:n} = x_{j:n} | X_{0:i} = x_{0:i}) - P(X_{j:n} = x_{j:n}) \\ & + \sum_{x_{j:n} \in A_{n,j}^{c}(x_{0:i})} P(X_{j:n} = x_{j:n}) - P(X_{j:n} = x_{j:n} | X_{0:i} = x_{0:i})] \\ \leqslant \sup_{x_{0:i}} & [P(X_{j:n} \in A_{n,j}(x_{0:i}) | X_{0:i} = x_{0:i}) - P(X_{j:n} \in A_{n,j}(x_{0:i})) \\ & - P(X_{j:n} \in A_{n,j}^{c}(x_{0:i}) | X_{0:i} = x_{0:i}) + P(X_{j:n} \in A_{n,j}^{c}(x_{0:i}))] \\ \leqslant & 2 \sup_{A \in \mathbb{F}^{i}_{-\infty}, B \in \mathbb{F}^{\infty}_{j}} | P(B|A) - P(A) | \leqslant 2\phi(j-i). \end{split}$$

Hence, according to the definition of the $\Phi'_{i,j}$ we have

$$\Phi'_{i,j} = \sup_{x_{0:i}, y_{0:i}} \|P(X_{j:n} \in \cdot | X_{0:i} = x_{0:i}) - P(X_{j:n} \in \cdot | X_{0:i} = y_{0:i})\|_{TV}$$

$$\leq 2 \sup_{x_{0:i}} \|P(X_{j:n} \in \cdot | X_{0:i} = x_{0:i}) - P(X_{j:n} \in \cdot)\|_{TV}$$

$$\leq 4 \sup_{A \in \mathbb{F}^{i}_{-\infty}, B \in \mathbb{F}^{\infty}_{j}} |P(B|A) - P(A)| \leq 4\phi(j-i).$$

And so we have that $\Delta_n \leq 1 + 4 \sum_{k=0}^{\infty} \phi(k) < \infty$. Moreover, according to the proof of Theorem 3, using the notations introduced in (8), we have

$$K(X_{1:n}) \leq C' + \log^*(m) + l \max_{j \leq l} K(a_j) + l^{(m+1)} \log^* n - m \log^* l - (n-m) \sum_{j=1}^{l^{m+1}} f_j^{m,n} \log Q_j^m.$$

Let $C_1(n) \triangleq C' + \log^*(m) + l \max_{j \leq l} K(a_j) + l^{(m+1)} \log^* n - m \log^* l$. Our goal it to find a concentration inequality for $(n-m) \sum_{j=1}^{l^{m+1}} f_j^{m,n} \log Q_j^m$. Toward this goal, we prove that this function is 1-Lipschitz and then use Lemma 9. Note that

$$(n-m)\sum_{j=1}^{l^{m+1}} f_j^{m,n} \log Q_j^m = \sum_{j=m+1}^n \sum_{k=1}^{l^{m+1}} I_{(X_{j-m:j}=a_k^m)} \log(Q_k^m),$$

where a_k^m is the k^{th} element of A^m . Let $x, x' \in A^n$ denote two vectors that only differ at the j^{th} -coordinate (i.e. $x_i = x_i'$, $\forall i \neq j$). Then, by the M-stability assumption of the theorem

 $|g(x) - g(x')| \leq M$ (note that g is the log-likelihood of $X_{m+1:n}$). Hence, g is M-Lipschitz for the Hamming metric. Lemma 9 implies that for every t > 0

$$P\Big((n-m)\sum_{j=1}^{l^{m+1}} f_j^{m,n} \log Q_j^m + (n-m)H(X_1|X_{0:-m+1}) \leqslant -t\Big) \leqslant 2e^{-\frac{t^2}{2nM^2\Delta^2}},$$

and

$$P\Big((n-m)\sum_{j=1}^{l^{m+1}} f_j^{m,n} \log Q_j^m + (n-m)H(X_1|X_{0:-m+1}) \geqslant t\Big) \leqslant 2e^{-\frac{t^2}{2nM^2\Delta^2}}.$$

It is straightforward to confirm that for every t, if $t - \frac{C_1(n)}{n} + \frac{m}{n}H(X_1|X_{0:-m+1}) > 0$, then

$$P\left(\frac{K(X_{1:n})}{n} - H(X_1|X_{0:-m+1}) \ge t\right)$$

$$\leq P\left(\frac{1}{n}((n-m)\sum_{j=1}^{l^{m+1}} f_j^{m,n} \log Q_j^m - [(n-m)H(X_1|X_{0:-m+1}) + mH(X_1|X_{0:-m+1})]) \ge t - \frac{C_1(n)}{n}\right)$$

$$\leq P\left(\frac{1}{n}((n-m)\sum_{j=1}^{l^{m+1}} f_j^{m,n} \log Q_j^m - (n-m)H(X_1|X_{0:-m+1})) \ge t - \frac{C_1(n)}{n} + \frac{m}{n}H(X_1|X_{0:-m+1})\right)$$

$$\leq e^{-\frac{n\left(t - \frac{C_1(n)}{n} + \frac{m}{n}H(X_1|X_{0:-m+1})\right)^2}{2M^2\Delta^2}}$$

$$\leq e^{-\frac{n\left(t - \frac{C_1(n)}{n} + \frac{m}{n}H(X_1|X_{0:-m+1})\right)^2}{2M^2\Delta^2}}$$
(36)

