

Large deviations for a class of nonhomogeneous Markov chains: K -word level results

Zach Dietz¹ and Sunder Sethuraman²

March 21, 2005

Abstract

In previous work, Dietz and Sethuraman (2005), large deviations with respect to additive functionals were established for a class of finite-state time-nonhomogeneous Markov chains whose connecting transition matrices $\{P_n\}$ converge to a general limit matrix P which includes some stochastic optimization algorithms. In this note, large deviations at the next level, that is with respect to K -word empirical measures for $K \geq 1$, are established. The rate functions found connect the “ K -word landscape” to features of the base convergence $P_n \rightarrow P$ in terms of an optimization over certain “routing” and “resting” costs which gives insight into how deviations are achieved.

Research supported in part by NSA grant H982300510041.

Key words and phrases: large deviations, nonhomogeneous, Markov, K -word.

Abbreviated title: LDP for some nonhomogeneous Markov chains II.

AMS (2000) subject classifications: Primary 60J10; secondary 60F10.

¹ Department of Mathematics, Tulane University, 6823 St. Charles Ave., New Orleans, LA 70118; zdietz@math.tulane.edu.

² Department of Mathematics, Iowa State University, 396 Carver Hall, Ames, IA 50011; sethuram@iastate.edu.

1 Introduction and Results

Recently, some large deviations principles and bounds were proved with respect to additive functionals for a class of finite state time-nonhomogeneous Markov chains [2]. The purpose of this note is to extend these results to K -word empirical distributions (Theorem 1.2) with a view toward applications, and also to explore more conceptually the notion of “entropy” in these models (cf. Ch. 3, 6 [1]). The rate functions found connect certain “ K -word landscapes” to the 1-word or “scalar” setting, and describe an optimization over certain “routing” and “resting costs” which, as in the “scalar” case

in [2], illuminates how deviations typically arise, and helps “categorize” different large deviation behaviors depending on the “strength” of nonhomogeneity in the system.

We recall now the setting in [2]. Let $\Sigma = \{1, 2, \dots, \tau\}$ be a finite space of $\tau \geq 1$ states, and let π and $P_n = \{p_n(i, j)\}$ for $n \geq 1$ be a distribution and a sequence of stochastic matrices on Σ . Define $\mathbb{P}_\pi = \mathbb{P}_\pi^{\{P_n\}}$ as the (nonhomogeneous) Markov measure on the sequence space Σ^∞ with Borel sets $\mathcal{B}(\Sigma^\infty)$ corresponding to initial distribution π and transition kernels $\{P_n\}$. That is, with respect to the coordinate process, X_0, X_1, \dots , the Markov property holds,

$$\mathbb{P}_\pi(X_{n+1} = j | X_0, X_1, \dots, X_{n-1}, X_n = i) = p_{n+1}(i, j)$$

for all $i, j \in \Sigma$ and $n \geq 0$. In this context, P_{n+1} controls “transitions” between times n and $n + 1$.

Let now $P = \{p(i, j)\}$ be a general stochastic matrix on Σ , and consider the collection $\mathbb{A}(P) = \{\mathbb{P}_\pi^{\{P_n\}} : P_n \rightarrow P\}$ where the convergence $P_n \rightarrow P$ is elementwise, that is $\lim_{n \rightarrow \infty} p_n(i, j) = p(i, j)$ for all $i, j \in \Sigma$. The collection \mathbb{A} can be thought of as perturbations of the time-homogeneous Markov chain run with P , and is a natural class in which to study how “nonhomogeneity” enters the large deviation picture. A basic question is whether and in what sense the large deviations differ from that under the P -time-homogeneous chain. The class \mathbb{A} includes many stochastic optimizations which involve reducible limits P such as simulated annealing and Metropolis procedures. For instance, in the Metropolis algorithm,

$$P_n(i, j) = \begin{cases} g(i, j) \exp\{-\beta_n(H(j) - H(i))_+\} & \text{for } j \neq i \\ 1 - \sum_{l \neq i} P_n(i, l) & \text{for } j = i \end{cases}$$

where g is an irreducible transition function, and β_n represents an “inverse temperature” parameter which diverges, $\beta_n \rightarrow \infty$. Here, the limit kernel $P = \lim_n P_n$ corresponds to steepest descent or “greedy” dynamics in that jumps from i to j when $H(j) > H(i)$ are not allowed. We refer to [2] for more discussion and associated references.

Let now $K \geq 1$ be an integer. The K -word empirical process lives on the space of K -tuples or K -words Σ^K and are the induced distributions of $\{\mathcal{L}_n^{(K)} : n \geq 1\}$ on Σ^K with respect to \mathbb{P}_π where

$$\mathcal{L}_n^{(K)} = \frac{1}{n} \sum_{i=1}^n \delta_{\langle X_i, \dots, X_{i+K-1} \rangle}$$

where $\delta_a(\cdot)$ is the indicator of a . Let us enumerate $\Sigma^K = \{\omega_1, \dots, \omega_{\tau^K}\}$, and let $\Omega^{(K)}$ be the collection of probability vectors on Σ^K . There is a 1 – 1 correspondence of $\mathcal{L}_n^{(K)}$ with an element in $\Omega^{(K)}$, namely the “vector form” of $\mathcal{L}_n^{(K)}$,

$$Z_n^{(K)} = \langle \mathcal{L}_n^{(K)}(\omega_i) : \omega_i \in \Sigma^K \rangle.$$

In turn, $Z_n^{(K)}$ can be identified with the following additive sum. That is, let $f^{(K)} : \Sigma^K \rightarrow \Omega^{(K)}$ be defined by

$$f^{(K)}(\vec{x}) = \langle 1_{\omega_1}(\vec{x}), \dots, 1_{\omega_{\tau^K}}(\vec{x}) \rangle.$$

Then,

$$Z_n^{(K)} = \frac{1}{n} \sum_{i=1}^n f^{(K)}(\langle X_i, \dots, X_{i+K-1} \rangle).$$

