ALEA, Lat. Am. J. Probab. Math. Stat. **18**, 887–905 (2021)
DOI: 10.30757/ALEA.v18-32

Rényi entropy and pattern matching for run-length encoded sequences

Jérôme Rousseau

Departamento de Matemática, Faculdade de Ciências da Universidade do Porto, Rua do Campo Alegre, 687, 4169-007 Porto, Portugal

Departamento de Matemática, Universidade Federal da Bahia, Av. Ademar de Barros s/n, 40170-110 Salvador, Brazil
E-mail address: jerome.rousseau@ufba.br
URL: http://www.sd.mat.ufba.br/∼jerome.rousseau

Abstract. In this note, we studied the asymptotic behaviour of the length of the longest common substring for run-length encoded sequences. When the original sequences are generated by an α-mixing process with exponential decay (or ψ-mixing with polynomial decay), we proved that this length grows logarithmically with a coefficient depending on the Rényi entropy of the pushforward measure. For Bernoulli processes and Markov chains, this coefficient is computed explicitly.

1. Introduction

Since Big Data seems to be the trending field of (at least) this decade, data compression algorithms have become a fundamental tool for data storage and are in the first lines of the battle between storage costs, computations costs and delays in data availability. For an introduction to data compression we refer the reader to Sayood (2017) and to the unavoidable Lempel-Ziv algorithms (Ziv and Lempel, 1977, 1978).

For sequences with long runs of the same value, Run-Length Encoding (RLE) is a simple and efficient lossless data compression method. More precisely, for a run of the same value, the algorithm stored the value and the length of the run. For example, the following binary sequence

$$0000111000000001100111111111100000000$$

Received by the editors June 13th, 2020; accepted February 11th, 2021.
2010 *Mathematics Subject Classification.* 60F15, 60G10, 94A17, 68P30, 37A25, 60J10.
Key words and phrases. Rényi entropy, longest common substring, string matching, data compression, mixing process.
This work was partially supported by CNPq, by FCT project PTDC/MAT-PUR/28177/2017, with national funds, and by CMUP (UIDB/00144/2020), which is funded by FCT with national (MCTES) and European structural funds through the programs FEDER, under the partnership agreement PT2020.

will be compressed as

$$(0,4)(1,3)(0,8)(1,2)(0,2)(1,9)(0,8).$$

Thus, this sequence of 37 characters will be represented after compression by a sequence of 14 characters.

RLE is typically used for image compression but has also application in image analysis (Hinds et al., 1990), texture analysis of volumetric data Xu et al. (2004) and has also been used for data compression of television signals Robinson and Cherry (1967) and fax transmission (Hunter and Robinson, 1980).

Since pattern (or string) matching problems are not only highly significant in computer sciences, information theory and probability (see e.g. Kontoyiannis and Suhov, 1994; Kontoyiannis et al., 1998; Dębowski, 2018; Abadi et al., 2018; Neuhauser, 1996; Collet et al., 2009) but also in biology Waterman (1995), geology Miall (2010) and linguistics (e.g. Dębowski, 2011 and references therein) among others, algorithms to solve string matching problems for RLE strings have been developed (see e.g. Apostolico et al., 1999; Freschi and Bogliolo, 2004; Hooshmand et al., 2018; Ahsan et al., 2012; Chen et al., 2010 and references within).

In this note, we will focus on a particular string matching problem: the longest common substring problem (or longest consecutive common subsequence problem). More precisely, we will concentrate on the asymptotics of the length of the longest common substring, i.e. for two sequences $\mathbf{x}$ and $\mathbf{y}$ drawn randomly from the same alphabet, the behaviour of

$$M_n(\mathbf{x},\mathbf{y}) = \max\left\{k \ : \mathbf{x}_i^{i+k-1} = \mathbf{y}_j^{j+k-1} \text{ for some } 0 \le i,j \le n-k\right\}$$

when $n \to \infty$.

In Barros et al. (2019), it was proved that for α-mixing process with exponential decay $M_n \sim \frac{2}{H_2(\mu)} \log n$ almost surely, where $H_2(\mu)$ is the Rényi entropy (see Definition 2.3) of the stationary measure μ. Similar results have been proved for more than two sequences Barros and Rousseau (2021) and for random sequences in random environment Rousseau (2021+).

In Coutinho et al. (2020), the authors wondered if the above mentioned result holds if the sequences are transformed following certain rules of modification. Thus, if f is a measurable function (called an encoder) transforming a sequence $\mathbf{x}$ into another sequence $f(\mathbf{x})$, they studied the behaviour of $M_n(f(\mathbf{x}), f(\mathbf{y}))$ and obtain a relation with the Rényi entropy of the pushforward measure $f_*\mu$.

A natural question would be to ask if we could apply the results presented in Coutinho et al. (2020) when the encoder is a compression algorithm and in particular the run-length encoder. Unfortunately, to obtain their main result, the authors needed that the encoder does not compress too much the sequences, an hypothesis which is not satisfied by the run-length encoder. Thus, we present here a different proof which allows us to prove, in Theorem 2.6, that, when f is the run-length encoder and the original sequences are generated by an α-mixing process with exponential decay (or ψ-mixing with polynomial decay), almost surely

$$M_n(f(\mathbf{x}), f(\mathbf{y})) \underset{n\to\infty}{\sim} \frac{2}{H_2(f_*\mu)} \log n.$$

We apply this result to Bernoulli processes (Example 3.1) and Markov chains (Examples 3.2 and 3.3), and, in these cases, compute explicitly $H_2(f_*\mu)$. We emphasize

that for Markov chains the computation are different whether there are two or more than two states.

Other examples of processes satisfying our mixing assumptions are Gibbs states of a Hölder-continuous potential Bowen (1975); Walters (1975), ARMA processes Mokkadem (1988), some renewal processes Abadi et al. (2015) and stationary determinantal process on the integer lattice Fan et al. (2019). We refer the reader to Doukhan (1994); Bradley (2007) for more examples and deep surveys on strong mixing conditions.

2. Longest common substring for RLE sequences

We consider a stationary stochastic process $X = (X_n)_{n\in\mathbb{N}}$ over a finite alphabet $\mathcal{A}$, with stationary measure μ. We will denote σ the left shift and, for $i \in \mathbb{N}$, $\sigma^i X = (X_{i+n})_{n\in\mathbb{N}}$. For $k \in \mathbb{N}$, we denote by $\mathcal{A}^k$ the set of cylinders or strings of length k and the length of a cylinder ω will be denoted $|\omega|$. When there is no ambiguity, cylinders of $\mathcal{A}^k$ will be denoted ω. We will use the notation x_i^{i+k-1} if we need to indicate its time of occurrence i. Moreover, $\mu(\omega)$ will denote the probability $\mu(X_i^{i+k-1} = \omega)$ (which is independent of i by stationarity).

We will be interested in some statistical properties of run-length encoded (RLE) sequences where the original sequences are generated by the stochastic process X.

Definition 2.1. Let $\mathcal{B} = \{(\alpha, k)\}_{\alpha\in\mathcal{A},k\in\mathbb{N}}$. We define the run-length encoder $f : \mathcal{A}^{\mathbb{N}} \to \mathcal{B}^{\mathbb{N}}$ by

$$f(\underbrace{\alpha_1 \dots \alpha_1}_{k_1}\underbrace{\alpha_2 \dots \alpha_2}_{k_2} \dots \underbrace{\alpha_n \dots \alpha_n}_{k_n} \dots) = (\alpha_1, k_1)(\alpha_2, k_2) \dots (\alpha_n, k_n) \dots$$

We observe that for all $i \in \mathbb{N}$, we consider that $\alpha_{i+1} \neq \alpha_i$.

We will focus our analysis on the length of the longest common substring of RLE sequences:

Definition 2.2. Given two sequences $\mathbf{x}, \mathbf{y}$, we define the n-length of the longest common substring by

$$M_n(\mathbf{x}, \mathbf{y}) = \max\left\{k \ : \mathbf{x}_i^{i+k-1} = \mathbf{y}_j^{j+k-1} \text{ for some } 0 \le i, j \le n-k\right\}.$$

and we will study the behaviour of the n-length of the longest common substring of the RLE sequences $f(\mathbf{x}), f(\mathbf{y})$

$$M_n^{RLE}(\mathbf{x}, \mathbf{y}) := M_n(f(\mathbf{x}), f(\mathbf{y})).$$

We will prove that M_n^{RLE} is linked with the Rényi entropy of the pushforward measure $f_*\mu$. We recall that $f_*\mu(.) = \mu(f^{-1}(.))$ and we observe that $f_*\mu$ is the law of the stochastic process $f(X)$ but is in general not stationary. We give now the definition of Rényi entropy:

Definition 2.3. For $k > 1$, the lower and upper Rényi entropies of order k of a measure P are defined as

$$\underline{H}_k(P) = -\varliminf_{n\to\infty} \frac{1}{(k-1)n} \log \sum_\omega P(\omega)^k \quad \text{and}$$

$$\overline{H}_k(P) = -\varlimsup_{n\to\infty} \frac{1}{(k-1)n} \log \sum_\omega P(\omega)^k \ ,$$

where the sums are taken over all cylinders ω of length n. When the limit exists we denote by $H_k(P)$ the common value. We note that the Rényi entropy of order k is also called generalized Rényi entropy and that the Rényi entropy of order 2 is often only called Rényi entropy.

