# Large Deviation Property of Waiting Times for Markov and Mixing Processes


Vatsal Sharan
Indian Institute of Technology Kanpur
Kanpur-208016
Email: vatsal@iitk.ac.in

Rakesh K Bansal
Indian Institute of Technology Kanpur
Kanpur-208016
Email: rkb@iitk.ac.in



***Abstract*—In this work, we study the asymptotic properties of the waiting time until the opening string in the realization of a process first appears in an independent realization of the same or a different process. We first establish that the normalized waiting time between two independent realizations of a single source obeys the large deviation property for a class of mixing processes. Using the method of Markov types, we extend the result to when both the sequences are realizations of two distinct irreducible and aperiodic Markov sources.**



***Index Terms*—Large deviation property, waiting times, Markov types, Markov processes, mixing processes**


## I. Introduction

For two processes $X$ and $Y$ having probability measures $P$ and $Q$, the *waiting time* $W_n$ is defined as the time until the opening string $x_1^n$ in the realization $x$ of the $X$ process first appears in an independent realization $y$ produced by the $Y$ process. Wyner and Ziv [1] initiated the study of the asymptotics of waiting times and showed that for $x$ and $y$ realizations of a stationary ergodic Markov source $(1/n)\log W_n \rightarrow H$ in probability, where $H$ is the entropy rate of the Markov source. Shields [2] proved that the result holds almost surely. Luczak and Szpankowski [3] analysed waiting times allowing distortion between the strings. Kontoyiannis [4] showed that for $X$ ergodic and $Y$ Markov, $(1/n)\log W_n \rightarrow H(P) + D(P \parallel Q)$ a.s., where $D(P \parallel Q)$ is the relative entropy rate, defined in Section IV. It seems natural to examine the conditions on $X$ and $Y$ under which the large deviation property holds for the normalized waiting time. We establish a class of mixing processes for which the large deviation property holds for Wyner and Ziv's result. We also show that the large deviation property holds for Kontoyiannis' result if $X$ and $Y$ both are irreducible and aperiodic Markov sources.

Abadi [5] proved an upper bound for the difference between the exponential distribution and the distribution of the first occurrence of a string in a stochastic process (hitting time) with a finite alphabet, which we shall use in our work. We use the idea that the waiting time conditioned on a given sequence is the same as it's hitting time. To prove the large deviation property for Kontoyiannis' result, we will also use the theory of *Markov types* introduced in 1981 by Davisson *et al.* [6].

The rest of the paper is organized as follows. Section II introduces the notation and previous results. In Section III we prove the large deviation property for the normalized waiting time between realizations of the same source satisfying certain mixing conditions. In Section IV we discuss the theory of Markov types and extend our result to realizations of two distinct irreducible and aperiodic Markov sources. Section V contains the concluding remarks and potential applications.

## II. Notation and Known Results

Let $\{X_n\}_{n=-\infty}^{n=\infty}$ and $\{Y_n\}_{n=-\infty}^{n=\infty}$ denote stationary and ergodic processes defined on $(\mathscr{A}_{-\infty}^{\infty}, \sigma, P)$ and $(\mathscr{A}_{-\infty}^{\infty}, \sigma, Q)$ respectively. $\mathscr{A}$ denotes the common set of alphabets which is taken to be finite and $\sigma$ is the sigma field generated by finite dimensional cylinders. $\mathscr{A}^n$ denotes the set of all $n$-long sequences over $\mathscr{A}$. $P$ and $Q$ are the probability measures. We denote $\{X_n\}_{n=-\infty}^{n=\infty}$ by $X$ and $\{Y_n\}_{n=-\infty}^{n=\infty}$ by $Y$. $X$ is called $\psi$-mixing if

$$\sup_{A\in\sigma_{-\infty}^{n}, B\in\sigma_{n+l}^{\infty}} \frac{|P(A\cap B) - P(A)P(B)|}{P(A)P(B)} \leq \psi(l) \qquad (1)$$

where $\psi(l)$ is a decreasing sequence converging to 0 and $\sigma_i^j$ denotes the sigma algebra generated by $X_i^j = X_i X_{i+1} ... X_j$. $X$ is called $\phi$-mixing if

$$\sup_{A\in\sigma_{-\infty}^{n}, B\in\sigma_{n+l}^{\infty}} \frac{|P(A\cap B) - P(A)P(B)|}{P(A)} \leq \phi(l) \qquad (2)$$

where $\phi(l)$ is a decreasing sequence converging to 0. For each block $B \in \mathscr{A}^n$, let $[B] = \{x : x_1^n = B\}$ denote the cylinder set defined by $B$. Define the waiting time (hitting time) to the cylinder set [B] by

$$\tau_B(x) = \inf\{i \geq 1 : \mathrm{T}^i(x) \in [B]\} \qquad (3)$$

where T is the left shift map defined by $(\mathrm{T}x)_k = x_{k+1}$. Let $\{x_n\}_{n=-\infty}^{n=\infty}$ and $\{y_n\}_{n=-\infty}^{n=\infty}$ denote particular realizations of $X$ and $Y$ respectively. Now define the waiting time between $x_1^n$ and $\{y_n\}_{n=-\infty}^{n=\infty}$ to be:

$$W_n = W_n(x_1^n, y) = \min\{j \geq 1 : x_1^n = y_j^{j+n-1}\} \qquad (4)$$

As a *dual* of waiting time $W_n$, match length $L_m = L_m(x, y)$ is defined as follows:

$$L_m = \max\{j \geq 1 : x_1^j = y_k^{k+j-1}, \text{for some } k \in \{1, ..., m\}\} \qquad (5)$$