We now discuss the underlying K -word process $X_n^{(K)} = \langle X_n, \dots, X_{n+K-1} \rangle$ for $n \geq 0$ on Σ^K . It is simple to verify that this process is a nonhomogeneous Markov chain associated with transition kernels $P_n^{(K)} = \{p_n^{(K)}(\vec{x}, \vec{y})\}$ and initial distribution $\pi^{(K)}$ on Σ^K given by

$$p_n^{(K)}(\vec{x}, \vec{y}) = p_{n+K-1}(x_K, y_K) \prod_{i=1}^{K-1} \delta_{x_{i+1}}(y_i) \quad \text{and} \quad \pi^{(K)}(\vec{x}) = \mathbb{P}_\pi(\langle X_0, \dots, X_{K-1} \rangle = \vec{x}).$$

The matrices $\{P_n^{(K)}\}$ converge to the limit $P^{(K)} = \{p^{(K)}(\vec{x}, \vec{y})\}$ where

$$p^{(K)}(\vec{x}, \vec{y}) = p(x_K, y_K) \prod_{i=1}^{K-1} \delta_{x_{i+1}}(y_i).$$

Therefore, the process measure $\mathbb{P}_{\pi^{(K)}}^{(K)} = \mathbb{P}_{\pi^{(K)}}^{\{P_n^{(K)}\}}$ with respect to $\pi^{(K)}$ and connecting K -word transition matrices $\{P_n^{(K)}\}$ belongs to $\mathbb{A}(P^{(K)})$. By 1 – 1 correspondence, results on $\{Z_n^{(K)}\}$ under $\mathbb{P}_{\pi^{(K)}}^{(K)}$ translate to results for $\{\mathcal{L}_n^{(K)}\}$ under \mathbb{P}_π . Of course, when $K = 1$, $\mathbb{P}_{\pi^{(1)}}^{(1)} = \mathbb{P}_\pi$ and so in this case we will suppress the superscript “(1).”

We discuss now some results in [2]. Let $d \geq 1$ be an integer and $h : \Sigma \rightarrow \mathbb{R}^d$ be a function. Let also $Z_n(h) = \sum_{i=1}^n h(X_i)$ for $n \geq 1$. Under an “initial ergodicity” Condition (SIE-1), and a combination of regularity Assumptions A,B and C on the initial measure π and approach $P_n \rightarrow P$, a large deviation principle for $\{Z_n(h)\}$ under \mathbb{P}_π was proved with respect to a certain rate function \mathbb{J} in Theorem 3.3 [2]. Also, depending on the decay of certain “ P_n -connection” probabilities which “bridge” P -reducible and other sets, three regimes of large deviation behavior were identified in Corollary 3.1 [2]. Roughly, given that the limit P is reducible, if the decay is too fast, too slow, or in an intermediate range, the large deviations are the same as under the time-homogeneous chain run with P , trivial, or non-trivial and involve the decay rates in terms of “routing costs.” We also remark some systems–“geometric” cooling algorithms and glassy physics dynamics–in which the “intermediate” speed results apply are mentioned in [2].

Certainly, by setting $d = \tau^K$ and $h = f^{(K)}$, and interpreting Conditions and Assumptions (SIE-1) and A,B and C in terms of $\{P_n^{(K)}\}$ and $\pi^{(K)}$ in place of $\{P_n\}$ and

π , we do obtain a large deviation principle for $\{Z_n^{(K)}\}$ under $\mathbb{P}_{\pi^{(K)}}^{(K)}$ and in turn the empirical measures $\{\mathcal{L}_n^{(K)}\}$ under \mathbb{P}_π as a simple application of [2]. However, for such a result to be useful, the relations between Condition (SIE-1), Assumptions A,B and C, and rate function \mathbb{J} with respect to $h = f^{(K)}$, $\{P_n^{(K)}\}$ and $\pi^{(K)}$, and the base measure π and base convergence $P_n \rightarrow P$ should be understood. In particular, how does the K -word process “landscape” relate to the 1-word or “scalar” process? In fact, the key contribution of this article is to make concrete these connections (Proposition 1.1), and to write the rate function in terms of “routing” and “resting” costs with respect to the “scalar” convergence $P_n \rightarrow P$ (cf. near Theorem 1.2).

We now state explicitly these conditions, assumptions, rate function \mathbb{J} and Theorem 3.3 [2]. First, we recall some terms and definitions (cf. sections 2,3 [2]).

Canonical Form for P . By reordering Σ if necessary, the stochastic matrix P may be put in form

$$P = \begin{bmatrix} U(0,0) & U(0,1) & \dots\dots & U(0,M_0) \\ 0 & S(1) & 0 & \dots & 0 \\ \vdots & 0 & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & \dots\dots\dots & 0 & S(M_0) \end{bmatrix}. \quad (1)$$

where $1 \leq M_0 \leq \mathfrak{r}$, and $S(1), \dots, S(M_0)$ are stochastic irreducible submatrices corresponding to disjoint subsets of recurrent states, and submatrices $U(0,0), \dots, U(0,M_0)$ correspond to transient states when they exist.

When there are transient states, the square block $U(0,0)$ itself may be decomposed as (cf. section 1.2 [3])

$$U(0,0) = \begin{bmatrix} R(1) & V(1,2) & \dots\dots\dots & V(1,N_0) \\ 0 & R(2) & V(2,3) & \dots & V(2,N_0) \\ \vdots & 0 & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & \dots\dots\dots & 0 & R(N_0) \end{bmatrix}$$

where $1 \leq N_0 \leq \mathfrak{r}-1$, and $R(i)$ is either the 1×1 zero matrix or an irreducible submatrix corresponding to a subset of transient states for $1 \leq i \leq N_0$. The $R(i) = [0]$ matrices are called degenerate transient, and the irreducible $R(i)$ are named *nondegenerate transient* since respectively returns to corresponding states are impossible, and possible under the time-homogeneous chain run with P .

We now insert the form for $U(0,0)$ into (1). When there are transient states, let $P(i) = R(i)$ for $1 \leq i \leq N_0$, and $P(i) = S(i - N_0)$ for $N_0 + 1 \leq i \leq N_0 + M_0$. When all states are recurrent, let $P(i) = S(i)$ for $1 \leq i \leq M_0$. Also, in the following, let $T(i, j)$ for $i < j$ denote the appropriate “connecting” submatrix $U(\cdot, \cdot)$ or $V(\cdot, \cdot)$. We remark that $T(i, j)$ is a matrix of zeroes for $N_0 + 1 \leq i < j \leq N_0 + M_0$.