The existence of the Rényi entropy of order k has not been proved for general stochastic processes. However, it was proved for Bernoulli processes, finite Markov chains (e.g. Rached, 1998), infinite Markov chains Ciuperca et al. (2011), Gibbs measures of a Hölder-continuous potential Haydn and Vaienti (2010), for ϕ-mixing measures Łuczak and Szpankowski (1997), for weakly ψ-mixing processes Haydn and Vaienti (2010) and for ψ_g-regular processes Abadi and Cardeño (2015).

Definition 2.4. The process X with stationary measure μ is α-mixing if there exists a function $\alpha : \mathbb{N} \to \mathbb{R}$ where $\alpha(g)$ converges to zero when g goes to infinity and such that

$$\sup_{A\in\mathcal{F}_0^n \ ; \ B\in\mathcal{F}_{n+g}^\infty} |\mu(A\cap B) - \mu(A)\mu(B)| \le \alpha(g) \ ,$$

for all $n \in \mathbb{N}$, where for $0 \le J \le L \le \infty$, $\mathcal{F}_J^L$ denotes the σ-algebra $\sigma(X_k, J \le k \le L)$.

When $\alpha(g)$ decreases exponentially fast to zero, we say that the process is α-mixing with exponential decay.

The process is ψ-mixing if there exists a function $\psi : \mathbb{N} \to \mathbb{R}$ where $\psi(g)$ converges to zero when g goes to infinity and such that

$$\sup_{A\in\mathcal{F}_0^n \ ; \ B\in\mathcal{F}_{n+g}^\infty} \left| \frac{\mu(A\cap B) - \mu(A)\mu(B)}{\mu(A)\mu(B)} \right| \le \psi(g),$$

for all $n \in \mathbb{N}$.

To obtain information on the growth length of the longest common substring for RLE sequences, we will need an assumption on the decay of the measure of cylinders:

(A) There exist $c > 0$ and $h > 0$, such that for any $n \in \mathbb{N}$ and any $a \in \mathcal{A}$

$$\mu(\underbrace{a \dots a}_{n}) \le ce^{-hn}.$$

We observe that in particular this assumption is always satisfied if the process is ψ-mixing with summable decay Galves and Schmitt (1997, Lemma 1).

First of all, without mixing assumption, we will prove an upper bound for the growth rate of the length of the longest common substring for RLE sequences:

Theorem 2.5. *If $\underline{H}_2(f_*\mu) > 0$ and if hypothesis (A) is satisfied, then for almost every $\mathbf{x}, \mathbf{y}$,*

$$\overline{\lim_{n\to\infty}} \frac{M_n^{RLE}(\mathbf{x},\mathbf{y})}{\log n} \le \frac{2}{\underline{H}_2(f_*\mu)}.$$

If the process is α-mixing with an exponential decay (or ψ-mixing with polynomial decay) and if for cylinders C_n of length n in $\mathcal{B}^n$, their preimage $f^{-1}C_n$ is of length at most $h(n)$ with $h(n) = o(n^\gamma)$ for some $\gamma > 0$, one could use the ideas of Coutinho et al. (2020) to get a lower bound. Nevertheless, the run-length encoder does not satisfy this last necessary assumption since preimage of cylinders under

f can have arbitrary length. Thus we present a different proof here to obtain the lower bound.

Theorem 2.6. *If $\underline{H}_2(f_*\mu) > 0$, hypothesis (A) is satisfied and the process is α-mixing with an exponential decay (or ψ-mixing with $\psi(g) = g^{-a}$ for some $a > 0$), then, for almost every realizations $\mathbf{x}, \mathbf{y}$,*

$$\underline{\lim}_{n\to\infty} \frac{M_n^{RLE}(\mathbf{x},\mathbf{y})}{\log n} \geq \frac{2}{\overline{H}_2(f_*\mu)}.$$

Thus, if the Rényi entropy exists, we get for almost every $\mathbf{x}, \mathbf{y}$,

$$\lim_{n\to\infty} \frac{M_n^{RLE}(\mathbf{x},\mathbf{y})}{\log n} = \frac{2}{H_2(f_*\mu)}.$$

Remark 2.7 (More than 2 sequences). One could wonder what will happen if we want to study the growth rate of the length of the longest common substring for k RLE sequences. Using the ideas presented in our proofs and in Barros and Rousseau (2021, Section 4), one could prove that, under the same assumptions of Theorem 2.6 and if the Rényi entropy of order k exists and is strictly positive, for almost every realizations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$,

$$\lim_{n\to\infty} \frac{M_n^{RLE}(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_k)}{\log n} = \frac{k}{(k-1)H_k(f_*\mu)}.$$

Theorem 2.5 and Theorem 2.6 will be proved in Section 4. We will now give examples satisfying our assumptions and where the Rényi entropy of the pushforward measure can be explicitly computed.

3. Examples

First of all, we will treat the case of Bernoulli processes and then of Markov chains. We emphasize that for Markov chains the situation and the computation are different when working with an alphabet of two symbols or an alphabet of more than two symbols.

3.1. *Bernoulli process.* Let us consider the alphabet $\mathcal{A} = \{a, b\}$ and the Bernoulli measure μ such that $\mu(a) = p$ and $\mu(b) = 1 - p$ with $0 < p < 1$. Hypothesis (A) can be easily checked and since this process is α-mixing with exponential decay, to apply our main theorem, we need to compute the Rényi entropy of the pushforward measure.

Let $n \in \mathbb{N}$. We assume that n is even (the odd case can be treated similarly). We observe that by definition of the run-length encoder, cylinders of length n can only have two types, i.e. the cylinder is of type 1 and $C_n = (a, k_1)(b, k_2)(a, k_3)\ldots(a, k_{n-1})(b, k_n)$ with $k_1, \ldots, k_n \in \mathbb{N}$ or the cylinder is of type 2 and $C_n = (b, k_1)(a, k_2)(b, k_3)\ldots(b, k_{n-1})(a, k_n)$ with $k_1, \ldots, k_n \in \mathbb{N}$.

It is important to notice that

$$f^{-1}\left((a, k_1)(b, k_2)\ldots(b, k_n)\right) = \underbrace{a\ldots a}_{k_1}\underbrace{b\ldots b}_{k_2}\ldots\underbrace{b\ldots b}_{k_n}a.$$

Indeed if the last symbol of C_n is (b, k_n), it does not only inform us that in the preimage we have a concatenation of k_n symbols b but also it imposes that this

concatenation must be followed by a symbol a, otherwise, if it was a symbol b, the last symbol of C_n would not be (b, k_n).

Thus, we have

$$\begin{aligned}
&\sum_{C_n} f_*\mu(C_n)^2 \\
&= \sum_{C_n \text{ of type 1}} f_*\mu(C_n)^2 + \sum_{C_n \text{ of type 2}} f_*\mu(C_n)^2 \\
&= \sum_{k_1,\dots,k_n} \mu(f^{-1}(a,k_1)(b,k_2)\dots(b,k_n))^2 \\
&+ \sum_{k_1,\dots,k_n\in\mathbb{N}} \mu(f^{-1}(b,k_1)(a,k_2)\dots(a,k_n))^2 \\
&= \sum_{k_1,\dots,k_n} \mu(\underbrace{a\dots a}_{k_1}\underbrace{b\dots b}_{k_2}\dots\underbrace{b\dots b}_{k_n}a)^2 + \mu(\underbrace{b\dots b}_{k_1}\underbrace{a\dots a}_{k_2}\dots\underbrace{a\dots a}_{k_n}b)^2 \\
&= \sum_{k_1,\dots,k_n} \left(\mu(a)^{k_1}\mu(b)^{k_2}\dots\mu(b)^{k_n}\mu(a)\right)^2 + \left(\mu(b)^{k_1}\mu(a)^{k_2}\dots\mu(a)^{k_n}\mu(b)\right)^2 \\
&= \sum_{k_1,\dots,k_n} p^{2k_1}(1-p)^{2k_2}\dots(1-p)^{2k_n}p^2 + (1-p)^{2k_1}p^{2k_2}\dots p^{2k_n}(1-p)^2 \\
&= p^2\left(\sum_{k=1}^{+\infty}\left(p^2\right)^k\right)^{n/2}\left(\sum_{k=1}^{+\infty}\left((1-p)^2\right)^k\right)^{n/2} \\
&+ (1-p)^2\left(\sum_{k=1}^{+\infty}\left(p^2\right)^k\right)^{n/2}\left(\sum_{k=1}^{+\infty}\left((1-p)^2\right)^k\right)^{n/2} \\
&= \left(p^2+(1-p)^2\right)\left(\frac{p^2}{1-p^2}\right)^{n/2}\left(\frac{(1-p)^2}{1-(1-p)^2}\right)^{n/2}.
\end{aligned}$$