**Abadi's Theorem [5]:** Let $X$ be $\psi$-mixing or $\phi$-mixing with $\phi$ summable i.e. $\sum_l \phi(l) < \infty$. Then, there exist constants

$C > 0$, $0 < \xi_1 < 1 < \xi_2 < \infty$, such that for all $n \in \mathbb{N}$, $A \in \mathscr{A}^n$ and $t > 0$, there exists $\xi_A \in [\xi_1, \xi_2]$, for which the following inequality holds:

$$|P(\tau_A > t) - e^{-\xi_A P(A)t}| \leq C\epsilon(A)e^{-\xi_A P(A)t}(\xi_A P(A)t \vee 1)$$

where $\epsilon(A) = \inf_{n \leq \Delta \leq 1/P(A)}[\Delta P(A) + *(\Delta)]$, $C > 0$ is a constant, $*$ is $\psi$ or $\phi$, $(a \vee b)$ denotes $\max(a, b)$ and $C\epsilon(A) \to 0$ as $n \to \infty$.

From now, we will denote $C\epsilon(A)$ by $c_A$. The following result is from [7] (Eq. 4) which we are stating as a Lemma for ready reference.

**Lemma 1:** For an exponentially $\phi$-mixing process (i.e. $\phi(l) \leq e^{-l}, \forall\, l \geq l_0$), for all $n$-long sequences $s = x_1^n \in \mathscr{A}^n$, there exist $D > 0$ and $z > 0$, s.t. $\forall\, n \geq n_0$

$$c_s \leq De^{-zn}$$

We will also make use of the following definition and fact later-

**Definition 1 [8]:** $X$ has exponential rates for entropy if for every $\epsilon > 0$,

$$P(\{x_1^n : 2^{-n(H(P)+\epsilon)} \leq P(x_1^n) \leq 2^{-n(H(P)-\epsilon)}\}) \geq 1 - e^{-k(\epsilon)n}$$

where $k(\epsilon) > 0\; \forall\, \epsilon > 0, k(0) = 0$.

**Fact 1 [8]:** An irreducible and aperiodic Markov source is exponentially $\phi$-mixing and has exponential rates for entropy.

## III. Waiting time for realizations of same source

Let $H$ denote the entropy rate of Source $S$ having measure $P$ and $\mathbb{P}$ denote the product measure $P \times P$.

**Lemma 2:** If $S$ satisfies $\psi$-mixing condition or $\phi$-mixing condition with summable coefficients and with exponential rates for entropy, then for two independent realizations, $x$ and $y$ of the source,

$$\mathbb{P}(\frac{\log W_n}{n} > H + \epsilon) \leq e^{-f(\epsilon)n} \qquad \forall n \geq N_1(\epsilon)$$

where $f(\epsilon) > 0\; \forall\, \epsilon > 0$ and $f(0) = 0$.

**Lemma 3:** If $S$ is exponentially $\phi$-mixing and has exponential rates for entropy, then for two independent realizations, $x$ and $y$ of the source,

$$\mathbb{P}(\frac{\log W_n}{n} < H - \epsilon) \leq e^{-g(\epsilon)n} \qquad \forall n \geq N_2(\epsilon)$$

where $g(\epsilon) > 0\; \forall\, \epsilon > 0$ and $g(0) = 0$.

Combining Lemma 2 and Lemma 3 we have

**Theorem 1:** If $S$ is exponentially $\phi$-mixing and has exponential rates for entropy, then for two independent realizations, $x$ and $y$ of the source,

$$\mathbb{P}(|\frac{\log W_n}{n} - H| > \epsilon) \leq 2e^{-h(\epsilon)n} \qquad \forall n \geq N(\epsilon)$$

where $h(\epsilon) = \min(f(\epsilon), g(\epsilon))$ and $N(\epsilon) = \max\{N_1(\epsilon), N_2(\epsilon)\}$.