To prove the upper bound, first set $\delta_n = \frac{1}{n^{\frac{1}{2}+\eta}}$. Similar to the proof we presented in Section 3.5.2, we can prove that

$$P\left(\frac{K(X_{1:n})}{n} < -\frac{\log(P(X_{1:n}))}{n} - \delta_n\right) \leqslant n\zeta e^{-n^{\frac{1}{2}-\eta}}.$$

Hence, if n satisfies $t + \frac{m}{n}H(X_1|X_{0:-m+1}) > 0$, then

$$P\left(\frac{K(X_{1:n})}{n} - H(X_{1}|X_{0:-m+1}) \leq -t\right)$$

$$\leq P\left(-\frac{\log(P(X_{1:n}))}{n} - \frac{n-m}{n}H(X_{1}|X_{0:-m+1}) \leq -t + \delta_{n} + \frac{m}{n}H(X_{1}|X_{0:-m+1})\right)$$

$$+ P\left(\frac{K(X_{1:n})}{n} < -\frac{\log(P(X_{1:n}))}{n} - \delta_{n}\right)$$

$$\leq P\left(-\frac{1}{n}[(n-m)\sum_{j=1}^{l^{m+1}} f_{j}^{m,n} \log Q_{j}^{m} - (n-m)H(X_{1}|X_{0:-m+1})] \leq -t + \frac{m}{n}H(X_{1}|X_{0:-m+1})\right)$$

$$+ P\left(\frac{K(X_{1:n})}{n} < -\frac{\log(P(X_{1:n}))}{n} - \delta_{n}\right)$$

$$\leq e^{-\frac{n(t-\frac{m}{n}H(X_{1}|X_{0:-m+1}) - \delta_{n})^{2}}{2M^{2}\Delta^{2}} + n\zeta e^{-n^{\frac{1}{2}-\eta}}.$$

$$(37)$$

To obtain the second inequality we used the fact that $-\log(P(X_{1:n})) = -\log(P(X_{1:m})) - \log(P(X_{m+1:n}|X_{1:m})) \ge -\log(P(X_{m+1:n}|X_{1:m}))$. Finally, The first term in the last line is similar to (36). Hence, by combining (36) and (37) we obtain

$$P\left(\left|\frac{K(X_{1:n})}{n} - H(X_1|X_{0:-m+1})\right| \geqslant t\right) \leqslant 2e^{-\frac{n\left(t - \frac{C_1(n)}{n} - \frac{m}{n}H(X_1|X_{0:-m+1}) - \delta_n\right)^2}{2M^2\Delta^2}} + n\zeta e^{-n^{\frac{1}{2} - \eta}}.$$

Finally, note that $C_1(n) = O_n(n^{-1}\log^*(n))$, and hence if we define $\gamma_n \triangleq \frac{C_1(n)}{n} + \frac{m}{n}H(X_1|X_{0:-m+1}) + \delta_n$ and $K_1 \triangleq 2M^2\Delta^2$, then we have

$$P\left(\left|\frac{K(X_{1:n})}{n} - H(X_1|X_{-m:0})\right| \geqslant t\right) \leqslant 2e^{-\frac{n(t-\gamma_n)^2}{K_1}} + n\zeta 2^{-n^{\frac{1}{2}-\eta}},$$

where $\gamma_n = O(n^{-(\frac{1}{2} - \eta)})$.

We now want to discuss the details of the proof of inequality 3.

For $n \in \mathbb{N}$, consider the two vectors $x, x' \in A^{n^2}$ such that $d_n(x, x') \leq 1$. If $d_n(x, x') = 0$, then we can easily see that $|K(x_{1:n}) - K(x'_{1:n})| = 0$. Hence, we assume that $d_n(x, x') = 1$. Suppose that $x_i \neq x'_i$. Then we can note that if the universal machine knows $x_{1:n}$ to know $x'_{1:n}$ it only need to know i and x'_i . Therefore

$$K(x'_{1:n}) \leq K(x_{1:n}) + C' + \log^*(n) + \max_i K(a_i),$$

where C' is a constant that depends only on the universal machine.

As the previous inequality is symetric in x, x' we obtain that $x_{1:n} \to \frac{1}{n}K(x_{1:n})$ is $\frac{C' + \log^*(n) + \max_i K(a_i)}{n}$ -Lipschitz.

Lemma 9 implies that for every t > 0

$$P\Big(\left|\frac{1}{n}K(X_{1:n}) - \mathbb{E}(\frac{1}{n}K(X_{1:n}))\right| > t\Big) \leqslant 2e^{-\frac{nt^2}{2(C' + \log^*(n) + \max_i K(a_i))^2 \Delta^2}}.$$

Moreover thanks to Kraft inequality and the positivity of the Kullback-Leiller divergence we have that $\mathbb{E}(\log(\frac{P(X_{1:n})}{2^{-K(X_{1:n})}})) \geqslant 0$, hence $H(X_{1:n}) \geqslant \mathbb{E}(K(X_{1:n}))$.