This implies that the Rényi entropy of the pushforward measure exists and we have

$$H_2(f_*\mu) = -\lim_{n\to\infty}\frac{1}{n}\log\sum_{C_n} f_*\mu(C_n)^2 = -\frac{1}{2}\log\left(\frac{p(1-p)}{(1+p)(2-p)}\right).$$

Finally, applying Theorem 2.6, we have for almost every realizations $\mathbf{x}, \mathbf{y}$

$$\lim_{n\to\infty}\frac{M_n^{RLE}(\mathbf{x},\mathbf{y})}{\log n} = \frac{4}{\log\left(\frac{(1+p)(2-p)}{p(1-p)}\right)}.$$

3.2. *Markov chain with two states.* Let us consider the alphabet $\mathcal{A} = \{a, b\}$ and the transition matrix $P = (p_{ij})_{i,j\in\mathcal{A}}$ with $p_{aa} = p$ and $p_{bb} = q$ where $0 < p, q < 1$. The stationary measure μ is ψ-mixing with exponential decay (see e.g. Bradley, 2007) and hypothesis (A) is satisfied (see e.g. Galves and Schmitt, 1997, Lemma 1). Thus to apply our theorem we will compute the Rényi entropy of the pushforward measure.

As in the Bernoulli case, assuming that n is even, cylinders of length n can only have two forms, i.e. $C_n = (a, k_1)(b, k_2)(a, k_3)\dots(a, k_{n-1})(b, k_n)$ or $C_n =$

$(b,k_1)(a,k_2)(b,k_3)\dots(b,k_{n-1})(a,k_n)$ with $k_1,\dots,k_n\in\mathbb{N}$. Thus, we have

$$\begin{aligned}
&\sum_{C_n} f_*\mu(C_n)^2\\
&= \sum_{k_1,\dots,k_n} \mu(\underbrace{a\dots a}_{k_1}\underbrace{b\dots b}_{k_2}\dots\underbrace{b\dots b}_{k_n}a)^2+\mu(\underbrace{b\dots b}_{k_1}\underbrace{a\dots a}_{k_2}\dots\underbrace{a\dots a}_{k_n}b)^2\\
&= \sum_{k_1,\dots,k_n}\left(\mu(a)p_{aa}^{k_1-1}p_{ab}p_{bb}^{k_2-1}\dots p_{ab}p_{bb}^{k_n-1}p_{ba}\right)^2\\
&+\left(\mu(b)p_{bb}^{k_1-1}p_{ba}p_{aa}^{k_2-1}\dots p_{ba}p_{aa}^{k_n-1}p_{ab}\right)^2\\
&= (\mu(a)^2+\mu(b)^2)\left(p_{ab}^{n/2}\right)^2\left(p_{ba}^{n/2}\right)^2\left(\sum_{k=1}^{\infty}(p_{aa}^2)^{k-1}\right)^{n/2}\left(\sum_{k=1}^{\infty}(p_{bb}^2)^{k-1}\right)^{n/2}\\
&= (\mu(a)^2+\mu(b)^2)p_{ab}^n p_{ba}^n\left(\frac{1}{1-p_{aa}^2}\right)^{n/2}\left(\frac{1}{1-p_{bb}^2}\right)^{n/2}\\
&= (\mu(a)^2+\mu(b)^2)\left(\frac{1-p}{1+p}\right)^{n/2}\left(\frac{1-q}{1+q}\right)^{n/2}
\end{aligned}$$

and the Rényi entropy is

$$H_2(f_*\mu)=-\frac{1}{2}\log\left(\frac{(1-p)(1-q)}{(1+p)(1+q)}\right).$$

Applying Theorem 2.6, we have for almost every realizations $\mathbf{x},\mathbf{y}$

$$\lim_{n\to\infty}\frac{M_n^{RLE}(\mathbf{x},\mathbf{y})}{\log n}=\frac{4}{\log\left(\frac{(1+p)(1+q)}{(1-p)(1-q)}\right)}.$$

3.3. *Markov chain with more than 2 states.* To study Markov chains with more than two states, we will use another strategy which cannot be used for two states. The idea is that when the original process X is a Markov chain with finite alphabet, the process $f(X)$ is a Markov chain with infinite alphabet. However, when working with only two states, this process is not aperiodic preventing us to compute the Rényi entropy using the results of Ciuperca et al. (2011) (which are based on Perron-Frobenius Theorem).

Let us consider the alphabet $\mathcal{A}=\{\alpha_i\}_{1\le i\le N}$ and the transition matrix $P=(p_{ij})_{1\le i,j\le N}$ with $0<p_{ij}<1$ for every $1\le i,j\le N$. The stationary measure μ is ψ-mixing with exponential decay (see e.g. Bradley, 2007) and hypothesis (A) is satisfied (see e.g. Galves and Schmitt, 1997, Lemma 1). Thus to apply our theorem we will compute the Rényi entropy of the pushforward measure.

First of all, we observe that the process $f(X)$ is also a Markov chain on the alphabet $\mathcal{B}=\{(\alpha,k)\}_{\alpha\in\mathcal{A},k\in\mathbb{N}}$ with transition matrix $Q=(q_{(\alpha_i,k)(\alpha_j,\ell)})_{1\le i,j\le N,k,\ell\in\mathbb{N}}$ and initial distribution $\pi=(\pi((\alpha_i,k)))_{1\le i\le N,k\in\mathbb{N}}$. By definition of the run-length encoder, we observe that for all $1\le i\le N$ and $k,\ell\in\mathbb{N}$

$$q_{(\alpha_i,k)(\alpha_i,\ell)}=\mathbb{P}\left(f(X)_{n+1}=(\alpha_i,\ell)|f(X)_n=(\alpha_i,k)\right)=0.$$

Moreover, for $i\neq j$ and $k,\ell\in\mathbb{N}$, we have

$$q_{(\alpha_i,k)(\alpha_j,\ell)} \quad = \quad \mathbb{P}\left(f(X)_{n+1}=(\alpha_j,\ell)|f(X)_n=(\alpha_i,k)\right)$$

$$= \mu(\underbrace{\alpha_i \dots \alpha_i}_{k}\underbrace{\alpha_j \dots \alpha_j}_{\ell}\alpha_j^c).\mu(\underbrace{\alpha_i \dots \alpha_i}_{k}\alpha_i^c)^{-1}$$

where α^c can be any symbol in $\mathcal{A} \setminus \{\alpha\}$. Thus, we obtain

$$q_{(\alpha_i,k)(\alpha_j,\ell)} = \frac{p_{ij}p_{jj}^{\ell-1}(1-p_{jj})}{(1-p_{ii})}.$$

Then, for all $1 \le i \le N$ and $k \in \mathbb{N}$, we have

$$\pi((\alpha_i,k)) = \mathbb{P}(f(X)_1 = (\alpha_i,k)) = \mu(\underbrace{\alpha_i \dots \alpha_i}_{k}\alpha_i^c) = \mu(\alpha_i)p_{ii}^{k-1}(1-p_{ii}).$$

Since, $f(X)$ is a Markov chain over a countable alphabet, we will compute $H_2(f_*\mathbb{P})$ using Theorem 2 in Ciuperca et al. (2011):

Theorem 3.1 (Theorem 2 Ciuperca et al., 2011)**.** *Let $Y = (Y_n)_{n\in\mathbb{N}}$ be an irreducible and aperiodic Markov chain with denumerable state space E, transition matrix $Q = (q(i,j))_{(i,j)\in E^2}$ and initial distribution $\pi = (\pi(i))_{i\in E}$. If we have*

(1.A) $\sup_{(i,j)\in E} q(i,j) < 1$;

(1.B) there exists $\sigma_0 < 1$ such that for all $s > \sigma_0$

$$\sup_{i\in E}\sum_{j\in E} q(i,j)^s < \infty$$

and

$$\sum_{i\in E}\pi(i)^s < \infty;$$

(1.C) for all $\varepsilon > 0$ and all $s > \sigma_0$, there exists some $A \subset E$ with a finite number of elements, such that

$$\sup_{i\in E}\sum_{j\in E\setminus A} q(i,j)^s < \varepsilon$$

then for the Markov measure ν with initial distribution π and transition matrix Q and for $k > 1$, the Rényi entropy of order k exists and

$$H_k(\nu) = -\log\lambda_k$$

where λ_k is the largest positive eigenvalue of the matrix $Q_k = \left((q(i,j))^k\right)_{(i,j)\in E}$.