*Hence if $S$ is an irreducible and aperiodic Markov source, then the normalized waiting time between two independent realizations of $S$ obeys the large deviation property.*

### A. Proof of Lemma 2

We define $A_n(\delta)$ as the following set of $n$-long sequences,

$$A_n(\delta) = \{x_1^n : 2^{-n(H+\delta)} \leq P(x_1^n) \leq 2^{-n(H-\delta)}\} \quad (6)$$

$$\begin{aligned}
&\mathbb{P}(\frac{\log W_n}{n} > H + \epsilon) = \mathbb{P}(W_n > 2^{n(H+\epsilon)}) \\
&= \sum_{s \in \mathscr{A}^n} P(s)\mathbb{P}(W_n > 2^{n(H+\epsilon)}|x_1^n = s) \\
&= \sum_{s \in A_n(\delta)} P(s)\mathbb{P}(W_n > 2^{n(H+\epsilon)}|x_1^n = s) \\
&\quad + \sum_{s \in A_n(\delta)^c} P(s)\mathbb{P}(W_n > 2^{n(H+\epsilon)}|x_1^n = s) \\
&\overset{(a)}{\leq} \sum_{s \in A_n(\delta)} P(s)e^{-\xi_s P(s)2^{n(H+\epsilon)}}(1 + c_s(\xi_s P(s)2^{n(H+\epsilon)} \vee 1) \\
&\quad + \sum_{s \in A_n(\delta)^c} P(s) \\
&\overset{(b)}{\leq} \sum_{s \in A_n(\delta)} P(s)e^{-\xi_1 P(s)2^{n(H+\epsilon)}}(1 + c_s(\xi_2 P(s)2^{n(H+\epsilon)} \vee 1) \\
&\quad + \sum_{s \in A_n(\delta)^c} P(s) \qquad (7)
\end{aligned}$$

where (a) follows from Abadi's Theorem (waiting time conditioned on $s$ is the same as the hitting time for $s$), (b) follows as $\xi_s \in [\xi_1, \xi_2]$. For $s \in A_n(\delta)$, we have

$$2^{n(\epsilon-\delta)} \leq 2^{n(H+\epsilon)}P(s) \leq 2^{n(\epsilon+\delta)}. \quad (8)$$

For every $\epsilon > 0$ choose $\delta = \frac{\epsilon}{2}$. Consequently, we have

$$2^{(\frac{n\epsilon}{2})} \leq 2^{n(H+\epsilon)}P(s) \leq 2^{(\frac{3n\epsilon}{2})}. \quad (9)$$

Since $c_s \to 0$ as $n \to \infty$, for a given $d > 0, c_s < d\; \forall\, n \geq n_0$. Also, from Definition 1, $P(A_n(\frac{\epsilon}{2})) \geq 1 - e^{-k(\frac{\epsilon}{2})n}$ for processes with exponential rates for entropy. Further, $\xi_2 2^{(\frac{3n\epsilon}{2})} > 1$ as $\xi_2 > 1$. Hence we get,

$$\begin{aligned}
&\mathbb{P}(\frac{\log W_n}{n} > H + \epsilon) \\
&\leq \sum_{s \in A_n(\frac{\epsilon}{2})} P(s)e^{-\xi_1 2^{\frac{n\epsilon}{2}}}(1 + d(\xi_2 2^{(\frac{3n\epsilon}{2})})) + e^{-k(\frac{\epsilon}{2})n} \\
&\leq e^{-\xi_1 2^{\frac{n\epsilon}{2}}}(1 + d(\xi_2 2^{(\frac{3n\epsilon}{2})})) + e^{-k(\frac{\epsilon}{2})n} \\
&\leq e^{-f(\epsilon)n} \qquad (10)
\end{aligned}$$

where $f(\epsilon)$ is a real positive valued function for all $\epsilon > 0$ and $f(0) = 0$. □

### B. Proof of Lemma 3

$A_n(\frac{\epsilon}{2})$ denotes the same set as considered earlier. For each $s \in A_n(\frac{\epsilon}{2})$

$$2^{-(\frac{3n\epsilon}{2})} \leq 2^{n(H-\epsilon)}P(s) \leq 2^{-(\frac{n\epsilon}{2})} \quad (11)$$

$$\begin{aligned}
&\mathbb{P}(\frac{\log W_n}{n} < H - \epsilon) \\
=&1 - \mathbb{P}(\frac{\log W_n}{n} \geq H - \epsilon) \\
=&1 - \mathbb{P}(W_n \geq 2^{n(H-\epsilon)}) \\
=&1 - \sum_{s \in \mathscr{A}^n} P(s)\mathbb{P}(W_n \geq 2^{n(H-\epsilon)} | x_1^n = s) \\
\overset{(a)}{\leq}&1 - \sum_{s \in A_n(\frac{\epsilon}{2})} P(s)e^{-\xi_s P(s) 2^{n(H-\epsilon)}}(1 - c_s(\xi_s P(s) 2^{n(H-\epsilon)} \vee 1)) \\
\overset{(b)}{\leq}&1 - \sum_{s \in A_n(\frac{\epsilon}{2})} P(s)e^{-\xi_2 2^{-(\frac{n\epsilon}{2})}}(1 - c_s(\xi_2 2^{-(\frac{n\epsilon}{2})} \vee 1)) \\
\overset{(c)}{\leq}&1 - \sum_{s \in A_n(\frac{\epsilon}{2})} P(s)e^{-\xi_2 2^{-(\frac{n\epsilon}{2})}}(1 - c_s) \\
\overset{(d)}{\leq}&1 - \sum_{s \in A_n(\frac{\epsilon}{2})} P(s)e^{-\xi_2 2^{-(\frac{n\epsilon}{2})}}(1 - De^{-zn}) \\
=&1 - e^{-\xi_2 2^{-(\frac{n\epsilon}{2})}}(1 - De^{-zn}) \sum_{s \in A_n(\frac{\epsilon}{2})} P(s) \quad (12)
\end{aligned}$$