Moreover we can use the upper-bound on the Kolmogorov complexity obtained in equation 8 to get that for all $m \in \mathbb{N}$

$$\mathbb{E}(K(X_{1:n})) \leq C' + \log^*(m) + l \max_{i \leq l} K(a_i) + l^{m+1} \log^* n + m \log^* l + (n-m)H(X_1|X_{0:-m+1}).$$

Hence

$$\left|\frac{1}{n}\mathbb{E}(K(X_{1:n})) - H(X_1|X_{0:-m+1})\right| \leqslant \frac{C' + \log^*(m) + l \max_{j \leqslant l} K(a_j) + l^{m+1} \log^* n + m \log^* l + mH(X_1)}{n}.$$

Therefore by defining $\gamma'(n) \triangleq \frac{C' + \log^*(m) + l \max_{j \leq l} K(a_j) + l^{m+1} \log^* n + m \log^* l + mH(X_1)}{n}$ we get that $\forall t > \gamma'(n)$

$$P\left(\left|\frac{1}{n}K(X_{1:n}) - \mathbb{E}\left(\frac{1}{n}K(X_{1:n})\right)\right| > t\right) \leqslant 2e^{-\frac{n(t-\gamma'(n))^2}{2(C'+\log^*(n)+\max_i K(a_i))^2\Delta^2}}.$$

3.6.1 Proof of Example 2.1

We first mention the following central-limit theorem for triangular arrays of martingales that will be later used in the proof.

Theorem 10. [24] Let $(S_{n,i}, F_i, 1 \le i \le k_n, n \ge 1)$ be a zero-mean, square integrable martingale array with differences $X_{n,i}$, and let η^2 be an a.s. finite random variable. Suppose that

$$\forall \epsilon > 0, \quad \sum_{i \leq k_n} \mathbb{E}(X_{n,i}^2 I_{|X_{n,i}| > \epsilon} | F_{i-1}) \xrightarrow{P} 0$$
$$\sum_{i \leq k_n} \mathbb{E}(X_{n,i}^2 | F_{i-1}) \xrightarrow{P} \eta^2.$$

Then $S_{n,k_n} \stackrel{d}{\to} Z$, where the characteristic function Z is $\mathbb{E}(e^{-\frac{1}{2}\eta^2t^2})$.

We review the roadmap of the proof. First we find an upper bound and lower-bound for the complexity of $X_{1:n}$ in terms of the $(\tau_k)_k$ and $(Y_k)_k$. Using this upper and lower bound we will prove that there is a function, f_n such that $\sqrt{n}(\frac{K(X_{1:n})}{n} - f_n(\{\tau_k\}_{k=-\infty}^{\infty}, \{Y_k\}_{k=-\infty}^{\infty})) \to 0$ almost surely. This implies that if the central-limit theorem holds, then the asymptotic distribution of $\sqrt{n}(H(X_1|X_{0:-\infty}) - f_n(\{\tau_k\}_{k=-\infty}^{\infty}, \{Y_k\}_{k=-\infty}^{\infty}))$ would also be Gaussian. We will then prove that this does not happen since there is a $\eta > 0$ such as: $n^{\frac{1}{2}-\eta}(H(X_1|X_{0:-\infty}) - f_n(\{\tau_k\}_{k=-\infty}^{\infty}, \{Y_k\}_{k=-\infty}^{\infty}))$ is not bounded in probability.

First to understand the proof we have to notice that the process $\{X_i\}_{i=-\infty}^{\infty}$ is constituted of different segments of random variables that comes from different distributions and those segments have different lengths, for example $X_{1:\tau_1-\theta}|\tau_1,\theta$ comes from a certain distribution and $X_{\tau_1-\theta+1:\tau_1-\theta+\tau_2}|\tau_1,\theta,\tau_2$ may come from another distribution. Let $\{L_i\}_i$ denote the i^{th} segments, e.g. $L_1 = X_{1:\tau_1-\theta}$. Define $l_1 \triangleq \tau_1 - \theta$, which is the length of the first segment, and for every i > 0 define

$$N_i \triangleq \max\{k : \text{ such that } l_1 + \tau_2 + \dots + \tau_k \leq i\}.$$

 N_i is maximum number of segments $\{L_k\}_k$, including the first one, that are entirely in $X_{1:i}$. Finally, define $l_{left}(i) \triangleq i - l_1 - \sum_k^{N_i} \tau_k$, which is the number of elements of $X_{1:i}$ that are not in any of the different L_k , for $k \leq N_i$.

To describe $X_{1:n}$ we may describe each segment $X_{1:l_1}$, $X_{l_1+1:l_1+\tau_2}$,..., $X_{l_1+\sum_k^{N_n}\tau_k+1:n}$. It is straightforward to confirm the following two facts: (i) if $Y_i = 1$ then the i^{th} segment can be described by the length of the segment and a constant cost, C, to indicate to the machine that it should produce an array of 0's. (ii) If $Y_i = 0$ then the i^{th} segment can be described by describing each element in that segment. Since we have $N_n + 1$ segments, it is straightforward to confirm that

$$K(X_{1:n} \mid N_n, l_1, \tau_2, \dots, \tau_{N_n}, l_{left(n)}, Y_1, \dots, Y_{N_n})$$

$$\leq C(N_n + 1) + \min(l_1, n) I_{Y_1 = 0} + \sum_{i \leq N_n} \tau_i I_{Y_i = 0} + l_{left}(n) I_{Y_{N_n + 1} = 0}.$$
(38)