The canonical decomposition of P then is the following:

$$P = \begin{bmatrix} P(1) & T(1,2) & \dots\dots\dots & T(1,N+M) \\ 0 & P(2) & T(2,3) & \dots & T(2,N+M) \\ \vdots & 0 & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & \dots\dots\dots & 0 & P(N+M) \end{bmatrix}.$$

Let $C_i \subset \Sigma$ be the subset which corresponds to $P(i)$ so that $P(i) = \{p(x, y) : x, y \in C_i\}$ for $1 \leq i \leq N + M$. Define also the sets

$$\begin{aligned} \mathcal{D} &= \{i : P(i) \text{ degenerate transient}\} \\ \mathcal{N} &= \{i : P(i) \text{ nondegenerate transient}\} \\ \mathcal{M} &= \{i : P(i) \text{ stochastic}\} \\ \mathcal{G} &= \mathcal{N} \cup \mathcal{M} (= \{i : P(i) \text{ nondegenerate transient or stochastic}\}) \end{aligned}$$

and let $N = |\mathcal{D}|$, $M = |\mathcal{G}|$. It will also be convenient to enumerate $\mathcal{G} = \{\zeta_1, \zeta_2, \dots, \zeta_M\}$.

Resting Costs. For $i \in \mathcal{G}$ and $\lambda \in \mathbb{R}^d$, let $\rho(C_i, \lambda, h)$ be the Perron-Frobenius eigenvalue of the tilted matrix $\Pi_{C_i, \lambda, h} = \{p(x, y)e^{\lambda h(y)} : x, y \in C_i\}$, and define $\mathbb{I}_i : \mathbb{R}^d \rightarrow [0, \infty]$ by

$$\mathbb{I}_i(x) = \sup_{\lambda \in \mathbb{R}^d} \{\lambda x - \log \rho(C_i, \lambda, h)\}. \quad (2)$$

From Proposition 2.1 [2], the function \mathbb{I}_i is the rate function for $\{Z_n(h)\}$ with respect to the time-homogeneous process run under P restricted to C_i , and represents a certain “resting” cost of being in C_i .

Routing Costs. When $N + M \geq 2$, define, for distinct $1 \leq i, j \leq N + M$,

$$t(n, (i, j)) = \max_{x \in C_i, y \in C_j} p_n(x, y).$$

Also, for $0 \leq k \leq N + M - 2$, let $L_k = \langle i = l_0, l_1, \dots, l_k, l_{k+1} = j \rangle$ be a $k + 2$ -tuple of distinct indices in $\{1, \dots, N + M\}$, and define the “routing” cost matrix \mathcal{U}_0 by

$$\mathcal{U}_0(i, j) = \max_{0 \leq k \leq N+M-2} \max_{L_k} \sum_{s=0}^k \overline{\lim}_{n \rightarrow \infty} \frac{1}{n} \log t(n, (l_s, l_{s+1})). \quad (3)$$

Rate Function \mathbb{J} . Let \mathbb{S}_M and Ω_M be the set of permutations, and collection of probability vectors on $\{1, 2, \dots, M\}$. For $\vec{v} \in \Omega_M$ and $z \in \mathbb{R}^d$, define the set of convex combinations, $D(M, \vec{v}, z) = \{\vec{x} = \langle x_1, \dots, x_M \rangle \in (\mathbb{R}^d)^M : \sum_{i=1}^M v_i x_i = z\}$. Then, for $\sigma \in \mathbb{S}_M$, $\vec{v} \in \Omega_M$, $\vec{x} \in (\mathbb{R}^d)^M$, and $z \in \mathbb{R}^d$, define the extended functions

$$C_{\vec{v}}(\sigma, \vec{x}) = \begin{cases} -\sum_{i=1}^{M-1} (\sum_{j=1}^i v_j) \mathcal{U}_0(\zeta_{\sigma(i)}, \zeta_{\sigma(i+1)}) + \sum_{i=1}^M v_i \mathbb{I}_{\zeta_{\sigma(i)}}(x_i) & \text{when } M \geq 2 \\ \mathbb{I}_{\zeta_1}(x_1) & \text{when } M = 1 \end{cases}$$

and

$$\mathbb{J}(z) = \inf_{\vec{v} \in \Omega_M} \inf_{\vec{x} \in D(M, \vec{v}, z)} \min_{\sigma \in \mathbb{S}_M} C_{\vec{v}}(\sigma, \vec{x}).$$

It is shown in Proposition 4.1 [2] that \mathbb{J} is a good rate function. The form of \mathbb{J} suggests when $M \geq 2$ that $Z_n(h)$ deviates to z when X_n typically visits sets $\{C_i : i \in \mathcal{G}\}$ in a certain order σ according to certain time-lengths \vec{v} so that the average z is achieved and “resting” and “routing” costs are minimized.

“*Initial Ergodicity*” Condition. The following condition prevents “blocking.”

Condition (SIE-1). Suppose $\pi(C_i) > 0$ for $i \in \mathcal{G}$, and when $p(s, t) > 0$ for $s, t \in C_i$ and $i \in \mathcal{G}$ then also $p_n(s, t) > 0$ for $n \geq 1$.

“*Regularity*” Assumptions on Convergence $P_n \rightarrow P$. Assumption A specifies that maximal “connection” probabilities in the “ $(1/n) \log$ ” sense have limits. Assumption B states that $\underline{\lim}_{n \rightarrow \infty} \frac{1}{n} \log t(n, (i, j))$ can be achieved in a systematic manner. Assumption C covers the case when P governs periodic motion but the P_n approach is slow enough to give a sense of “primitivity.” We note Assumptions A and B include the case all limits exist $\lim (1/n) \log p_n(x, y)$ for $x \in C_i, y \in C_j$ and distinct $1 \leq i, j \leq N + M$; also Assumption C includes the case that all $P(i)$ for $i \in \mathcal{G}$ are positive submatrices (cf. Proposition 3.1 [2]). See [2] for more discussion and some counter-examples.

Assumption A. Suppose $\lim_{n \rightarrow \infty} \frac{1}{n} \log t(n, (i, j))$ exists in the extended sense for all distinct $1 \leq i, j \leq N + M$.