First, we observe that $f(X)$ is irreducible and aperiodic since $0 < p_{ij} < 1$ for every $1 \le i,j \le N$.

We also observe that if the alphabet $\mathcal{A}$ as only two symbols, $f(X)$ is periodic of period 2, thus we cannot apply Ciuperca et al. (2011).

We will now check the other assumptions to apply their results. Since $0 < p_{ij} < 1$ for every $1 \le i,j \le N$, we have

$$\sup_{1\le i,j\le N,k,\ell\in\mathbb{N}} q_{(\alpha_i,k)(\alpha_j,\ell)} = \sup_{1\le i,j\le N,k,\ell\in\mathbb{N}} \frac{p_{ij}p_{jj}^{\ell-1}(1-p_{jj})}{(1-p_{ii})} \le \sup_{1\le j\le N}(1-p_{jj}) < 1$$

and Assumption 1.A is satisfied.

Let $s > 0$. We have for any $1 \le i,j \le N$ and $k \in \mathbb{N}$

$$\sum_{\ell\in\mathbb{N}} q^s_{(\alpha_i,k)(\alpha_j,\ell)} = \sum_{\ell\in\mathbb{N}}\left(\frac{p_{ij}p_{jj}^{\ell-1}(1-p_{jj})}{(1-p_{ii})}\right)^s$$

$$= \left(\frac{p_{ij}(1-p_{jj})}{(1-p_{ii})}\right)^s \frac{1}{1-p_{jj}^s}.$$

Thus,

$$\begin{aligned}
&\sup_{1\le i\le N,k\in\mathbb{N}} \sum_{1\le j\le N,\ell\in\mathbb{N}} q^s_{(\alpha_i,k)(\alpha_j,\ell)} \\
&= \sup_{1\le i\le N} \sum_{1\le j\le N} \sum_{\ell\in\mathbb{N}} q^s_{(\alpha_i,k)(\alpha_j,\ell)} \\
&= \sup_{1\le i\le N} \sum_{1\le j\le N} \left(\frac{p_{ij}(1-p_{jj})}{(1-p_{ii})}\right)^s \frac{1}{1-p_{jj}^s} < +\infty. \qquad (3.1)
\end{aligned}$$

Moreover,

$$\begin{aligned}
\sum_{1\le i\le N,k\in\mathbb{N}} \pi((\alpha_i,k))^s &= \sum_{1\le i\le N,k\in\mathbb{N}} \left(\mu(\alpha_i)p_{ii}^{k-1}(1-p_{ii})\right)^s \\
&= \sum_{1\le i\le N} \frac{\mu(\alpha_i)^s(1-p_{ii})^s}{1-p_{ii}^s} < +\infty. \qquad (3.2)
\end{aligned}$$

(3.1) and (3.2) imply that Assumption 1.B is satisfied.

Let $\varepsilon > 0$, $s > 0$ and define

$$M = \sup_{1\le i\le N} \sum_{1\le j\le N} \left(\frac{p_{ij}(1-p_{jj})}{(1-p_{ii})}\right)^s.$$

Since $0 < p_{ij} < 1$ for every $1 \le i, j \le N$, it exists $m \in \mathbb{N}$ such that for all $1 \le j \le N$, we have

$$\frac{p_{jj}^{ms}}{1-p_{jj}^s} < \frac{\varepsilon}{M}.$$

Let $A = \{(\alpha_i,k)\}_{1\le i\le N,1\le k<m}$. We observe that A has a finite number of elements and that

$$\begin{aligned}
\sup_{1\le i\le N,k\in\mathbb{N}} \sum_{(\alpha_j,\ell)\in\mathcal{B}\setminus A} q^s_{(\alpha_i,k)(\alpha_j,\ell)} &= \sup_{1\le i\le N} \sum_{1\le j\le N} \sum_{\ell=m}^{\infty} q^s_{(\alpha_i,k)(\alpha_j,\ell)} \\
&= \sup_{1\le i\le N} \sum_{1\le j\le N} \left(\frac{p_{ij}(1-p_{jj})}{(1-p_{ii})}\right)^s \frac{p_{jj}^{ms}}{1-p_{jj}^s} \\
&< M.\frac{\varepsilon}{M} = \varepsilon.
\end{aligned}$$

Thus, Assumption 1.C is satisfied.

Finally, since $f(X)$ satisfies all the assumptions of Theorem 3.1, $H_2(f_*\mu)$ exists and

$$H_2(f_*\mu) = -\log\lambda$$

where λ is the largest positive eigenvalue of the matrix $Q_2 = \left(\left(q_{(\alpha_i,k)(\alpha_j,\ell)}\right)^2\right)_{1\le i,j\le N,k,\ell\in\mathbb{N}}$.

Applying Theorem 2.6, we have for almost every realizations $\mathbf{x}, \mathbf{y}$

$$\lim_{n\to\infty} \frac{M_n^{RLE}(\mathbf{x},\mathbf{y})}{\log n} = \frac{2}{-\log\lambda}.$$

4. Proof of the main results

To prove our theorems, a natural tool one would like to use is the stationarity of the measure. Since $f_*\mu$ is not stationary, we will need to use the stationarity of μ. In order to do so, we will introduce a new object

$$\tilde{M}_n(\mathbf{x},\mathbf{y}) = \max\left\{k \ : f(\sigma^i\mathbf{x})_1^k = f(\sigma^j\mathbf{y})_1^k \text{ for some } 0 \le i,j \le n-k\right\}.$$

First of all, we will explain how $\tilde{M}_n$ and M_n^{RLE} are related.

Lemma 4.1. *For every sequences* $\mathbf{x},\mathbf{y}$ *and every* $n \in \mathbb{N}$

$$M_n^{RLE}(\mathbf{x},\mathbf{y}) \ge \tilde{M}_n(\mathbf{x},\mathbf{y}) - 1. \tag{4.1}$$

Moreover, if $|f(\mathbf{x}_1^n)| \ge u(n)$ *and* $|f(\mathbf{y}_1^n)| \ge u(n)$ *for some* $u(n) \in \mathbb{N}$ *then*

$$M_{u(n)}^{RLE}(\mathbf{x},\mathbf{y}) \le \tilde{M}_n(\mathbf{x},\mathbf{y}). \tag{4.2}$$

Proof: For two sequences $\mathbf{x},\mathbf{y}$, assume that $\tilde{M}_n(\mathbf{x},\mathbf{y}) = \ell$. Thus, by definition, there exist $0 \le i,j \le n-\ell$ such that $f(\sigma^i\mathbf{x})_1^\ell = f(\sigma^j\mathbf{y})_1^\ell$.

Moreover, by definition of the run-length encoder f, we observe that the cylinder $f(\sigma^i\mathbf{x})_2^\ell$ always appears at some position in the cylinder $f(\mathbf{x})_1^n$, more precisely it exists $0 \le i' \le n-\ell$ such that $f(\sigma^i\mathbf{x})_2^\ell = f(\mathbf{x})_{i'}^{i'+\ell-2}$. Identically, it exists $0 \le j' \le n-\ell$ such that $f(\sigma^j\mathbf{y})_2^\ell = f(\mathbf{y})_{j'}^{j'+\ell-2}$.

Thus $f(\mathbf{x})_{i'}^{i'+\ell-2} = f(\mathbf{y})_{j'}^{j'+\ell-2}$ which implies that $M_n^{RLE}(\mathbf{x},\mathbf{y}) \ge \ell - 1 = \tilde{M}_n(\mathbf{x},\mathbf{y}) - 1$ and (4.1) is proved.

Now, assume that if $|f(\mathbf{x}_1^n)| \ge u(n)$ and $|f(\mathbf{y}_1^n)| \ge u(n)$ for some $u(n) \in \mathbb{N}$. One can observe that by definition of the run-length encoder, we must have $u(n) \le n$. Let us assume that $M_{u(n)}^{RLE}(\mathbf{x},\mathbf{y}) = \ell$. Thus, by definition, there exist $0 \le i,j \le u(n)-\ell$ such that $f(\mathbf{x})_i^{i+\ell-1} = f(\mathbf{y})_j^{j+\ell-1}$.

Moreover, since $|f(\mathbf{x}_1^n)| \ge u(n)$, it exists $0 \le i' \le n-\ell$ such that $f(\mathbf{x})_i^{i+\ell-1} = f(\sigma^{i'}\mathbf{x})_1^\ell$. Identically, it exists $0 \le j' \le n-\ell$ such that $f(\mathbf{y})_j^{j+\ell-1} = f(\sigma^{j'}\mathbf{y})_1^\ell$.