Note that for the full-description of $X_{1:n}$ we should also describe the following to the universal machine: (i) $(l_1, \tau_{1:N_n}, l_{left}(n))$, (ii) $(Y_{1:N_n+1})$. Hence it is straightforward to check the following upper bound for the Kolmogorov complexity of $X_{1:n}$:

$$K(X_{1:n}) \leq (N_n + 1)(1 + C + \log^*(n)) + \min(l_1, n)I_{Y_1=0} + \sum_{i \leq N_n} \tau_i I_{Y_i=0} + l_{left}(n)I_{Y_{N_n+1}=0}.$$
(39)

Before we proceed to simplify the above upper bound, let me find a lower bound for the Kolmogorov's complexity of $X_{1:n}$ as well. Define the vector V_n in the following way: take all the segments of $X_{1:n-l_{left}(n)}$ that are coming from \tilde{Y} and concatenate them to obtain the vector V_n . Note that if the Universal computer has access to $X_{1:n}$, then it only requires the following information to construct V_n : the values of Y_1, \ldots, Y_{N_n} and $l_1, \tau_2, \tau_3, \ldots, \tau_{N_n}$. Hence, it is straightforward to show that

$$K(V_n) \le K(X_{1:n}) + (N_n + 1)(1 + \log^* n + C).$$
 (40)

It is intuitively clear that since V_n has iid Bern(1/2) elements its Kolmogorov complexity should be concentrated around its length. Below we prove this intuition:

Lemma 11. Let l_n denote the length of V_n . If $\delta_n = n^{-2/3}$, then

$$\mathbb{P}(K(V_n) \leqslant l_n - n\delta_n | l_n) \to 0,$$

as $n \to \infty$.

Proof. First for a certain l_n we can describe V_n by :

- (i) Describe the length of the sequence: l_n , with a cost of at most $\log^*(l_n) + C$.
- (ii) Describe each of the l_n elements of the sequence, with a cost of at most l_n .
- (iii) Telling it how to build the sequence, with a cost of C', where C' is a constant that depends only on the universal machine.

Hence:

$$K(V_n) \leqslant C' + \log^*(l_n) + l_n.$$

And so: $P(K(V_n) \ge 2l_n|l_n) \to 0$. Then, we have

$$P(K(V_n) \leq -\log(P(V_n|l_n)) - n\delta_n|l_n)$$

$$\leq P(K(V_n) \geq 2l_n|l_n) + P(K(V_n) \leq 2l_n, K(V_n) \leq -\log(P(V_n|l_n)) - n\delta_n|l_n)$$

$$\leq P(K(V_n) \geq 2l_n|l_n) + \sum_{i=1}^{2l_n} \sum_{\substack{n \text{ as } K(n) = i}} 2^{\log(P(V|l_n))} \leq P(K(V_n) \geq 2l_n|l_n) + 2l_n 2^{n\delta_n} \to 0.$$

Please note that to pass from the second-line to the third we have used Lemma 6. Finally,

$$\mathbb{P}(K(V_n) \leqslant l_n - n\delta_n | l_n) \to 0.$$

Indeed knowing l_n , V_n is a sequence of iid bernouilli($\frac{1}{2}$) and so: $-\log(P(V_n)|l_n) = l_n$

We should note that: $l_n = (l_1 \wedge n)I_{Y_1=0} + \sum_{i \leq N_n} \tau_i I_{Y_i=0} + l_{left}(n)I_{Y_{N_n+1}=0}$. Combing (39), (40), and Lemma 11 we obtain the following upper and lower bounds for $K(X_{1:n})$:

$$K(X_{1:n}) \leq (N_n + 1)(1 + C + \log^*(n)) + l_n.$$

 $K(X_{1:n}) \geq l_n - (N_n + 1)(1 + C + \log^*(n)) - n\delta_n,$

$$(41)$$

where the lower bound holds with probability converging to 1. Our next goal is to show that with probability converging to one

$$\frac{1}{\sqrt{n}} \left(K(X_{1:n}) - \min(l_1, n) I_{Y_1 = 0} - \sum_{i \le N_n} \tau_i I_{Y_i = 0} - l_{left}(n) I_{Y_{N_n + 1} = 0} \right) \to 0.$$
 (42)

It is straightforward to confirm that $\frac{n\delta_n}{\sqrt{n}} \to 0$. Hence, we only have to prove that $N_n(\log^*(n) +$ $(C+1)/\sqrt{n} \to 0$ (which is going to be true if $N_n \log^*(n)/\sqrt{n} \to 0$). Toward this goal define $S_n \triangleq \sum_{i=2}^n \tau_i$. Since S_n is a sum of iid variables, it is straightforward to confirm that

$$\frac{S_n}{S_n^{1+u}} \xrightarrow{a.s} 0,$$

 $0 < u < \frac{\frac{1}{2} - \frac{\epsilon}{4}}{\frac{1}{2} - \epsilon} - 1$. Moreover as $\frac{S_n}{S_n^{1+u}} = \frac{1}{S_n^u} \leqslant \frac{1}{\tau_1^u} \in L^1$, by dominated convergence theorem we also obtain the L^1 convergence. Then we have that by exchangeability of the $(\tau_i)_{i \leqslant n} |S_n|$ that

$$\mathbb{E}\left(\frac{S_n}{S_n^{1+u}}\right) = \mathbb{E}\left(\mathbb{E}\left(\frac{\sum_{i=1}^n \tau_i}{S_n^{1+u}}|S_n\right)\right) = \mathbb{E}\left(\frac{n\tau_1}{S_n^{1+u}}\right) \xrightarrow{P} 0.$$