Assumption B. Suppose for all distinct $1 \leq i, j \leq N + M$ there exists an element $a = a(i, j) \in C_i$ and a sequence $\{b_n = b_n(i, j)\} \subset C_j$ such that

$$\underline{\lim}_{n \rightarrow \infty} \frac{1}{n} \log t(n, (i, j)) = \lim_{n \rightarrow \infty} \frac{1}{n} \log p_n(a, b_n)$$

that is $\underline{\lim}_{n \rightarrow \infty} (1/n) \log t(n, (i, j))$ is achieved on a fixed departing point $a \in C_i$.

Assumption C. Define $P^*(i) = \{p^*(s, t) : s, t \in C_i\}$ by

$$p^*(s, t) = \begin{cases} p(s, t) & \text{when } p(s, t) > 0 \\ 1 & \text{when } \underline{\lim} (1/n) \log p_n(s, t) = 0 \text{ and } p(s, t) = 0 \\ 0 & \text{otherwise} \end{cases}$$

Suppose that $P^*(i)$ is primitive for $i \in \mathcal{G}$, that is for some power $k \geq 1$, $(P^*(i))^k$ is composed of all positive entries.

We now state Theorem 3.3 in [2].

Theorem 1.1 *Suppose $\{P_n\}$ and π satisfy Condition (SIE-1) and Assumption A, and also either Assumptions B or C. Then, with respect to good rate function \mathbb{J} , and Borel sets $\Gamma \subset \mathbb{R}^d$, we have*

$$\begin{aligned} - \inf_{z \in \Gamma^o} \mathbb{J}(z) &\leq \underline{\lim}_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}_\pi(Z_n(h) \in \Gamma^o) \\ &\leq \overline{\lim}_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}_\pi(Z_n(h) \in \bar{\Gamma}) \leq - \inf_{z \in \bar{\Gamma}} \mathbb{J}(z). \end{aligned}$$

We remark, following Corollary 3.1 [2], in the case $M \geq 2$, when “routing cost” $\mathcal{U}_0 \equiv -\infty$ or $\mathcal{U}_0 \equiv 0$, that is when the P_n -connection probabilities decay “too fast” or “too slow,” the rate function \mathbb{J} reduces to that under the time-homogeneous chain run with P or a trivial one. When $-\infty < \mathcal{U}_0 < 0$, \mathbb{J} nontrivially includes the values \mathcal{U}_0 .

To extend Theorem 1.1 to the K -word process, we now “lift” to the K -word level several of the “scalar” or $K = 1$ definitions and quantities.

K-word Canonical Form. We put $P^{(K)}$ in canonical form

$$P^{(K)} = \begin{bmatrix} P^{(K)}(1) & T^{(K)}(1,2) & \dots\dots\dots & T^{(K)}(1, R_0^{(K)}) \\ 0 & P^{(K)}(2) & T^{(K)}(2,3) & \dots & T^{(K)}(2, R_0^{(K)}) \\ \vdots & 0 & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & \dots\dots\dots & 0 & P^{(K)}(R_0^{(K)}) \end{bmatrix}$$

where $R_0^{(K)}$ is the total number of “degenerate” and “nondegenerate” blocks in the decomposition. Analogous to the “scalar” case $K = 1$, let $\mathcal{D}^{(K)}$, $\mathcal{N}^{(K)}$, and $\mathcal{M}^{(K)}$ denote degenerate transient, nondegenerate transient, and stochastic subblocks respectively, and $\mathcal{G}^{(K)} = \mathcal{N}^{(K)} \cup \mathcal{M}^{(K)}$. Let $C_i^{(K)} \subset \Sigma^K$ correspond to $P^{(K)}(i)$ so that $P^{(K)}(i) = \{p^{(K)}(\vec{x}, \vec{y}) : \vec{x}, \vec{y} \in C_i^{(K)}\}$ for $1 \leq i \leq R_0^{(K)}$. We say elements of $C_i^{(K)}$ are degenerate transient, or non-degenerate when $i \in \mathcal{D}^{(K)}$, or $\mathcal{G}^{(K)}$, and note, as in the “scalar” case, when $i \in \mathcal{G}^{(K)}$ that $C_i^{(K)}$ is irreducible.

Of course, when $K = 1$, $R_0^{(1)} = N + M$ and $C_i^{(1)} = C_i$ for $1 \leq i \leq N + M$. When $K \geq 2$, although the structure of $P^{(K)}$ seems involved, it turns out that the K -word non-degenerate sets $\{C_i^{(K)} : i \in \mathcal{G}^{(K)}\}$ are exactly the irreducible components of $(C_i)^K \subset \Sigma^K$ for $i \in \mathcal{G}$. For $i \in \mathcal{G}$, let

$$\bar{C}_i^{(K)} = \{\vec{x} \in (C_i)^K : p(x_1, x_2)p(x_2, x_3) \cdots p(x_{K-1}, x_K) > 0\}.$$

A simplifying case is when $P(i)$ is positive in which case $\bar{C}_i^{(K)} = (C_i)^K$.

In section 3, we prove the following “ $P^{(K)}$ -landscape” proposition.

Proposition 1.1 (I) *A K -word $\vec{x} = \langle x_1, \dots, x_K \rangle$ is degenerate transient exactly when x_s does not lead to x_t for some $1 \leq s, t \leq K$ with respect to the time-homogeneous process with transition matrix P .*

(II) *The sets $\mathcal{G}^{(K)}$ and \mathcal{G} are in 1 : 1 correspondence and $|\mathcal{G}^{(K)}| = M$, and in particular, when $K \geq 2$, for each $i \in \mathcal{G}^{(K)}$ there is a unique $j \in \mathcal{G}$ where $C_i^{(K)} = \bar{C}_j^{(K)}$.*

We now enumerate $\mathcal{G}^{(K)} = \{\zeta_1^{(K)}, \dots, \zeta_M^{(K)}\}$. Define also $\phi : \{1, \dots, R_0^{(K)}\} \rightarrow \{1, \dots, N + M\}$ as the index where for all words $\vec{x} \in C_i^{(K)}$ the last letter $x_K \in C_{\phi(i)}$. Indeed, ϕ is well defined clearly on the collection of singletons $\mathcal{D}^{(K)}$, and also on $\mathcal{G}^{(K)}$

as $C_i^{(K)} \subset (C_j)^K$ for some $j \in \mathcal{G}$, noting Proposition 1.1 (II). We also observe that ϕ restricted to $\mathcal{G}^{(K)}$ is a bijection onto \mathcal{G} , and when $K = 1$ is simply the identity function.