Thus, $f(\sigma^{i'}\mathbf{x})_1^\ell = f(\sigma^{j'}\mathbf{y})_1^\ell$ which implies that $\tilde{M}_n(\mathbf{x},\mathbf{y}) \ge \ell = M_{u(n)}^{RLE}(\mathbf{x},\mathbf{y})$ and (4.2) is proved.

□

Proof of Theorem 2.5: The proof of this theorem follows in part the lines of the proof of the Theorem 7 in Barros et al. (2019), nevertheless some subtle modifications are necessary since $f_*\mu$ is not stationary.

Let $\varepsilon > 0$ and denote

$$k_n = \left\lceil \frac{2\log n + \log\log n}{\underline{H_2}(f_*\mathbb{P}) - \varepsilon} \right\rceil.$$

We define $u(n) = \frac{n}{\log n^\delta}$ with $\delta > \frac{1}{h}$ and

$$V_n = \{\mathbf{x}, |f(\mathbf{x}_1^n)| \ge u(n)\}.$$

First of all, using (4.2), we observe that

$$\begin{aligned}\mathbb{P}(M_{u(n)}^{RLE} \ge k_n) &\le \mathbb{P}((V_n \times V_n) \cap M_{u(n)}^{RLE} \ge k_n) + \mathbb{P}((V_n \times V_n)^c) \\ &\le \mathbb{P}(\tilde{M}_n \ge k_n) + 3\mu(V_n^c).\end{aligned} \tag{4.3}$$

Firstly, we will estimate $\mathbb{P}(\tilde{M}_n \geq k_n)$. For $0 \leq i,j \leq n-1$ we define the following event

$$A_{i,j} = \{f(\sigma^i X)_1^{k_n} = f(\sigma^j Y)_1^{k_n}\}$$

and the following random variable

$$S_n = \sum_{i,j=0,\dots,n-1} \mathbb{1}_{A_{i,j}}. \tag{4.4}$$

It follows from our definitions and Markov's inequality that

$$\mathbb{P}\left(\tilde{M}_n \geq k_n\right) = \mathbb{P}\left(S_n \geq 1\right) \leq \mathbb{E}\left(S_n\right).$$

Moreover, we use the stationarity of μ to get

$$\begin{aligned}
\mathbb{E}\left(S_n\right) &= \sum_{0\leq i,j\leq n-1} \sum_{\omega\in\mathcal{B}^{k_n}} \mathbb{P}\left(f(\sigma^i X)_1^{k_n} = f(\sigma^j Y)_1^{k_n} = \omega\right) \\
&= \sum_{0\leq i,j\leq n-1} \sum_{\omega\in\mathcal{B}^{k_n}} \mu\left(f(\sigma^i X)_1^{k_n} = \omega\right) \mu\left(f(\sigma^j Y)_1^{k_n} = \omega\right) \\
&= n^2 \sum_{\omega\in\mathcal{B}^{k_n}} \mu\left(f^{-1}\omega\right)^2 = n^2 \sum_{\omega\in\mathcal{B}^{k_n}} f_*\mu\left(\omega\right)^2. \qquad (4.5)
\end{aligned}$$

Thus, for n large enough, by definition of $\underline{H}_2(f_*\mathbb{P})$ and of k_n, we have

$$\mathbb{P}\left(\tilde{M}_n \geq k_n\right) \leq n^2 \sum_{\omega\in\mathcal{B}^{k_n}} f_*\mu\left(\omega\right)^2 \leq n^2 e^{-k_n(\underline{H}_2(f_*\mathbb{P})-\varepsilon)} \leq \frac{1}{\log n}. \tag{4.6}$$

Let us now estimate $\mu(V_n^c) = \{\mathbf{x}, |f(\mathbf{x}_1^n)| < u(n)\}$. Let $\mathbf{x} \in V_n^c$, by definition of the run-length encoder, one can notice that since $|f(\mathbf{x}_1^n)| < u(n)$, it exists $a \in \mathcal{A}$ such that the cylinder $a...a$ of length $\left\lceil \frac{n}{u(n)} \right\rceil$ appears at some position in the cylinder $\mathbf{x}_1^n$, more precisely it exists $0 \leq i \leq n - \left\lceil \frac{n}{u(n)} \right\rceil$ such that $a...a = \mathbf{x}_{i+1}^{i+\lceil n/u(n)\rceil}$. Thus, we obtain

$$\begin{aligned}
\mu(V_n^c) &\leq \mu\left(\bigcup_{a\in\mathcal{A}} \bigcup_{0\leq i\leq n-\left\lceil \frac{n}{u(n)} \right\rceil} \sigma^{-i} a\dots a\right) \qquad (4.7) \\
&\leq \sum_{a\in\mathcal{A}} \sum_{0\leq i\leq n-\left\lceil \frac{n}{u(n)} \right\rceil} \mu\left(\sigma^{-i} a\dots a\right) \\
&= \left(n - \left\lceil \frac{n}{u(n)} \right\rceil + 1\right) \sum_{a\in\mathcal{A}} \mu\left(a\dots a\right).
\end{aligned}$$

Thus, using assumption (A) and since $u(n) = \frac{n}{\log n^\delta}$ with $\delta > \frac{1}{h}$, we obtain

$$\mu(V_n^c) \leq c|\mathcal{A}|ne^{-hn/u(n)} \leq c|\mathcal{A}|\frac{1}{n} \tag{4.8}$$

where $|\mathcal{A}|$ denotes the cardinality of $\mathcal{A}$.

Thus, combining (4.3), (4.6) and (4.8), we obtain

$$\mathbb{P}(M_{u(n)}^{RLE} \geq k_n) \leq \mathcal{O}\left((\log n)^{-1}\right).$$

Choosing a subsequence $(n_\ell)_{\ell\in\mathbb{N}}$ such that $n_\ell = \lceil e^{\ell^2}\rceil$ we have that $\sum_\ell \mathbb{P}\left(M^{RLE}_{u(n_\ell)} \geq k_{n\ell}\right) < +\infty$. Thus, by the Borel-Cantelli lemma, we have almost surely, if ℓ is large enough,

$$M^{RLE}_{u(n_\ell)} < k_{n_\ell}$$

and then

$$\frac{M^{RLE}_{u(n_\ell)}}{\log n_\ell} \leq \frac{1}{\underline{H}_2(f_*\mu) - \varepsilon}\left(2 + \frac{1+\log\log n_\ell}{\log n_\ell}\right).$$

Taking the limit superior in this inequality and observing that $(M^{RLE}_n)_n$, $(u(n))_n$ and $(n_\ell)_\ell$ are increasing, that $\lim_{\ell\to\infty} \frac{\log n_\ell}{\log n_{\ell+1}} = 1$ and that $\lim_{n\to\infty} \frac{\log u(n)}{\log n} = 1$, we obtain almost surely

$$\underline{\lim}_{n\to\infty} \frac{M^{RLE}_n}{\log n} = \underline{\lim}_{\ell\to\infty} \frac{M^{RLE}_{u(n)}}{\log n} = \underline{\lim}_{\ell\to\infty} \frac{M^{RLE}_{u(n_\ell)}}{\log n_\ell} \leq \frac{2}{\underline{H}_2(f_*\mathbb{P}) - \varepsilon}.$$

And the theorem is proved since ε can be chosen arbitrarily small.

□

Proof of Theorem 2.6: We will proved here the theorem when the process is α-mixing with exponential decay, the ψ-mixing case can be obtained similarly by a slight modification. Without loss of generality, we will also assume that $\alpha(g) = e^{-g}$.

Let $\varepsilon > 0$ and define

$$k_n = \left\lfloor \frac{2\log n + b\log\log n}{\overline{H_2}(f_*\mu) + \varepsilon} \right\rfloor$$

where b is a constant to be chosen.