Therefore $\frac{n\tau_1}{(S_n-\tau_1)^{1+u}} \xrightarrow{L_1} 0$, which implies that

$$\mathbb{E}\left(\frac{n\tau_1}{(S_n - \tau_1)^{1+u}}\right) \geqslant nP\left(\frac{\tau_1}{(S_n - \tau_1)^{1+u}} \geqslant 1\right)$$

$$= n\mathbb{E}\left(P(\tau_1 > (S_n - \tau_1)^{1+u}|S_n - \tau_1)\right)$$

$$\geqslant K'n\mathbb{E}\left(S_n^{-(1+u)(\frac{1}{2} + \epsilon)}\right) \to 0,$$

where we have used the fact that there is a constant K' such that any fixed b, $P(|\tau_1| > b) \ge K'b^{-(\frac{1}{2}-\epsilon)}$

Hence.

$$\mathbb{E}(S_n^{-(\frac{1}{2}-\epsilon)(1+u)}) = o(\frac{1}{n}). \tag{43}$$

By employing Markov inequality we obtain $S_n^{-1} = o_p(n^{-(1+u)^{-1}(\frac{1}{2}-\epsilon)^{-1}})$. Note that if we have $m \triangleq |n^{\frac{1}{2}-\frac{\epsilon}{4}}|$.

$$\begin{split} & \mathbb{P}(N_n > n^{1/2 - \epsilon/4}) \leqslant \mathbb{P}(S_{\lfloor n^{\frac{1}{2} - \frac{\epsilon}{4}} \rfloor} \leqslant n) \\ & \leqslant \mathbb{P}(S_m \leqslant m^{(\frac{1}{2} - \frac{\epsilon}{4})^{-1}}) \leqslant \mathbb{P}(1 \leqslant S_m^{-1}(m^{(\frac{1}{2} - \frac{\epsilon}{4})^{-1}})) \to 0 \end{split}$$

Where the last equation comes from Equation 43 and $(\frac{1}{2} - \frac{\epsilon}{4})^{-1} < \frac{1}{(\frac{1}{2} - \epsilon)(1+u)}$. Hence, it is straightforward to conclude that

$$\frac{N_n \log^*(n)}{\sqrt{n}} \to 0. \tag{44}$$

This completes the proof of (42).

It is straightforward to prove that the entropy rate of this process is 1/2. Hence, we would like to show that

$$\sqrt{n}\left(\frac{K(X_1,X_2,\ldots,X_n)}{n}-\frac{1}{2}\right),$$

is $\omega(1)$. Suppose that this is not the case, then by using Prohorov's theorem the sequence is tight and the sequence $\sqrt{n}(\frac{K(X_1,X_2,\dots,X_n)}{n}-\frac{1}{2})$ will have a subsequence that converges almost surely. To simplify the notation, instead of working with the convergent subsequence we assume that the entire sequence converges in distribution. Since

$$\sqrt{n} \left(\frac{\min(l_1, n) I_{Y_1 = 0} - \sum_{i \leq N_n} \tau_i I_{Y_i = 0} - l_{left}(n) I_{Y_{N_n + 1} = 0}}{n} - \frac{1}{2} \right)$$

$$= \sqrt{n} \left(\frac{K(X_1, X_2, \dots, X_n)}{n} - \frac{1}{2} \right)$$

$$+ \frac{1}{\sqrt{n}} \left(K(X_{1:n}) - \min(l_1, n) I_{Y_1 = 0} - \sum_{i \leq N_n} \tau_i I_{Y_i = 0} - l_{left}(n) I_{Y_{N_n + 1} = 0} \right)$$
(45)

and according to (42):

$$\frac{1}{\sqrt{n}} \left(K(X_{1:n}) - \min(l_1, n) I_{Y_1 = 0} - \sum_{i \le N_n} \tau_i I_{Y_i = 0} - l_{left}(n) I_{Y_{N_n + 1} = 0} \right) \xrightarrow{P} 0,$$

and we have assumed that $\sqrt{n}(\frac{K(X_1,X_2,...,X_n)}{n}-\frac{1}{2})$ converges in distribution, we can use Slutsky's theorem and claim that $\sqrt{n}\left(\frac{\min(l_1,n)I_{Y_1=0}-\sum_{i\leqslant N_n}\tau_iI_{Y_i=0}+l_{left}(n)I_{Y_{N_n+1}=0}}{n}-\frac{1}{2}\right)$ converges in distribution. Note that $l_1I_{Y_1=0}<\tau_1$ and $l_{left}(n)I_{Y_{N_n+1}=0}<\tau_{N_n+1}$. Therefore,

$$\frac{\min(l_1, n)I_{Y_1=0} + l_{left}(n)I_{Y_{N_n+1}=0}}{\sqrt{n}} \xrightarrow{a.s.} 0.$$