K-Word Resting Costs. We apply the definition (2) with $h = f^{(K)}$ and denote by $\mathbb{I}_i^{(K)} : \Omega^{(K)} \rightarrow [0, \infty]$ for $i \in \mathcal{G}^{(K)}$ the function,

$$\mathbb{I}_i^{(K)}(x) = \sup_{\lambda \in \mathbb{R}^{t^K}} \{ \langle \lambda, x \rangle - \log \rho(C_i^{(K)}, \lambda, f^{(K)}) \}.$$

As in the “scalar” case, $\mathbb{I}_i^{(K)}$ is the rate function for $\{Z_n^{(K)}\}$ with respect to the $P^{(K)}$ -time-homogeneous process restricted to $C_i^{(K)}$ or more simply to $(C_{\phi(i)})^K \supset C_i^{(K)}$. Indeed, when $K = 1$, $(C_{\phi(i)})^1 = C_i^{(1)} = C_i$; when $K \geq 2$, $C_i^{(K)} = \bar{C}_{\phi(i)}^{(K)}$ (Proposition 1.1 (II)) and any word in $(C_{\phi(i)})^K \setminus \bar{C}_{\phi(i)}^{(K)}$ is (1) $P^{(K)}$ -transient to which returns are impossible (Lemma 3.1), and (2) leads, as $C_{\phi(i)}$ is P -irreducible, to $\bar{C}_{\phi(i)}^{(K)}$.

Moreover, when $i \in \mathcal{M}^{(K)}$, $\mathbb{I}_i^{(K)}(x)$ can be cast as a relative entropy of x with respect to $P^{(K)}$ restricted to $(C_{\phi(i)})^K$ (cf. section 6.5.2 [1]). When $i \in \mathcal{N}^{(K)}$, $C_i^{(K)}$ consists of $P^{(K)}$ -transient states and $P^{(K)}(i)$ is irreducible but strictly substochastic, and so $\rho(C_i^{(K)}, 0, f^{(K)}) < 1$ and hence $\mathbb{I}_i^{(K)}$ takes strictly positive values.

K-Word Routing Costs. When $R_0^{(K)} \geq 2$, define for distinct $1 \leq i, j \leq R_0^{(K)}$ that

$$t^{(K)}(n, (i, j)) = \max_{\vec{x} \in C_i^{(K)}, \vec{y} \in C_j^{(K)}} p_n^{(K)}(\vec{x}, \vec{y}).$$

Following the definition of the “scalar” cost \mathcal{U}_0 , for $0 \leq k \leq R_0^{(K)} - 2$, let $L_k^{(K)} = \langle i = l_0, l_1, \dots, l_k, l_{k+1} = j \rangle$ be a $k + 2$ -tuple of distinct indices in $\{1, \dots, R_0^{(K)}\}$. Define

$$\mathcal{U}_0^{(K)}(i, j) = \max_{0 \leq k \leq R_0^{(K)} - 2} \max_{L_k^{(K)}} \prod_{s=0}^k \overline{\lim} \frac{1}{n} t^{(K)}(n, (l_s, l_{s+1})),$$

and also importantly the “scalar” cost associated in a sense to “last letter evolution,”

$$\tilde{\mathcal{U}}_0^{(K)}(i, j) = \mathcal{U}_0(\phi(i), \phi(j)).$$

K-Word Rate Functions. We now define a rate function $\mathbb{J}_U^{(K)}$ with respect to a non-positive cost matrix $U = \{U(i, j)\}$. For probability vector $\vec{v} \in \Omega_M$, permutation $\sigma \in \mathbb{S}_M$, $\vec{x} \in (\Omega^{(K)})^M$ and cost U , let

$$C_{\vec{v}, U}^{(K)}(\sigma, \vec{x}) = \begin{cases} - \sum_{i=1}^{M-1} (\sum_{j=1}^i v_j) U(\zeta_{\sigma(i)}^{(K)}, \zeta_{\sigma(i+1)}^{(K)}) + \sum_{i=1}^M v_i \mathbb{I}_{\zeta_{\sigma(i)}^{(K)}}^{(K)}(x_i) & \text{for } M \geq 2 \\ \mathbb{I}_{\zeta_1^{(K)}}^{(K)}(x_1) & \text{for } M = 1. \end{cases}$$

Also, for $z \in \Omega^{(K)}$, let $D^{(K)}(M, \vec{v}, z) = \{\vec{x} \in (\Omega^{(K)})^M : \sum_{i=1}^M v_i x_i = z\}$, and define $\mathbb{J}_U^{(K)} : \Omega^{(K)} \rightarrow [0, \infty]$ by

$$\mathbb{J}_U^{(K)}(z) = \inf_{\vec{v} \in \Omega_M} \inf_{\vec{x} \in D^{(K)}(M, \vec{v}, z)} \min_{\sigma \in \mathbb{S}_M} C_{\vec{v}, U}^{(K)}(\sigma, \vec{x}). \quad (4)$$

Let now $\mathbb{J}^{(K)} = \mathbb{J}_{\tilde{\mathcal{U}}_0^{(K)}}^{(K)}$. As in the “scalar” case, $\mathbb{J}^{(K)}(z)$ is a good rate function and when $M \geq 2$ represents a certain optimization over “routing costs” $\tilde{\mathcal{U}}_0^{(K)}$ and “resting costs” or relative entropies $\{\mathbb{I}_i^{(K)} : i \in \mathcal{G}^{(K)}\}$ given to K -word paths, whose empirical distribution is z , which visit sets $\{C_i^{(K)} : \mathcal{G}^{(K)}\}$ in order σ according to time-lengths \vec{v} . When $M = 1$, the K -word evolution spends most of its time in the single $P^{(K)}$ -stochastic set, and $\mathbb{J}^{(K)}$, not involving “routing costs,” reduces to $\mathbb{I}_{\zeta_1^{(K)}}$.

Our main result is the following.