First of all, using (4.1) and (4.4), we observe that

$$\mathbb{P}(M^{RLE}_n < k_n - 1) \leq \mathbb{P}(\tilde{M}_n < k_n) = \mathbb{P}(S_n = 0)$$

thus, by Chebyshev's inequality

$$\mathbb{P}\left(M^{RLE}_n < k_n - 1\right) \leq \frac{\mathrm{var}\,(S_n)}{\mathbb{E}\,(S_n)^2}. \tag{4.9}$$

For the variance of S_n, we use the following lemma (which will be proved after the proof of the theorem):

Lemma 4.2. *Under the assumptions of Theorem 2.6, for $g \in \mathbb{N}$, we have*

$$\begin{aligned}\mathrm{var}(S_n) \quad \leq \quad & 4(g+k_n)\mathbb{E}(S_n)^{3/2} + 4(g+k_n)^2\mathbb{E}(S_n) + 2n^4\alpha(g+k_n-k_n^3) \\ & +4n^3(g+k_n)\alpha(g+k_n-k_n^3) + \left(2n^4 + 4n^3(g+k_n)\right)ce^{-hk_n^2}.\end{aligned}$$

Since $k_n = \mathcal{O}(\log n)$, for $\beta > 3$ and $g = (\log n)^\beta$ we have

$$(g + k_n - k_n^3) \sim (\log n)^\beta.$$

Thus, since $\alpha(g) = e^{-g}$ and since $\log(n^5) = o\left((\log n)^\beta\right)$, we obtain

$$2n^4\alpha(g+k_n-k_n^2) = \mathcal{O}(n^{-1}) \tag{4.10}$$

and

$$4n^3(g+k_n)\alpha(g+k_n-k_n^2) = \mathcal{O}(n^{-1}). \tag{4.11}$$

By definition of k_n, for n large enough we have $hk_n^2 \geq 5\log n$, thus we obtain

$$\left(2n^4 + 4n^3(g+k_n)\right) ce^{-hk_n^2} = \mathcal{O}(n^{-1}). \tag{4.12}$$

Thus, combining (4.9) together with Lemma 4.2, (4.10), (4.11) and (4.12), we have

$$\mathbb{P}\left(M_n^{RLE} < k_n - 1\right) \leq \frac{\mathrm{var}\left(S_n\right)}{\mathbb{E}\left(S_n\right)^2} \leq \frac{4(g+k_n)}{\mathbb{E}\left(S_n\right)^{1/2}} + \frac{4(g+k_n)^2}{\mathbb{E}\left(S_n\right)} + \mathcal{O}(n^{-1}).$$

By (4.5) and by definitions of k_n and the Rényi entropy, we have

$$\mathbb{E}\left(S_n\right) = n^2 \sum_{\omega \in \mathcal{B}^{k_n}} f_*\mu\left(\omega\right)^2 \geq n^2 e^{-k_n(\underline{H}_2(f_*\mu)+\varepsilon)} \geq (\log n)^{-b}$$

and recalling that $(g+k_n) \sim (\log n)^\beta$, one can choose $b \ll -1$ to obtain

$$\mathbb{P}\left(M_n^{RLE} < k_n - 1\right) \leq \mathcal{O}\left((\log n)^{-1}\right).$$

Choosing a subsequence $(n_\ell)_{\ell\in\mathbb{N}}$ such that $n_\ell = \lceil e^{\ell^2} \rceil$ we have that $\sum_\ell \mathbb{P}\left(M_{n_\ell}^{RLE} < k_{n\ell} - 1\right) < +\infty$. Thus, by the Borel-Cantelli lemma, we have almost surely, if ℓ is large enough,

$$M_{n_\ell}^{RLE} \geq k_{n\ell} - 1$$

and then

$$\frac{M_{n_\ell}^{RLE}}{\log n_\ell} \geq \frac{1}{\overline{H}_2(f_*\mu)+\varepsilon}\left(2 + b\frac{1+\log\log n_\ell}{\log n_\ell}\right) - \frac{1}{\log n_\ell}.$$

Taking the limit inferior in this inequality and observing that $(M_n^{RLE})_n$ and $(n_\ell)_\ell$ are increasing and that $\lim\limits_{\ell\to\infty} \dfrac{\log n_\ell}{\log n_{\ell+1}} = 1$, we obtain almost surely

$$\varliminf_{n\to\infty} \frac{M_n^{RLE}}{\log n} = \varliminf_{\ell\to\infty} \frac{M_{n_\ell}^{RLE}}{\log n_\ell} \geq \frac{2}{\overline{H}_2(f_*\mu)+\varepsilon}.$$

And the theorem is proved since ε can be chosen arbitrarily small.

□

Proof of Lemma 4.2: To estimate the variance of S_n, we observe that

$$\mathrm{var}\left(S_n\right) = \sum_{0\leq i,i',j,j'\leq n-1} \mathbb{E}\left(\mathbb{1}_{A_{i,j}}\mathbb{1}_{A_{i',j'}}\right) - \mathbb{E}\left(S_n\right)^2. \tag{4.13}$$

Let $g \in \mathbb{N}$. Firstly, we assume that $i' - i > g + k_n$ and $j' - j > g + k_n$ (the case $i - i' > g + k_n$ and $j - j' > g + k_n$ can be treated identically), then we have

$$\begin{aligned}
&\mathbb{E}\left(\mathbb{1}_{A_{i,j}}\mathbb{1}_{A_{i',j'}}\right)\\
&= \sum_{\omega,\omega'\in\mathcal{B}^{k_n}} \mathbb{P}\left(f(\sigma^i X)_1^{k_n} = f(\sigma^j Y)_1^{k_n} = \omega, f(\sigma^{i'} X)_1^{k_n} = f(\sigma^{j'} Y)_1^{k_n} = \omega'\right)\\
&= \sum_{\omega,\omega'\in\mathcal{B}^{k_n}} \mu\left(f(\sigma^i X)_1^{k_n} = \omega, f(\sigma^{i'} X)_1^{k_n} = \omega'\right)\mu\left(f(\sigma^j Y)_1^{k_n} = \omega, f(\sigma^{j'} Y)_1^{k_n} = \omega'\right)\\
&= \sum_{\omega,\omega'\in\mathcal{B}^{k_n}} \mu\left(f(X)_1^{k_n} = \omega, f(\sigma^{i'-i} X)_1^{k_n} = \omega'\right)\mu\left(f(Y)_1^{k_n} = \omega, f(\sigma^{j'-j} Y)_1^{k_n} = \omega'\right)
\end{aligned}$$

by stationarity of μ.

To use the mixing property, we need to work with cylinders whose preimage under f does not have a length too large so that the gap is preserved. Thus, we define the set

$$\mathcal{Z}_n = \left\{\omega \in \mathcal{B}^{k_n} : |f^{-1}\omega| \leq k_n^3\right\}$$

and, using the α-mixing, we obtain

$$\begin{aligned}
&\sum_{\substack{\omega\in\mathcal{Z}_n\\ \omega'\in\mathcal{B}^{k_n}}} \mu\left(f(X)_1^{k_n}=\omega, f(\sigma^{i'-i}X)_1^{k_n}=\omega'\right)\mu\left(f(Y)_1^{k_n}=\omega, f(\sigma^{j'-j}Y)_1^{k_n}=\omega'\right)\\
&\leq \sum_{\substack{\omega\in\mathcal{Z}_n\\ \omega'\in\mathcal{B}^{k_n}}} \left[\mu\left(f(X)_1^{k_n}=\omega\right)\mu\left(f(X)_1^{k_n}=\omega'\right)+\alpha(g+k_n-k_n^3)\right]\\
&\quad\times\mu\left(f(Y)_1^{k_n}=\omega, f(\sigma^{j'-j}Y)_1^{k_n}=\omega'\right)\\
&\leq \sum_{\substack{\omega\in\mathcal{Z}_n\\ \omega'\in\mathcal{B}^{k_n}}} \mu\left(f(X)_1^{k_n}=\omega\right)\mu\left(f(X)_1^{k_n}=\omega'\right)\left[\mu\left(f(Y)_1^{k_n}=\omega\right)\mu\left(f(Y)_1^{k_n}=\omega'\right)\right.\\
&\qquad\left.+\alpha(g+k_n-k_n^3)\right]\\
&\quad+\sum_{\omega'\in\mathcal{B}^{k_n}}\alpha(g+k_n-k_n^3)\mu\left(f(\sigma^{j'-j}Y)_1^{k_n}=\omega'\right)\\
&\leq 2\alpha(g+k_n-k_n^3)+\left[\sum_{\omega\in\mathcal{B}^{k_n}} f_*\mu(\omega)^2\right]^2. \qquad (4.14)
\end{aligned}$$

When the length of the preimage cylinders is too large (i.e. $\omega \notin \mathcal{Z}_n$), we cannot use the mixing property, however, we observe that

$$\begin{aligned}
&\sum_{\substack{\omega\in\mathcal{Z}_n^C\\ \omega'\in\mathcal{B}^{k_n}}} \mu\left(f(X)_1^{k_n}=\omega, f(\sigma^{i'-i}X)_1^{k_n}=\omega'\right)\mu\left(f(Y)_1^{k_n}=\omega, f(\sigma^{j'-j}Y)_1^{k_n}=\omega'\right)\\
&\leq \sum_{\omega\in\mathcal{Z}_n^C}\mu\left(f(X)_1^{k_n}=\omega\right)\sum_{\omega'\in\mathcal{B}^{k_n}}\mu\left(f(Y)_1^{k_n}=\omega, f(\sigma^{j'-j}Y)_1^{k_n}=\omega'\right)\\
&\leq \sum_{\omega\in\mathcal{Z}_n^C}\mu\left(f(X)_1^{k_n}=\omega\right)\mu\left(f(Y)_1^{k_n}=\omega\right)\\
&= \sum_{\omega\in\mathcal{Z}_n^C}\mu\left(f^{-1}\omega\right)^2.
\end{aligned}$$