(a) follows from Abadi's Theorem and ignoring the contribution of the negative term due to sequences belonging to $A_n(\frac{\epsilon}{2})^c$, (b) follows as $\xi_s \in [\xi_1, \xi_2]$, (c) follows as $\xi_2 2^{-(\frac{n\epsilon}{2})} < 1$ eventually, (d) follows from Lemma 1. By Definition 1, $P(A_n(\frac{\epsilon}{2})) \geq 1 - e^{-k(\frac{\epsilon}{2})n}$ for processes with exponential rates for entropy. So,

$$\begin{aligned}
&\mathbb{P}(\frac{\log W_n}{n} < H - \epsilon) \\
&\leq 1 - e^{-\xi_2 2^{-(\frac{n\epsilon}{2})}}(1 - De^{-zn})(1 - e^{k(\frac{\epsilon}{2})n}) \\
&\leq 1 - e^{-\xi_2 2^{-(\frac{n\epsilon}{2})}}(1 - C'e^{-u(\epsilon)n}) \\
&\overset{(e)}{\leq} 1 - (1 - \xi_2 2^{-(\frac{n\epsilon}{2})})(1 - C'e^{-u(\epsilon)n}) \\
&\leq e^{-g(\epsilon)n} \quad (13)
\end{aligned}$$

(e) follows as $e^{-x} > 1 - x \; \forall \; x > 0$. Here $C' > 0$, $u(\epsilon)$ and $g(\epsilon)$ are positive valued functions for $\epsilon > 0$, $g(0) = 0$. □

## IV. Waiting time for realizations of two distinct Markov Sources

Let $X$ and $Y$ be two *irreducible* and *aperiodic* Markov sources distributed according to the measures $P$ and $Q$ respectively and taking values from a finite state space $\mathscr{A} = \{a_1, ..., a_K\}$. Let $\mathbb{P}$ denote the product measure $P \times Q$. $P_n$ and $Q_n$ denote the finite dimensional marginals of $P$ and $Q$ respectively. We assume $P_n \ll Q_n$, otherwise there will exist finite strings $x_1^n$ such that $P(x_1^n) > 0$ but $Q(x_1^n) = 0$ and the waiting time will be infinite with non-zero probability. $\Gamma_P = \{p_{jk}\} = Pr\{X_i = a_k | X_{i-1} = a_j\}$ is the transition probability matrix for $P$ and $\Gamma_Q$ is the transition probability matrix for $Q$. Let $p_s(0)$ denote the initial probability of state $a_s$ for source with measure $P$. There exists a unique stationary distribution for $P$ denoted by $G = \{g_1, ..., g_K\}$ such that $G\Gamma_P = G$. Let the probability distribution defined by the $j^{th}$ row of $\Gamma_P$ and $\Gamma_Q$ be denoted by $P^j$ and $Q^j$ respectively. The entropy rate $H(P)$ of the Markov source is defined as $\sum_{j=1}^{j=K} g_j H(P^j)$. The relative entropy rate $D(P \parallel Q)$ is defined as $\sum_{j=1}^{K} g_j D_{KL}(P^{(j)} \parallel Q^{(j)})$ with $D_{KL}(. \parallel .)$ being the usual definition of Kullback-Leibler divergence. For $X$ ergodic and $Y$ Markov, the following result has been established in [4],

$$\lim_{n \to \infty} \frac{1}{n} \log W_n = H(P) + D(P \parallel Q) \qquad \mathbb{P} - a.s. \quad (14)$$