Theorem 1.2 *Suppose the base sequence $\{P_n\}$ and initial distribution π satisfy Condition (SIE-1) and Assumption A, and also either Assumption B or C. Then, for Borel sets $\Gamma \subset \Omega^{(K)}$, we have*

$$\begin{aligned} - \inf_{z \in \Gamma^o} \mathbb{J}^{(K)}(z) &\leq \underline{\lim}_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}_{\pi^{(K)}}^{(K)}(Z_n^{(K)} \in \Gamma^o) \\ &\leq \overline{\lim}_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}_{\pi^{(K)}}^{(K)}(Z_n^{(K)} \in \bar{\Gamma}) \leq - \inf_{z \in \bar{\Gamma}} \mathbb{J}^{(K)}(z). \end{aligned}$$

We remark, by the form of $\mathbb{J}^{(K)}$, which involves “scalar routing cost” $\mathcal{U}_0(\phi(i), \phi(j))$, one can deduce three types of large deviation behaviors in the same way as discussed earlier after Theorem 1.1 for the “scalar” process. Namely, in the case $M \geq 2$, when $\mathcal{U}_0 \equiv -\infty$, $\mathcal{U}_0 \equiv 0$ or $-\infty < \mathcal{U}_0 < 0$, one obtains the same behavior as under the time-homogeneous chain run with $P^{(K)}$, a trivial behavior, or non-trivial deviations involving the values of \mathcal{U}_0 . We note also Condition (SIE-1) may be relaxed to a more general Condition (SIE) (cf. section 3 [2]) with some modifications.

In addition, we remark that it is an exercise to obtain process level large deviations for $\mathcal{L}_n^\infty = (1/n) \sum_{i=1}^n \delta_{\langle X_i, X_{i+1}, \dots \rangle}$ from the K -word deviations through a “projective limit approach” along the same lines as discussed in section 6.5.2 [1] for time-homogenous Markov chains.

We now comment on the structure of the article. In the next section, we outline the proof of Theorem 1.2. In section 3, the $P^{(K)}$ landscape Proposition 1.1 is proved. In sections 4 and 5, we give proof for certain steps in the outline.

2 Proof-Outline of Theorem 1.2

Step 1. The following addresses when Condition (SIE-1) holds for the K -word process and is proved in section 4.

Proposition 2.1 *Suppose $\{P_n\}$ and π satisfy (SIE-1). Then, $\{P_n^{(K)}\}$ and $\pi^{(K)}$ also satisfy (SIE-1).*

Step 2. We now extend the definition of “scalar routing” cost \mathcal{T}_1 to the K -word process (cf. near (2.10) [2]). Following the definition of $\underline{\gamma}^1(n, (i, j))$ (cf. (2.10) [2]), when $M \geq 2$, let $i, j \in \mathcal{G}^{(K)}$ be distinct, and $0 \leq k \leq R_0^{(K)} - 2$, and let $L_k = \langle i = l_0, l_1, \dots, l_k, l_{k+1} = j \rangle$ be a $k + 2$ -tuple of distinct indices in $\{1, \dots, R_0^{(K)}\}$. Let also

$$1 \leq q_0, q_{k+1} \leq \mathfrak{r}^K, \text{ and when } k > 1 \text{ and } 1 \leq s \leq k \text{ let } 1 \leq q_s \leq \mathfrak{r}^K + 1 \quad (5)$$

and call $Q_k = \langle q_0, \dots, q_{k+1} \rangle$. Let $\vec{x}^0 = \langle \vec{x}_1^0, \dots, \vec{x}_{q_0}^0 \rangle$ and $\vec{x}^{k+1} = \langle \vec{x}_1^{k+1}, \dots, \vec{x}_{q_{k+1}}^{k+1} \rangle$ be vectors in $(C_i^{(K)})^{q_0}$ and $(C_j^{(K)})^{q_{k+1}}$ respectively. For $k \geq 1$, and $0 \leq m \leq k + 1$, let $\{\vec{x}^m = \langle \vec{x}_1^m, \dots, \vec{x}_{q_m}^m \rangle\}_{m=1}^k$ be a collection such that $\vec{x}^m \in (C_{l_m}^{(K)})^{q_m}$ and let $V_k = \langle \vec{x}^0, \vec{x}^1, \dots, \vec{x}^{k+1} \rangle$. Define, for $\vec{y} \in C_i^{(K)}$ and $\vec{z} \in C_j^{(K)}$, that

$$\underline{\gamma}^{(K),1}(n, \vec{y}, \vec{z}) = \max_k \max_{L_k} \max_{Q_k} \max_{V_k} \mathbb{P}_{(n-1, \vec{y})}^{(K)}(\vec{X}_{n, n+r(k+1)+1}^{(K)} = \langle \vec{x}^0, \vec{x}^1, \dots, \vec{x}^{k+1}, \vec{z} \rangle)$$

where $r(u) = \sum_{m=0}^u q_m$ for $0 \leq u \leq k + 1$, $\vec{X}_{l,m}^{(K)} = \langle X_l^{(K)}, \dots, X_m^{(K)} \rangle$ and

$$\mathbb{P}_{(l-1, \vec{w})}^{(K)}(\vec{X}_{l,m}^{(K)} = \langle \vec{w}_0, \dots, \vec{w}_{m-l} \rangle) = \mathbb{P}_{\pi^{(K)}}^{(K)}(\vec{X}_{l,m}^{(K)} = \langle \vec{w}_0, \dots, \vec{w}_{m-l} \rangle \mid X_{l-1}^{(K)} = \vec{w})$$

for $l \leq m$ and $\vec{w}, \vec{w}_0, \dots, \vec{w}_{m-l} \in \Sigma^K$. Let also

$$\underline{\gamma}^{(K),1}(n, (i, j)) = \inf_{\vec{y} \in C_i^{(K)}, \vec{z} \in C_j^{(K)}} \underline{\gamma}^{(K),1}(n, \vec{y}, \vec{z})$$

and define $\mathcal{T}_1^{(K)}(i, j) = \underline{\lim}(1/n) \log \underline{\gamma}^{(K),1}(n, (i, j))$.

The following is Theorems 3.1 and 3.2 (i) [2] with respect to $h = f^{(K)}$ and $Z_n(h) = Z_n^{(K)}$ under $\mathbb{P}_{\pi^{(K)}}^{(K)}$.