Using the same argument as the one leading to (4.7), we observe that if $\omega \in \mathcal{Z}_n^C$ then $|\omega| = k_n$ and $|f^{-1}\omega| > k_n^3$, thus there exist $a \in \mathcal{A}$ and $\kappa \in \mathbb{N}$ such that $f^{-1}\omega \subset \sigma^{-\kappa}a\ldots a$ where $|a\ldots a| \geq \frac{k_n^3}{k_n} = k_n^2$. Thus, one can use assumption (A) to observe that $\mu(f^{-1}\omega) \leq \mu(\sigma^{-\kappa}a\ldots a) = \mu(a\ldots a) \leq ce^{-hk_n^2}$. Since this inequality is valid for any $\omega \in Z_n^C$, we obtain

$$\sum_{\substack{\omega\in\mathcal{Z}_n^C\\ \omega'\in\mathcal{B}^{k_n}}} \mu\left(f(X)_1^{k_n}=\omega, f(\sigma^{i'-i}X)_1^{k_n}=\omega'\right)\mu\left(f(Y)_1^{k_n}=\omega, f(\sigma^{j'-j}Y)_1^{k_n}=\omega'\right)$$

$$\leq \sum_{\omega\in\mathcal{Z}_n^C} \mu\left(f^{-1}\omega\right)^2 \leq ce^{-hk_n^2} \sum_{\omega\in\mathcal{Z}_n^C} \mu\left(f^{-1}\omega\right) \leq ce^{-hk_n^2}. \tag{4.15}$$

Thus, (4.14) together with (4.15) gives us that when $i'-i > g+k_n$ and $j'-j > g+k_n$

$$\mathbb{E}\left(\mathbb{1}_{A_{i,j}}\mathbb{1}_{A_{i',j'}}\right) \leq 2\alpha(g+k_n-k_n^3) + \left[\sum_{\omega\in\mathcal{B}^{k_n}} f_*\mu\left(\omega\right)^2\right]^2 + ce^{-hk_n^2}.$$

We observe that when $i'-i > g+k_n$ and $j-j' > g+k_n$ (the case $i-i' > g+k_n$ and $j'-j > g+k_n$ can be treated identically) then we can obtain (4.14) only if we restrict our sum to $\omega' \in \mathcal{Z}_n$, thus the estimate for $\mathbb{E}\left(\mathbb{1}_{A_{i,j}}\mathbb{1}_{A_{i',j'}}\right)$ will be slightly different and we will have

$$\mathbb{E}\left(\mathbb{1}_{A_{i,j}}\mathbb{1}_{A_{i',j'}}\right) \leq 2\alpha(g+k_n-k_n^3) + \left[\sum_{\omega\in\mathcal{B}^{k_n}} f_*\mu\left(\omega\right)^2\right]^2 + 2ce^{-hk_n^2}.$$

Thus, since $\operatorname{card}\{0 \leq i,j,i',j' \leq n-1 \text{ s.t. } |i'-i| > g+k_n \text{ and } |j'-j| > g+k_n\} \leq n^4$, we have

$$\sum_{\substack{|i'-i|>g+k_n\\ |j'-j|>g+k_n}} \mathbb{E}\left(\mathbb{1}_{A_{i,j}}\mathbb{1}_{A_{i',j'}}\right) \tag{4.16}$$

$$\leq n^4\left(2\alpha(g+k_n-k_n^3) + \left[\sum_{\omega\in\mathcal{B}^{k_n}} f_*\mu\left(\omega\right)^2\right]^2 + 2ce^{-hk_n^2}\right)$$

$$= n^4\left(2\alpha(g+k_n-k_n^3) + 2ce^{-hk_n^2}\right) + \mathbb{E}(S_n)^2. \tag{4.17}$$

Now, we assume that $i'-i > g+k_n$ and $0 \leq j'-j \leq g+k_n$ (the other cases such that $|i'-i| > g+k_n$ and $|j'-j| \leq g+k_n$ or such that $|i'-i| \leq g+k_n$ and $|j'-j| > g+k_n$ can be treated identically), using the mixing property as in (4.14) and then, using Hölder's inequality, we have

$$\sum_{\substack{\omega\in\mathcal{Z}_n\\ \omega'\in\mathcal{B}^{k_n}}} \mu\left(f(X)_1^{k_n}=\omega, f(\sigma^{i'-i}X)_1^{k_n}=\omega'\right)\mu\left(f(Y)_1^{k_n}=\omega, f(\sigma^{j'-j}Y)_1^{k_n}=\omega'\right)$$

$$\leq \alpha(g+k_n-k_n^3)$$

$$+ \sum_{\substack{\omega\in\mathcal{Z}_n\\ \omega'\in\mathcal{B}^{k_n}}} \mu\left(f(X)_1^{k_n}=\omega\right)\mu\left(f(X)_1^{k_n}=\omega'\right)\mu\left(f(Y)_1^{k_n}=\omega, f(\sigma^{j'-j}Y)_1^{k_n}=\omega'\right)$$

$$\leq \alpha(g+k_n-k_n^3)$$

$$+ \int_{\mathcal{A}^{\mathbb{N}}} \mu\left(f(X)_1^{k_n}=f(y)_1^{k_n}\right)\mu\left(f(\sigma^{j'-j}X)_1^{k_n}=f(\sigma^{j'-j}y)_1^{k_n}\right)d\mu(y)$$

$$\leq \alpha(g+k_n-k_n^3)$$

$$+ \left[\int_{\mathcal{A}^{\mathbb{N}}} \mu\left(f(X)_1^{k_n}=f(y)_1^{k_n}\right)^2 d\mu(y)\right]^{1/2}\left[\int_{\mathcal{A}^{\mathbb{N}}} \mu\left(f(\sigma^{j'-j}X)_1^{k_n}=f(\sigma^{j'-j}y)_1^{k_n}\right)^2 d\mu(y)\right]^{1/2}$$

$$= \alpha(g+k_n-k_n^3) + \sum_{\omega\in\mathcal{B}^{k_n}} f_*\mu(\omega)^3$$
$$\leq \alpha(g+k_n-k_n^3) + \left[\sum_{\omega\in\mathcal{B}^{k_n}} f_*\mu(\omega)^2\right]^{3/2}$$

where the last inequality comes from the subadditivity of the map $x \mapsto x^{3/2}$.

For the terms with $\omega \notin \mathcal{Z}_n$ we will use the estimate (4.15). Thus, for $|i'-i| > g+k_n$ and $|j'-j| \leq g+k_n$ we have

$$\mathbb{E}\left(\mathbb{1}_{A_{i,j}}\mathbb{1}_{A_{i',j'}}\right) \leq \alpha(g+k_n-k_n^3) + \left[\sum_{\omega\in\mathcal{B}^{k_n}} f_*\mu(\omega)^2\right]^{3/2} + ce^{-hk_n^2}.$$

Moreover, since $\mathrm{card}\{0 \leq i,j,i',j' \leq n-1 \text{ s.t. } |i'-i| > g+k_n \text{ and } |j'-j| \leq g+k_n\} \leq 2n^3(g+k_n)$, we have

$$\begin{aligned}
&\sum_{\substack{|i'-i|>g+k_n\\ |j'-j|\leq g+k_n}} \mathbb{E}\left(\mathbb{1}_{A_{i,j}}\mathbb{1}_{A_{i',j'}}\right) + \sum_{\substack{|i'-i|\leq g+k_n\\ |j'-j|>g+k_n}} \mathbb{E}\left(\mathbb{1}_{A_{i,j}}\mathbb{1}_{A_{i',j'}}\right)\\
&\leq 4n^3(g+k_n)\left(\alpha(g+k_n-k_n^3) + \left[\sum_{\omega\in\mathcal{B}^{k_n}} f_*\mu(\omega)^2\right]^{3/2} + ce^{-hk_n^2}\right)\\
&= 4n^3(g+k_n)\left(\alpha(g+k_n-k_n^3) + ce^{-hk_n^2}\right) + 4(g+k_n)\mathbb{E}(S_n)^{3/2}. \qquad (4.18)
\end{aligned}$$