We summarize the necessary results on Markov types here, the reader can refer to Davisson *et al.* [6] for a more detailed treatment. A Markov type $m_n$ is defined as the set of all $n$-long sequences over a finite alphabet having the same transition counts from letter to letter. For $x_1^n = a_{i_1}...a_{i_n}$, $n_{jk}(x)$ is defined as the number of transitions from $a_j$ to $a_k$, with $a_{i_n}$ preceding $a_{i_1}$. The Markov composition of $x_1^n$ is the square matrix $A(x_1^n)$ formed by the entries $n_{jk}(x_1^n)$. Two sequences are said to belong to the same Markov type if they have the same Markov composition. As a Markov type is defined by $K^2$ entries of $A(x_1^n)$ each taking $(n+1)$ possible values, the number of Markov types of order $n$ is not larger than $(n+1)^{K^2}$. We define $n_j = \sum_{k=1}^{K} n_{jk}$. It follows that $\sum_{k=1}^{K} n_j = n$. The empirical matrix $\Phi(x_1^n)$ is obtained by normalizing $A(x_1^n)$ as follows: the all-zero rows of $A(x_1^n)$ are left untouched and the others are normalized by dividing each element $n_{jk}$ by $n_j$. The $j$th row of $\Phi$ is denoted by $F^j$. The empirical probability distribution $F(x_1^n)$ is defined as $\frac{n_j}{n} = f_j \; \forall \; 1 \leq j \leq K$. The empirical entropy $H(\Phi|F)$ is defined as $\sum_{j=1}^{j=K} f_j H(F^j)$. The empirical relative entropy $D(\Phi \parallel \Gamma_P | F)$ is defined as $D(\Phi \parallel \Gamma_P | F) = \sum_{j=1}^{K} f_j D_{KL}(F^{(j)} \parallel \Gamma_P^{(j)})$ where $\Gamma_P^{(j)}$ is the $j$th row of $\Gamma_P$. We also define the number of "forbidden" transitions for a Markov type $m_n$ as $V_P(m_n) = \sum_{\{(j,k): p_{jk} = 0\}} n_{jk}(x_1^n)$ for any sequence $x_1^n$ having Markov type $m_n$ and measure $P$.

As $P(x_1^n) > 0$ we restrict our attention to sequences $x_1^n = x_1 ... x_n$ having Markov type $m_n$ such that $V_P(m_n) \in \{0, 1\}$ and $p_{x_n x_1} = 0$ if $V_P(m_n) = 1$. Also, as $P_n \ll Q_n$, $V_Q(m_n) \in \{0, 1\}$ and $q_{x_n x_1} = 0$ if $V_Q(m_n) = 1$. These conditions ensure that all variables used are well-defined.

Next we present Lemmas 4(a), 4(b), 5(a) and 5(b) which are directly based on [6]. Let $\alpha_{1,P} = \min_j p_j(0)$, $\alpha_{2,P} = \max_j p_j(0)$ and $\beta_{1,P} = \min_{\{(j,k): p_{jk} \neq 0\}} p_{jk}$, $\beta_{2,P} = \max_{(j,k)} p_{jk}$.

**Lemma 4(a) :** The probability of a sequence $x_1^n = x_1...x_n$ where $p_{x_n x_1} > 0$, characterized by empirical matrix $\Phi(x_1^n)$ and probability distribution $F$ satisfies

$$\frac{\alpha_{1,P}}{\beta_{2,P}} 2^{-n(H(\Phi|F)+D(\Phi\|\Gamma_P|F))} \leq P(x_1^n) \leq \frac{\alpha_{2,P}}{\beta_{1,P}} 2^{-n(H(\Phi|F)+D(\Phi\|\Gamma_P|F))}$$

**Proof:** From Eq. 9 of [6],

$$P(x_1^n) = \frac{p_{x_1}(0)}{p_{x_n x_1}} 2^{-n(H(\Phi|F)+D(\Phi\|\Gamma_P|F))}$$

Result follows from definition of $\alpha_{1,P}, \alpha_{2,P}, \beta_{1,P}, \beta_{2,P}$. □

**Lemma 4(b) :** If $p_{x_n x_1} = 0$ and $n$ is larger than $r = 2^{K^2}$ then it is possible to replace the last $(r-1)$ letters $x_{n-r+2}...x_n$ of $x_1^n$ by $b_1...b_{r-1}$ such that $P(x_1|x_{n-r+1}b_1...b_{r-1}) > 0$ (refer to [6]). Set $y_1^n = x_1...x_{n-r+1}b_1...b_{r-1}$. Then,

$$P(y_1^n)\left(\frac{\beta_{1,P}}{\beta_{2,P}}\right)^{r-1} \leq P(x_1^n) \leq P(y_1^n)\left(\frac{\beta_{2,P}}{\beta_{1,P}}\right)^{r-1}$$

**Proof:** Follows from the construction and the definition of $\beta_{1,P}$ and $\beta_{2,P}$. □

**Lemma 5(a) [6]:** The probability of a Markov type $m_n$ induced by a $n$-long sequence $x_1^n = x_1...x_n$ with $p_{x_n x_1} > 0$ satisfies,

$$n^{-K}(n+1)^{-K^2}\alpha_{1,P} 2^{-nD(\Phi\|\Gamma_P|F)} \leq P(m_n) \leq K\beta_{1,P}^{-1} 2^{-nD(\Phi\|\Gamma_P|F)}$$

**Lemma 5(b) [6]:** If $p_{ts} = 0$ it is possible to construct a new sequence having the Markov type $m_n^*$ by the method of Lemma 4(b) such that there exists $\sigma_n > 0$ independent of $m_n$ and $m_n^*$ with $\lim_{n\to\infty} \sigma_n = 0$ so that,

$$P(m_n) \leq 2^{n\sigma_n} P(m_n^*)$$

Next we develop a statement similar to Definition 1 in this context where $P$ and $Q$ are distinct.