Proposition 2.2 *When $\{P_n^{(K)}\}$ and $\pi^{(K)}$ satisfy Condition (SIE-1), we have with respect to Borel sets $\Gamma \subset \Omega^{(K)}$ that*

$$\begin{aligned} - \inf_{z \in \Gamma^o} \mathbb{J}_{\mathcal{T}_1^{(K)}}^{(K)}(z) &\leq \underline{\lim}_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}_{\pi^{(K)}}^{(K)}(Z_n^{(K)} \in \Gamma^o) \\ &\leq \overline{\lim}_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}_{\pi^{(K)}}^{(K)}(Z_n^{(K)} \in \bar{\Gamma}) \leq - \inf_{z \in \bar{\Gamma}} \mathbb{J}_{\mathcal{U}_0^{(K)}}^{(K)}(z). \end{aligned}$$

Step 3. It will now be helpful, for technical reasons, to define a new “scalar routing” cost \mathcal{T}_2 . Indeed, the only difference with $\mathcal{T}_1 = \mathcal{T}_1^{(1)}$, in the case $K = 1$, is that the vectors $\vec{x}^i = \langle x_1^i, \dots, x_{q_i}^i \rangle$ of elements in C_{l_i} for $1 \leq i \leq k$ which form V_k must now consist of *distinct* elements. Denote by $\underline{\gamma}^2(n, \vec{y}, \vec{z})$, $\underline{\gamma}^2(n, (i, j))$ and $\mathcal{T}_2(i, j) = \underline{\lim}(1/n) \log \underline{\gamma}^2(n, (i, j))$ the corresponding quantities made with this change.

As \mathcal{T}_1 is a maximization over more choices, $\mathcal{T}_1 \geq \mathcal{T}_2$. The following bounds of K -level costs in terms of “scalar” costs is proved in section 5.

Proposition 2.3 When $M \geq 2$, we have, for distinct $i, j \in \mathcal{G}^{(K)}$, that

$$(I) \mathcal{U}_0^{(K)}(i, j) \leq \mathcal{U}_0(\phi(i), \phi(j)) \quad \text{and} \quad (II) \mathcal{T}_1^{(K)}(i, j) \geq \mathcal{T}_2(\phi(i), \phi(j)).$$

Step 4. The cost \mathcal{T}_2 is only a small perturbation of \mathcal{T}_1 . In fact, the exact same proof of Proposition 4.10 [2] with respect to \mathcal{T}_1 , gives the same bounds with respect to \mathcal{T}_2 , namely under Assumptions B or C we have $\mathcal{T}_2 \geq \mathcal{T}_0$ where \mathcal{T}_0 is another “lower” cost defined as \mathcal{U}_0 (cf. (3)) except with $\underline{\lim}(1/n) \log t(n, (i, j))$ in place of $\overline{\lim}(1/n) \log t(n, (i, j))$. When in addition Assumption A holds, by definition $\mathcal{T}_0 = \mathcal{U}_0$, and the next result follows.

Proposition 2.4 When $\{P_n\}$ satisfies Assumption A, and either Assumptions B or C, and $M \geq 2$, we have for distinct $s, t \in \mathcal{G}$, that $\mathcal{T}_2(s, t) \geq \mathcal{U}_0(s, t)$.

Hence, under Assumption A and either Assumptions B or C, when $M \geq 2$, by Propositions 2.3 and 2.4, substituting into the definition of $\mathbb{J}_U^{(K)}$ (cf. (4)),

$$-\mathbb{J}_{\mathcal{U}_0}^{(K)} \geq -\mathbb{J}_{\mathcal{U}_0^{(K)}}^{(K)} \quad \text{and} \quad -\mathbb{J}_{\mathcal{T}_1^{(K)}}^{(K)} \geq -\mathbb{J}_{\mathcal{U}_0}^{(K)}.$$

Then, Theorem 1.2 follows in the case $M \geq 2$, and also trivially when $M = 1$, from Propositions 2.1 and 2.2. \square

3 $P^{(K)}$ Landscape: Proof of Proposition 1.1

Proof of Proposition 1.1 (I). Suppose $K \geq 2$ as the claim is just the definition when $K = 1$. Now, by definition, \vec{x} is *not* degenerate transient exactly when \vec{x} leads to itself with respect to the time-homogeneous process run with $P^{(K)}$, or in other words exactly when there is a sequence $\vec{x} = \vec{y}^0, \vec{y}^1, \dots, \vec{y}^l, \vec{y}^{l+1} = \vec{x}$ such that $\prod_{j=0}^l p^{(K)}(\vec{y}^j, \vec{y}^{j+1}) > 0$. But $p^{(K)}(\vec{x}, \vec{y}^1) \cdots p^{(K)}(\vec{y}^l, \vec{x}) > 0$ exactly when $\vec{y}^1 = \langle x_2, \dots, x_K, y_K^1 \rangle$ and $p(x_K, y_K^1) > 0$, $\vec{y}^2 = \langle x_3, \dots, x_K, y_K^1, y_K^2 \rangle$ and $p(y_K^1, y_K^2) > 0$, all the way to $\vec{y}^l = \langle y_1^l, x_1, \dots, x_{K-1} \rangle$ and $p(x_{K-2}, x_{K-1}) > 0$, and also $p(x_{K-1}, x_K) > 0$. Hence, when \vec{x} is not degenerate transient, x_s leads to x_t for all $1 \leq s, t \leq K$ with respect to the P -time-homogeneous evolution. Conversely, if x_s does not lead to x_t for some $1 \leq s, t \leq K$ under P , we cannot construct a path with positive probability from \vec{x} to itself under $P^{(K)}$. \square

Before proving part (II), we need the following statement.

Lemma 3.1 When $K \geq 2$, we have for $i \in \mathcal{G}$ that $\bar{C}_i^{(K)}$ is an irreducible set with respect to $P^{(K)}$, and, if non-empty, $(C_i)^K \setminus \bar{C}_i^{(K)}$ consists of degenerate transient vectors.