Hence

$$\sqrt{n} \left(\frac{\min(l_1, n) I_{Y_1=0} + l_{left}(n) I_{Y_{N_n+1}=0}}{n} - \frac{1}{2} \right) \to 0,$$

and our analyses reduces to the analysis of $\sqrt{n} \left(\frac{\sum_{i \leq N_n} \tau_i I_{Y_i=0}}{n} - \frac{1}{2} \right)$. Note that

$$\sqrt{n} \left(\mathbb{E} \left(\frac{\sum_{i \leq N_n} \tau_i I_{Y_i = 0}}{n} | N_n, \tau_1, \tau_2, \dots, \tau_{N_n} \right) - 0.5 \right) \\
= \sqrt{n} \left(\mathbb{E} \left(\frac{\sum_{i \leq N_n} \tau_i I_{Y_i = 0}}{n} | N_n, \tau_1, \tau_2, \dots, \tau_{N_n} \right) - 0.5 \right) \\
= \sqrt{n} \left(\frac{1}{2} \frac{\sum_{i \leq N_n} \tau_i}{n} - \frac{1}{2} \right) = \frac{\sqrt{n}}{2} \frac{\sum_{i \leq N_n} \tau_i - n}{n} = \frac{l_1 + l_{left}(n)}{2\sqrt{n}} \xrightarrow{a.s.} 0.$$

Hence we discuss the limiting distribution of the following quantity:

$$\sqrt{n} \left(\frac{\sum_{i \leqslant N_n} \tau_i I_{Y_i = 0}}{n} - \mathbb{E} \left(\frac{\sum_{i \leqslant N_n} \tau_i I_{Y_i = 0}}{n} | N_n, \tau_1, \tau_2, \dots, \tau_{N_n} \right) \right).$$

Toward that goal we will first introduce the following sigma-fields:

$$F_l \triangleq \sigma(\tau_i, I_{i \leq N_l} I_{Y_i = 0}, i \in \mathbb{N}),$$

and the processes

$$Y_l^n \triangleq \frac{1}{\sqrt{\sum_i \tau_i^2 I_{i \leqslant N_n}}} \sum_i \tau_i I_{i \leqslant N_l} (I_{Y_i = 0} - \frac{1}{2}).$$

It is straightforward to see that $((Y_l^n, F_l)_l)_n$ is a triangular array of martingales. The corresponding martingale differences are given by

$$X_{n,i} \triangleq \frac{1}{\sqrt{\sum_{i} \tau_{i}^{2} I_{i \leq N_{n}}}} \sum_{i=1}^{\infty} \tau_{j} I_{N_{i-1} < j \leq N_{i}} (I_{Y_{i}=0} - \frac{1}{2}).$$

We would now like to use Theorem 10. It is straightforward to check that

$$\frac{1}{\sum_i \tau_i^2 I_{i \leqslant N_n}} \sum_{i=1}^n \mathbb{E} \big((\sum_j \tau_j I_{N_{i-1} < j \leqslant N_i} \big(I_{Y_i = 0} - \frac{1}{2} \big))^2 \big| F_{i-1} \big) = \frac{1}{4}.$$

Furthermore, we have to prove the following claim:

$$\forall \epsilon > 0, \ \sum_{i \leqslant n} \mathbb{E}\left(\left|\frac{\sum_{j} \tau_{j} I_{N_{i-1} < j \leqslant N_{i}} (I_{Y_{i}=0} - \frac{1}{2})}{\sqrt{\sum_{i} \tau_{i}^{2} I_{i \leqslant N_{n}}}}\right|^{2} I_{\left|\frac{\sum_{j} \tau_{j} I_{N_{i-1} < j \leqslant N_{i}} (I_{Y_{i}=0} - \frac{1}{2})}{\sqrt{\sum_{i} \tau_{i}^{2} I_{i \leqslant N_{n}}}}\right| > \epsilon} |F_{i-1}) \xrightarrow{P} 0.$$
 (46)