**Lemma 6:** Consider irreducible and aperiodic Markov sources $X$ and $Y$ distributed according to the measures $P$ and $Q$ respectively with $P_n \ll Q_n$. Let $\gamma_{1,Q} = \frac{\alpha_{1,Q}}{\beta_{2,Q}}\left(\frac{\beta_{1,Q}}{\beta_{2,Q}}\right)^{r-1}$ and $\gamma_{2,Q} = \frac{\alpha_{2,Q}}{\beta_{1,Q}}\left(\frac{\beta_{2,Q}}{\beta_{1,Q}}\right)^{r-1}$. Then for $\epsilon > 0$,

$$P(\{(x_1^n) : \gamma_{1,Q} 2^{-n(H(P)+D(P\|Q)+\epsilon)} \leq Q(x_1^n) \leq \gamma_{2,Q} 2^{-n(H(P)+D(P\|Q)-\epsilon)}\}) \geq 1 - e^{-k(\epsilon)n}$$

where $k(\epsilon) > 0\ \forall \epsilon > 0, k(0) = 0$.

**Proof:** Let $\chi_n(\epsilon)$ be the class of all Markov types such that

$$|(H(\Phi_{m_n}|F_{m_n}) + D(\Phi_{m_n} \parallel \Gamma_Q|F_{m_n})) - (H(P) + D(P \parallel Q))| \leq \epsilon \quad \text{if} \quad V_Q(m_n) = 0 \tag{15}$$

$$|(H(\Phi_{m_n^*}|F_{m_n^*}) + D(\Phi_{m_n^*} \parallel \Gamma_Q|F_{m_n^*})) - (H(P) + D(P \parallel Q))| \leq \epsilon \quad \text{if} \quad V_Q(m_n) = 1 \tag{16}$$

where if $V_Q(m_n) = 1$, for every sequence $x_1^n \in m_n$ there exists a sequence obtained by the same construction as in Lemma 4(b) belonging to a Markov type $m_n^*$ such that $m_n^*$ satisfies Eq. 16. Let $B_n(\epsilon)$ be a set of $n$-long sequences defined as,

$$B_n(\epsilon) = \{x_1^n : \gamma_{1,Q} 2^{-n(H(P)+D(P\|Q)+\epsilon)} \leq Q(x_1^n) \leq \gamma_{2,Q} 2^{-n(H(P)+D(P\|Q)-\epsilon)}\} \tag{17}$$

From Lemma 4(a), the probability $Q(x_1^n)$ of any sequence $x_1^n = x_1...x_n$ such that $q_{x_n x_1} \neq 0$ having Markov type $m_n \in \chi_n(\epsilon)$ is bounded by,

$$\frac{\alpha_{1,Q}}{\beta_{2,Q}} 2^{-n(H(\Phi_{m_n}|F_{m_n})+D(\Phi_{m_n}\|\Gamma_Q|F_{m_n}))} \leq Q(x_1^n) \leq \frac{\alpha_{2,Q}}{\beta_{1,Q}} 2^{-n(H(\Phi_{m_n}|F_{m_n})+D(\Phi_{m_n}\|\Gamma_Q|F_{m_n}))} \tag{18}$$

Also from Lemma 4(a), Lemma 4(b) and the definition of $\gamma_{1,Q}$ and $\gamma_{2,Q}$, the probability $Q(x_1^n)$ of any sequence $x_1^n = x_1...x_n$ such that $q_{x_n x_1} = 0$ having Markov type $m_n \in \chi_n(\epsilon)$, is bounded by,

$$\gamma_{1,Q} 2^{-n(H(\Phi_{m_n^*}|F_{m_n^*})+D(\Phi_{m_n^*}\|\Gamma_Q|F_{m_n^*}))} \leq Q(x_1^n) \leq \gamma_{2,Q} 2^{-n(H(\Phi_{m_n^*}|F_{m_n^*})+D(\Phi_{m_n^*}\|\Gamma_Q|F_{m_n^*}))} \tag{19}$$