Finally, when $|i'-i| \leq g+k_n$ and $|j'-j| \leq g+k_n$, we just observe that

$$\begin{aligned}
\mathbb{E}\left(\mathbb{1}_{A_{i,j}}\mathbb{1}_{A_{i',j'}}\right) &\leq \mathbb{E}\left(\mathbb{1}_{A_{i,j}}\right)\\
&= \sum_{\omega\in\mathcal{B}^{k_n}} f_*\mu(\omega)^2
\end{aligned}$$

and since $\mathrm{card}\{0 \leq i,j,i',j' \leq n-1 \text{ s.t. } |i'-i| \leq g+k_n \text{ and } |j'-j| \leq g+k_n\} \leq 4n^2(g+k_n)^2$, we have

$$\sum_{\substack{|i'-i|\leq g+k_n\\ |j'-j|\leq g+k_n}} \mathbb{E}\left(\mathbb{1}_{A_{i,j}}\mathbb{1}_{A_{i',j'}}\right) \leq 4n^2(g+k_n)^2 \sum_{\omega\in\mathcal{B}^{k_n}} f_*\mu(\omega)^2 = 4(g+k_n)^2\mathbb{E}(S_n). \qquad (4.19)$$

Combining together the estimates (4.13), (4.17), (4.18) and (4.19), we obtain

$$\begin{aligned}
\mathrm{var}(S_n) \leq\ & 4(g+k_n)\mathbb{E}(S_n)^{3/2} + 4(g+k_n)^2\mathbb{E}(S_n) + 2n^4\alpha(g+k_n-k_n^3)\\
&+4n^3(g+k_n)\alpha(g+k_n-k_n^3) + \left(2n^4 + 4n^3(g+k_n)\right)ce^{-hk_n^2}
\end{aligned}$$

and the lemma is proved. □

Acknowledgements. The author would like to thank Rodrigo Lambert for useful discussions and the referee for useful suggestions and corrections to improve the paper.

References

Abadi, M., Cardeño, L., and Gallo, S. Potential well spectrum and hitting time in renewal processes. *J. Stat. Phys.*, **159** (5), 1087–1106 (2015). MR3345411.

Abadi, M., Gallo, S., and Rada-Mora, E. A. The shortest possible return time of β-mixing processes. *IEEE Trans. Inform. Theory*, **64** (7), 4895–4906 (2018). MR3819346.

Abadi, M. N. and Cardeño, L. Rényi entropies and large deviations for the first match function. *IEEE Trans. Inform. Theory*, **61** (4), 1629–1639 (2015). MR3332970.

Ahsan, S. B., Aziz, S. P., and Rahman, M. S. Longest common subsequence problem for run-length-encoded strings. In *2012 15th International Conference on Computer and Information Technology (ICCIT)*, pp. 36–41 (2012). DOI: 10.1109/ICCITechn.2012.6509736.

Apostolico, A., Landau, G. M., and Skiena, S. Matching for run-length encoded strings. *J. Complexity*, **15** (1), 4–16 (1999). MR1675807.

Barros, V., Liao, L., and Rousseau, J. On the shortest distance between orbits and the longest common substring problem. *Adv. Math.*, **344**, 311–339 (2019). MR3897435.

Barros, V. and Rousseau, J. Shortest Distance Between Multiple Orbits and Generalized Fractal Dimensions. *Ann. Henri Poincaré* (2021). DOI: 10.1007/s00023-021-01039-y.

Bowen, R. *Equilibrium states and the ergodic theory of Anosov diffeomorphisms.* Lecture Notes in Mathematics, Vol. 470. Springer-Verlag, Berlin-New York (1975). MR0442989.

Bradley, R. C. *Introduction to strong mixing conditions. Vol. 2.* Kendrick Press, Heber City, UT (2007). ISBN 0-9740427-7-3. MR2325295.

Chen, K.-Y., Hsu, P.-H., and Chao, K.-M. Hardness of comparing two run-length encoded strings. *J. Complexity*, **26** (4), 364–374 (2010). MR2669663.

Ciuperca, G., Girardin, V., and Lhote, L. Computation and estimation of generalized entropy rates for denumerable Markov chains. *IEEE Trans. Inform. Theory*, **57** (7), 4026–4034 (2011). MR2840440.

Collet, P., Giardina, C., and Redig, F. Matching with shift for one-dimensional Gibbs measures. *Ann. Appl. Probab.*, **19** (4), 1581–1602 (2009). MR2538081.

Coutinho, A., Lambert, R., and Rousseau, J. Matching strings in encoded sequences. *Bernoulli*, **26** (3), 2021–2050 (2020). MR4091100.

Dębowski, Ł. On the vocabulary of grammar-based codes and the logical consistency of texts. *IEEE Trans. Inform. Theory*, **57** (7), 4589–4599 (2011). MR2840476.

Dębowski, Ł. Maximal repetition and zero entropy rate. *IEEE Trans. Inform. Theory*, **64** (4, part 1), 2212–2219 (2018). MR3782249.

Doukhan, P. *Mixing. Properties and examples*, volume 85 of *Lecture Notes in Statistics*. Springer-Verlag, New York (1994). ISBN 0-387-94214-9. MR1312160.

Fan, S., Liao, L., and Qiu, Y. Stationary determinantal processes: ψ-mixing property and L^q-dimensions. *ArXiv Mathematics e-prints* (2019). arXiv: 1911.04718.

Freschi, V. and Bogliolo, A. Longest common subsequence between run-length-encoded strings: a new algorithm with improved parallelism. *Inform. Process. Lett.*, **90** (4), 167–173 (2004). MR2050885.

Galves, A. and Schmitt, B. Inequalities for hitting times in mixing dynamical systems. *Random Comput. Dynam.*, **5** (4), 337–347 (1997). MR1483874.

Haydn, N. and Vaienti, S. The Rényi entropy function and the large deviation of short return times. *Ergodic Theory Dynam. Systems*, **30** (1), 159–179 (2010). MR2586350.

Hinds, S. C., Fisher, J. L., and D'Amato, D. P. A document skew detection method using run-length encoding and the Hough transform. In *[1990] Proceedings. 10th International Conference on Pattern Recognition*, volume i, pp. 464–468 vol.1 (1990). DOI: 10.1109/ICPR.1990.118147.

Hooshmand, S., Tavakoli, N., Abedin, P., and Thankachan, S. V. On computing average common substring over run length encoded sequences. *Fund. Inform.*, **163** (3), 267–273 (2018). MR3879106.

Hunter, R. and Robinson, A. H. International digital facsimile coding standards. *Proceedings of the IEEE*, **68** (7), 854–867 (1980). DOI: 10.1109/PROC.1980.11751.

Kontoyiannis, I., Algoet, P. H., Suhov, Y. M., and Wyner, A. J. Nonparametric entropy estimation for stationary processes and random fields, with applications to English text. *IEEE Trans. Inform. Theory*, **44** (3), 1319–1327 (1998). MR1616653.

Kontoyiannis, I. and Suhov, Y. Prefixes and the entropy rate for long-range sources. In *Probability, statistics and optimisation*, Wiley Ser. Probab. Math. Statist. Probab. Math. Statist., pp. 89–98. Wiley, Chichester (1994). MR1320744.

Łuczak, T. and Szpankowski, W. A suboptimal lossy data compression based on approximate pattern matching. *IEEE Trans. Inform. Theory*, **43** (5), 1439–1451 (1997). MR1476778.

Miall, A. *The geology of stratigraphic sequences*. Second edition (2010). ISBN 978-3-642-05026-8. DOI: 10.1007/978-3-642-05027-5.

Mokkadem, A. Mixing properties of ARMA processes. *Stochastic Process. Appl.*, **29** (2), 309–315 (1988). MR958507.

Neuhauser, C. A phase transition for the distribution of matching blocks. *Combin. Probab. Comput.*, **5** (2), 139–159 (1996). MR1400960.

Rached, Z. Rényi's Entropy Rate For Discrete Markov Sources (1998). Master of Science Project.

Robinson, A. H. and Cherry, C. Results of a prototype television bandwidth compression scheme. *Proceedings of the IEEE*, **55** (3), 356–364 (1967). DOI: 10.1109/PROC.1967.5493.

Rousseau, J. Longest common substring for random subshifts of finite type (2021+). To appear in *Ann. Inst. H. Poincaré Probab. Stat.*

Sayood, K. *Introduction to Data Compression*. fifth edition (2017). ISBN 978-0-12-415796-5. DOI: 10.1016/C2010-0-69630-1.

Walters, P. Ruelle's operator theorem and g-measures. *Trans. Amer. Math. Soc.*, **214**, 375–387 (1975). MR412389.

Waterman, M. *Introduction to Computational Biology: Maps, Sequences and Genomes*. Chapman and Hall, London (1995). ISBN 9780412993916.

Xu, D., Kurani, A. S., Furst, J. D., and Raicu, D. S. Run-length Encoding for Volumetric Texture. In *The 4th IASTED International Conference on Visualization, Imaging, and Image Processing* (2004).

Ziv, J. and Lempel, A. A universal algorithm for sequential data compression. *IEEE Trans. Inform. Theory*, **IT-23** (3), 337–343 (1977). MR530215.

Ziv, J. and Lempel, A. Compression of individual sequences via variable-rate coding. *IEEE Trans. Inform. Theory*, **24** (5), 530–536 (1978). MR507465.