Toward this goal, note that

$$\begin{split} & \sum_{i \leqslant n} \mathbb{E} \left(\Big| \frac{\sum_{j} \tau_{j} I_{N_{i-1} < j \leqslant N_{i}} (I_{Y_{i} = 0} - \frac{1}{2})}{\sqrt{\sum_{i} \tau_{i}^{2} I_{i \leqslant N_{n}}}} \Big|^{2} I_{\Big| \frac{\sum_{j} \tau_{j} I_{N_{i-1} < j \leqslant N_{i}} (I_{Y_{i} = 0} - \frac{1}{2})}{\sqrt{\sum_{i} \tau_{i}^{2} I_{i \leqslant N_{n}}}} \Big| > \epsilon \right| F_{i-1} \right) \\ & \stackrel{(a)}{=} \sum_{i \leqslant n} \frac{1}{\sum_{i} \tau_{i}^{2} I_{i \leqslant N_{n}}} \mathbb{E} (\Big| \sum_{j} \tau_{j} I_{N_{i-1} < j \leqslant N_{i}} (I_{Y_{i} = 0} - \frac{1}{2}) \Big|^{2} I_{\Big| \frac{\sum_{j} \tau_{j} I_{N_{i-1} < j \leqslant N_{i}} (I_{Y_{i} = 0} - \frac{1}{2})}{\sqrt{\sum_{i} \tau_{i}^{2} I_{i \leqslant N_{n}}}} \Big| > \epsilon \Big| F_{i-1}) \\ & \stackrel{(b)}{\leqslant} \sum_{i \leqslant N_{n}} \frac{1}{\sum_{i} \tau_{i}^{2} I_{i \leqslant N_{n}}} \tau_{i}^{2} P (\Big| \frac{\sum_{j} \tau_{j} I_{N_{i-1} < j \leqslant N_{i}} (I_{Y_{i} = 0} - \frac{1}{2})}{\sqrt{\sum_{i} \tau_{i}^{2} I_{i \leqslant N_{n}}}} \Big| > \epsilon \Big| F_{i-1}) \\ & \stackrel{\leqslant}{\leqslant} \sum_{i \leqslant N_{n}} \frac{1}{\sum_{i} \tau_{i}^{2} I_{i \leqslant N_{n}}} \tau_{i}^{2} I_{\frac{\tau_{i}^{2}}{\sum_{i} \tau_{i}^{2} I_{i \leqslant N_{n}}} > \epsilon^{2} \leqslant \max_{i} I_{\frac{\tau_{i}^{2}}{\sum_{i} \tau_{i}^{2} I_{i \leqslant N_{n}}}} \geqslant \epsilon^{2} \leqslant \max_{i} I_{\frac{\tau_{i}^{2}}{\sum_{i} \tau_{i}^{2} I_{i \leqslant N_{n}}} \geqslant \epsilon^{2}} \leqslant \max_{i} I_{\frac{\tau_{i}^{2}}{\sum_{i} \tau_{i}^{2} I_{i \leqslant N_{n}}} \geqslant \epsilon^{2}} \leqslant \max_{i} I_{\frac{\tau_{i}^{2}}{\sum_{i} \tau_{i}^{2} I_{i \leqslant N_{n}}} \geqslant \epsilon^{2}} \leqslant \max_{i} I_{\frac{\tau_{i}^{2}}{\sum_{i} \tau_{i}^{2} I_{i \leqslant N_{n}}} \geqslant \epsilon^{2}} \leqslant \max_{i} I_{\frac{\tau_{i}^{2}}{\sum_{i} \tau_{i}^{2} I_{i \leqslant N_{n}}} \geqslant \epsilon^{2}} \leqslant \max_{i} I_{\frac{\tau_{i}^{2}}{\sum_{i} \tau_{i}^{2} I_{i \leqslant N_{n}}} \geqslant \epsilon^{2}} \leqslant \max_{i} I_{\frac{\tau_{i}^{2}}{\sum_{i} \tau_{i}^{2} I_{i \leqslant N_{n}}} \geqslant \epsilon^{2}} \leqslant \max_{i} I_{\frac{\tau_{i}^{2}}{\sum_{i} \tau_{i}^{2} I_{i \leqslant N_{n}}} \geqslant \epsilon^{2}} \leqslant \max_{i} I_{\frac{\tau_{i}^{2}}{\sum_{i} \tau_{i}^{2} I_{i \leqslant N_{n}}} \geqslant \epsilon^{2}} \leqslant \max_{i} I_{\frac{\tau_{i}^{2}}{\sum_{i} \tau_{i}^{2} I_{i \leqslant N_{n}}} \geqslant \epsilon^{2}} \leqslant \max_{i} I_{\frac{\tau_{i}^{2}}{\sum_{i} \tau_{i}^{2} I_{i \leqslant N_{n}}} \geqslant \epsilon^{2}} \leqslant \max_{i} I_{\frac{\tau_{i}^{2}}{\sum_{i} \tau_{i}^{2} I_{i \leqslant N_{n}}} \geqslant \epsilon^{2}} \leqslant \max_{i} I_{\frac{\tau_{i}^{2}}{\sum_{i} \tau_{i}^{2} I_{i \leqslant N_{n}}} \geqslant \epsilon^{2}} \leqslant \max_{i} I_{\frac{\tau_{i}^{2}}{\sum_{i} \tau_{i}^{2} I_{i \leqslant N_{n}}} \geqslant \epsilon^{2}} \leqslant \max_{i} I_{\frac{\tau_{i}^{2}}{\sum_{i} \tau_{i}^{2} I_{i \leqslant N_{n}}} \geqslant \epsilon^{2}} \leqslant \max_{i} I_{\frac{\tau_{i}^{2}}{\sum_{i} \tau_{i}^{2} I_{i \leqslant N_{n}}} \geqslant \epsilon^{2}} \leqslant \max_{i} I_{\frac{\tau_{i}^{2}}{\sum_{i} \tau_{i}^{2} I_{i \leqslant N_{n}}}} \leqslant \epsilon^{2}}$$