Hence if $m_n \in \chi_n(\epsilon)$, $x_1^n \in B_n(\epsilon)$ for all $x_1^n$ having Markov type $m_n$. Let us calculate the probability of $\tau_n(\epsilon) = \Omega^n - \chi_n(\epsilon)$, where $\Omega^n$ is the class of all Markov types of length $n$. Let $\eta_n(\epsilon)$ be the subset of $\tau_n(\epsilon)$ with $V_P(m_n) = 0$ for $m_n \in \eta_n(\epsilon)$, $\rho_n(\epsilon)$ be the subset of $\tau_n(\epsilon)$ with $V_P(m_n) = 1$ for $m_n \in \rho_n(\epsilon)$. With slight abuse of notation, we denote $\tau_n(\epsilon)$ by $\tau_n$, $\eta_n(\epsilon)$ by $\eta_n$ and $\rho_n(\epsilon)$ by $\rho_n$. As the number of Markov types of order $n$ is bounded by $(n+1)^{K^2}$,

$$P(\tau_n) \leq (n+1)^{K^2} \max_{m_n \in \tau_n} P(m_n) \tag{20}$$

If the maximization in Eq. 20 is achieved for a Markov type $M_n$ such that $V_P(M_n) = 0$ then we can write,

$$\begin{aligned} P(\tau_n) &\leq (n+1)^{K^2} P(M_n) \\ &\overset{(a)}{\leq} (n+1)^{K^2} K\beta_{1,P}^{-1} 2^{-nD(\Phi_{M_n}\|\Gamma_P|F_{M_n})} \\ &= (n+1)^{K^2} K\beta_{1,P}^{-1} 2^{-n \min_{m_n\in\eta_n} D(\Phi_{m_n}\|\Gamma_P|F_{m_n})} \end{aligned} \tag{21}$$

Here (a) follows from Lemma 5(a). If the maximization in Eq. 20 is achieved for a Markov type $M_n$ such that $V_P(M_n) = 1$ then we can write,

$$\begin{aligned} P(\tau_n) &\leq (n+1)^{K^2} P(M_n) \\ &\overset{(b)}{\leq} (n+1)^{K^2} 2^{n\sigma_n} P(M_n^*) \\ &\overset{(c)}{\leq} (n+1)^{K^2} 2^{n\sigma_n} K\beta_{1,P}^{-1} 2^{-nD(\Phi_{M_n^*}\|\Gamma_P|F_{M_n^*})} \\ &\overset{(d)}{\leq} (n+1)^{K^2} 2^{n\sigma_n} K\beta_{1,P}^{-1} 2^{-n \min_{m_n\in\eta_n} D(\Phi_{m_n}\|\Gamma_P|F_{m_n})} \end{aligned} \tag{22}$$

In (b) we use Lemma 5(b) choosing $M_n^*$ corresponding to a sequence $x_1^n \in M_n$ such that no Markov type $M_n^*$ satisfying Eq. 16 can be constructed for $x_1^n$. (c) follows from Lemma 5(a) as $V_Q(M_n^*) = 0$. (d) follows as $M_n^* \in \eta_n$. Therefore, in general we can write,

$$P(\tau_n) \leq (n+1)^{K^2} 2^{n\sigma_n} K\beta_{1,P}^{-1} 2^{-n \min\limits_{m_n \in \eta_n} D(\Phi_{m_n} \| \Gamma_P | F_{m_n})}$$

From definition, it follows that

$$\begin{aligned} &|(H(\Phi_{m_n}|F_{m_n}) + D(\Phi_{m_n} \parallel \Gamma_Q|F_{m_n})) \\ &- (H(P) + D(P \parallel Q))| > \epsilon \quad \forall \quad m_n \in \eta_n \end{aligned} \tag{23}$$

By continuity of $H(P)$ and $D(P \parallel Q)$ as functions of $P$ and $Q$ because we are dealing with *finite* alphabets and Markov setup, it follows that

$$\begin{aligned} &|(H(\Phi_{m_n}|F_{m_n}) + D(\Phi_{m_n} \parallel \Gamma_Q|F_{m_n})) \\ &- (H(P) + D(P \parallel Q))| > \epsilon \implies D(\Phi_{m_n} \parallel \Gamma_P|F_{m_n}) > h(\epsilon) \end{aligned} \tag{24}$$

where $h(\epsilon) > 0$ for $\epsilon > 0$. Hence we get,

$$\begin{aligned} P(\tau_n) &\leq (n+1)^{K^2} 2^{n\sigma_n} K\beta_{1,P}^{-1} 2^{-nh(\epsilon)} \\ &= 2^{-n(h(\epsilon) - \sigma_n - \frac{1}{n}(K^2 \log(n+1) + \log(K\beta_{1,P}^{-1}))} \\ &\leq 2^{-ng(\epsilon)} \qquad (\because \lim_{n\to\infty} \sigma_n = 0) \\ &= e^{-nk(\epsilon)} \\ \implies & P(\chi_n(\epsilon)) \geq 1 - e^{-nk(\epsilon)} \\ \implies & P(B_n(\epsilon)) \geq 1 - e^{-nk(\epsilon)} \end{aligned} \tag{25}$$

where $k(\epsilon), g(\epsilon) > 0 \; \forall \; \epsilon > 0$, $k(0) = g(0) = 0$. □

By imitating the proofs of Lemmas 2 and 3 while making use of Lemma 6 instead of Definition 1, we can obtain Lemmas 7 and 8 respectively.

**Lemma 7:** For irreducible and aperiodic Markov sources $X$ and $Y$ distributed according to the measures $P$ and $Q$ respectively with $P_n \ll Q_n$,

$$\mathbb{P}(\frac{\log W_n}{n} > H(P) + D(P \parallel Q) + \epsilon) \leq e^{-f(\epsilon)n}$$

for all $n \geq N_1(\epsilon)$, where $f(\epsilon) > 0 \; \forall \; \epsilon > 0$ and $f(0) = 0$.

**Lemma 8:** For irreducible and aperiodic Markov sources $X$ and $Y$ distributed according to the measures $P$ and $Q$ respectively with $P_n \ll Q_n$,

$$\mathbb{P}(\frac{\log W_n}{n} < H(P) + D(P \parallel Q) - \epsilon) \leq e^{-g(\epsilon)n}$$

for all $n \geq N_2(\epsilon)$, where $g(\epsilon) > 0 \; \forall \; \epsilon > 0$ and $g(0) = 0$.

Combining Lemma 7 and Lemma 8 we have

**Theorem 2:** For irreducible and aperiodic Markov sources $X$ and $Y$ distributed according to the measures $P$ and $Q$ respectively with $P_n \ll Q_n$,

$$\mathbb{P}(|\frac{\log W_n}{n} - (H(P) + D(P \parallel Q))| > \epsilon) \leq 2e^{-h(\epsilon)n}$$

for all $n \geq N(\epsilon)$, where $h(\epsilon) = \min(f(\epsilon), g(\epsilon))$ and $N(\epsilon) = \max\{N_1(\epsilon), N_2(\epsilon)\}$.

*Hence for irreducible and aperiodic Markov sources $X$ and $Y$ distributed according to the measures $P$ and $Q$ respectively with $P_n \ll Q_n$ the normalized waiting time obeys the large deviation property.*

**Remark:** A statement on the exponential convergence of $\dfrac{\log m}{L_m} \to H(P) + D(P \parallel Q)$ as $\log m \to \infty$ under the same constraints and an analogous result for Theorem 1, can be easily established. The reader can refer to the technique used in Corollary 2 and Corollary 3 of [9] for more details.

## V. Conclusion

In this work, we studied the asymptotic behavior of the normalized waiting times. We established the large deviation property of the normalized waiting times between independent realizations of a certain class of mixing processes and extended the result to realizations of two distinct irreducible and aperiodic Markov sources using the method of Markov types. Using ideas from [10] and [11], a relative entropy estimator for the change point detection problem can be developed and can be shown to converge exponentially using the large deviation property. This result aids the understanding of the performance of idealized (referring to infinite database) Lempel-Ziv Coding and can also be applied to classify a new DNA template. [12]

## References


[1] A. D. Wyner and J. Ziv, "Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data compression," *Information Theory, IEEE Transactions on*, vol. 35, no. 6, pp. 1250–1258, 1989.

[2] P. C. Shields, "Waiting times: positive and negative results on the wyner-ziv problem," *Journal of Theoretical Probability*, vol. 6, no. 3, pp. 499–519, 1993.

[3] T. Luczak and W. Szpankowski, "A suboptimal lossy data compression based on approximate pattern matching," *Information Theory, IEEE Transactions on*, vol. 43, no. 5, pp. 1439–1451, 1997.

[4] I. Kontoyiannis, "Asymptotic recurrence and waiting times for stationary processes," *Journal of Theoretical Probability*, vol. 11, no. 3, pp. 795–811, 1998.

[5] M. Abadi, "Sharp error terms and neccesary conditions for exponential hitting times in mixing processes," *Annals of probability*, pp. 243–264, 2004.

[6] L. Davisson, G. Longo, and A. Sgarro, "The error exponent for the noiseless encoding of finite ergodic markov sources," *Information Theory, IEEE Transactions on*, vol. 27, no. 4, pp. 431–438, 1981.

[7] D. H. Kim, "Convergence of recurrence of blocks for mixing processes," *Osaka Journal of Mathematics*, vol. 49, no. 1, pp. 1–20, 2012.

[8] P. Shields, "The ergodic theory of discrete sample paths," 1996.

[9] S. Jain and R. K. Bansal, "On large deviation property of recurrence times," in *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*, 2013, pp. 2880–2884.

[10] D. R. Juvvadi and R. K. Bansal, "Sequential change detection using estimators of entropy amp; divergence rate," in *Communications (NCC), 2013 National Conference on*, 2013, pp. 1–5.

[11] T. Jacob and R. Bansal, "Sequential change detection based on universal compression algorithms," in *Information Theory, 2008. ISIT 2008. IEEE International Symposium on*, 2008, pp. 1108–1112.

[12] A. Dembo and I. Kontoyiannis, "Source coding, large deviations, and approximate pattern matching," *Information Theory, IEEE Transactions on*, vol. 48, no. 6, pp. 1590–1615, 2002.