Note that to obtain Equality (a) we used the fact that τ_1, τ_2, \ldots are F_i measurable and hence so is N_n . To obtain Inequality (b) we used the fact that for a fixed i the difference between N_{i-1} and N_i is at most one and also $|I_{Y_i=0}-\frac{1}{2}|<1$. Finally, it is straightforward to see that $P(\max_i \frac{\tau_i^2}{\sum_i \tau_i^2 I_{i \leqslant N_n}} \geqslant \epsilon^2) \to 0$, which proves (46). According to Theorem Theorem 10 we have

$$\frac{1}{\sqrt{\sum_i \tau_i^2 I_{i \leqslant N_n}}} \sum_i \tau_i I_{i \leqslant N_n} (I_{Y_i = 0} - \frac{1}{2}) \xrightarrow{d} N(0, \frac{1}{4}).$$

Hence if

$$\sqrt{n} \left(\frac{\sum_{i \leqslant N_n} \tau_i I_{Y_i = 0}}{n} - \mathbb{E} \left(\frac{\sum_{i \leqslant N_n} \tau_i I_{Y_i = 0}}{n} | N_n, \tau_1, \tau_2, \dots, \tau_{N_n} \right) \right)$$

converges to a non-degenerate distribution we need: $\sum_i \tau_i^2 I_{i\leqslant N_n} = \Theta(n)$. However, by Cauchy-Swartz we can easily see that: $\sum_i \tau_i^2 I_{i\leqslant N_n} \geqslant \frac{1}{N_n} (\sum_{i\leqslant N_n} \tau_i)^2 = \frac{1}{N_n} (n-l_1-left(n))^2$. Therefore $\frac{\sum_i \tau_i^2 I_{i\leqslant N_n}}{n} \geqslant \frac{\frac{1}{N_n} (n-l_1-left(n))^2}{N_n \times n} \to \infty$. This contradiction proves that the speed of convergence is slower than $n^{-\frac{1}{2}}$.

References

- [1] Ray J Solomonoff. Complexity-based induction systems: comparisons and convergence theorems. Information Theory, IEEE Transactions on, 24(4):422–432, 1978.
- [2] David Leigh Donoho. The Kolmogorov sampler. Department of Statistics, Stanford University, 2002.
- [3] Shirin Jalali and Arian Maleki. Minimum complexity pursuit. In Communication, Control, and Computing (Allerton), 2011 49th Annual Allerton Conference on, pages 1764-1770. IEEE, 2011.
- [4] Andrew R Barron and Thomas M Cover. Minimum complexity density estimation. Information Theory, IEEE Transactions on, 37(4):1034–1054, 1991.
- [5] Alexander K Zvonkin and Leonid A Levin. The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms. Russian Mathematical Surveys, 25(6):83–124, 1970.

- [6] Ray J Solomonoff. A formal theory of inductive inference. part i. *Information and control*, 7(1):1–22, 1964.
- [7] Ray J Solomonoff. A formal theory of inductive inference. part ii. *Information and control*, 7(2):224–254, 1964.
- [8] Andrei N Kolmogorov. Three approaches to the quantitative definition of information'. *Problems of information transmission*, 1(1):1–7, 1965.
- [9] Gregory J Chaitin. On the simplicity and speed of programs for computing infinite sets of natural numbers. *Journal of the ACM (JACM)*, 16(3):407–422, 1969.
- [10] Gregory J Chaitin. On the length of programs for computing finite binary sequences. *Journal* of the ACM (JACM), 13(4):547–569, 1966.
- [11] Gregory J Chaitin. A theory of program size formally identical to information theory. Journal of the ACM (JACM), 22(3):329–340, 1975.
- [12] Gregory J Chaitin. Information-theoretic limitations of formal systems. *Journal of the ACM* (*JACM*), 21(3):403–424, 1974.
- [13] David G Willis. Computational complexity and probability constructions. *Journal of the ACM (JACM)*, 17(2):241–259, 1970.
- [14] Juris Hartmanis. Generalized kolmogorov complexity and the structure of feasible computations. In *Foundations of Computer Science*, 1983., 24th Annual Symposium on, pages 439–445. IEEE, 1983.
- [15] Michael Sipser. A complexity theoretic approach to randomness. In *Proceedings of the fifteenth annual ACM symposium on Theory of computing*, pages 330–335. ACM, 1983.
- [16] Vladimir V V'yugin. Ergodic theorems for individual random sequences. *Theoretical Computer Science*, 207(2):343–361, 1998.
- [17] AA Brudno. Entropy and the complexity of the trajectories of a dynamic system. *Trudy Moskovskogo Matematicheskogo Obshchestva*, 44:124–149, 1982.
- [18] Thomas M Cover and Thomas Joy A. *Elements of Information Theory*. John Wiley & Sons, 2012.
- [19] J. Sunklodas. Approximation of distributions of sums of weakly dependent random variables by the normal distribution. *Lithuanian Mathematical Journal*, 27(4):359–368, 10 1987.
- [20] I.A Ibragimov. Independent and stationary sequences of random variables. 1971.
- [21] L.Kontorovich & K. Ramanan. Concentration inequalities for dependent random variables via the martingale method. *Annals of Probility*, 2008.
- [22] Randal Douc and Catherine Matias. Asymptotics of the maximum likelihood estimator for general hidden markov models. *Bernouilli*, 7(3):381–420, 2001.
- [23] Igal Sason. On reverse pinkser inequalities.
- [24] Peter Hall and Christopher C Heyde. Martingale limit theory and its application. Academic press., 2014.