Proof. First, let $\vec{x}^1, \vec{x}^2 \in \bar{C}_i^{(K)}$. By definition, $\prod_{l=1}^{K-1} p(x_l^u, x_{l+1}^u) > 0$ for $u = 1, 2$. Also, as $P(i)$ is irreducible, x_K^1 leads to x_1^2 under the time-homogeneous process run with P . Hence, \vec{x}^1 leads to \vec{x}^2 under the time-homogeneous process run with $P^{(K)}$,

and so $\bar{C}_i^{(K)}$ is irreducible. Second, by the definition of a vector $\vec{x} \in (C_i)^K \setminus \bar{C}_i^{(K)}$, we have \vec{x} cannot lead to itself through any path $\vec{y}^1, \dots, \vec{y}^l$ as $p^{(K)}(\vec{x}, \vec{y}^1) \cdots p^{(K)}(\vec{y}^l, \vec{x}) \leq \prod_{i=1}^{K-1} p(x_i, x_{i+1}) = 0$, and so is degenerate transient. \square

Proof of Proposition 1.1 (II). Suppose $K \geq 2$ as the 1 : 1 correspondence is trivial for $K = 1$. Let $i \in \mathcal{G}^{(K)}$, and first observe $C_i^{(K)} \subset (C_j)^K$ some $1 \leq j \leq N + M$. Indeed, $C_i^{(K)}$ does not contain any degenerate transient words. And, any $\vec{x} \in \Sigma^K$ for which $x_s \in C_u$ and $x_t \in C_v$ for $1 \leq u < v \leq N + M$ and some $1 \leq s, t \leq K$ is such that x_s does not lead to x_t under the P -time-homogeneous evolution by inspecting the canonical decomposition of P . Hence, by Proposition 1.1 (I), \vec{x} is $P^{(K)}$ -degenerate transient, and so cannot belong to $C_i^{(K)}$.

Next, sets $(C_j)^K$ for $j \in \mathcal{D}$ are singletons concentrated on words of form $\langle x, x \dots, x \rangle$, where x does not lead to itself under P , and so are $P^{(K)}$ -degenerate transient by Proposition 1.1 (I). Then, $C_i^{(K)} \subset (C_j)^K$ for some unique $j \in \mathcal{G}$ as the sets $\{C_j : j \in \mathcal{G}\}$ are disjoint. Moreover, as $C_i^{(K)}$ is also irreducible, by Lemma 3.1, $C_i^{(K)} = \bar{C}_j^{(K)}$. \square

4 Condition (SIE-1): Proof of Proposition 2.1

Suppose $K \geq 2$ as the claim follows by assumption when $K = 1$. We now show that $\pi^{(K)}(C_i^{(K)}) > 0$ for all $i \in \mathcal{G}^{(K)}$. Let $i \in \mathcal{G}^{(K)}$ and, by Proposition 1.1 (II), write $C_i^{(K)} = \bar{C}_j^{(K)}$ for $j = \phi(i) \in \mathcal{G}$. As $\pi(C_j) > 0$ by assumption, let $x_0 \in C_j$ be such that $\pi(x_0) > 0$. Also, as C_j is irreducible, let $x_1, \dots, x_{K-1} \in C_j$ where $p(x_0, x_1) \cdots p(x_{K-2}, x_{K-1}) > 0$. Then, $\pi^{(K)}(C_i^{(K)}) = \mathbb{P}_\pi(\langle X_0, \dots, X_{K-1} \rangle \in \bar{C}_j^{(K)}) \geq \pi(x_0) \prod_{l=0}^{K-2} p(x_l, x_{l+1}) > 0$.

For the second part of Condition (SIE-1), let $i \in \mathcal{G}^{(K)}$, and $\vec{x}, \vec{y} \in C_i^{(K)}$. Now, if $p^{(K)}(\vec{x}, \vec{y}) > 0$, then $x_{k+1} = y_k$ for all $1 \leq k \leq K - 1$, and $p(x_K, y_K) > 0$. Also, as $x_K, y_K \in C_j$ where $j = \phi(i) \in \mathcal{G}$, we have $p_n(x_K, y_K) > 0$ for $n \geq 1$. Hence, $p_n^{(K)}(\vec{x}, \vec{y}) = p_{n+K-1}(x_K, y_K) > 0$ for $n \geq 1$. \square

5 Routing Costs: Proof of Proposition 2.3

Proof of Proposition 2.3 (I). Let $\mathcal{U}_0^{(K)}(i, j)$ be attained on $0 \leq k \leq R_0^{(K)} - 2$ and distinct indices $L_k = \langle i = l_0, l_1, \dots, l_k, l_{k+1} = j \rangle$. We now cull indices related to $\tilde{\mathcal{U}}_0^{(K)}(i, j) = \mathcal{U}_0(\phi(i), \phi(j))$.

1. As $i, j \in \mathcal{G}^{(K)}$ are distinct, we have $\phi(i) \neq \phi(j)$, and so there exists a largest integer $0 < s \leq k + 1$ such that $\phi(l_s) \neq \phi(l_0)$ but $\phi(l_{s-1}) = \phi(l_0)$. Let $s(0) = s$, and denote $l'_0 = \phi(l_0)$ and $l'_1 = \phi(l_{s(0)})$. If $\phi(l_{s(0)}) = \phi(j)$, we stop the process.

2. For $i \geq 1$, let $s(i)$ be the largest integer $s(i-1) < s \leq l_{k+1}$ such that $\phi(l_s) \neq \phi(l_{s(i-1)})$ but $\phi(l_{s-1}) = \phi(l_{s(i-1)})$. Let $l'_{i+1} = \phi(l_{s(i)})$. If $\phi(l_{s(i)}) = \phi(j)$, we stop, and otherwise we repeat step 2.

This algorithm stops at index $l'_{k'+1} = \phi(l_{k+1}) = \phi(j)$ with $k' \leq k$, and selects distinct indices $\langle \phi(i) = l'_0, \dots, l'_{k'+1} = \phi(j) \rangle$ in the range of ϕ so that also $0 \leq k' \leq N + M - 2$.

Now, for $0 \leq i \leq k'$, let $t^{(K)}(n, (l_{s(i)-1}, l_{s(i)}))$ be evaluated on some vectors $\vec{x} \in C_{l_{s(i)-1}}^{(K)}$ and $\vec{y} \in C_{l_{s(i)}}^{(K)}$, and so $x_K \in C_{l'_i}$ and $y_K \in C_{l'_{i+1}}$, where

$$t^{(K)}(n, l_{s(i)-1}, l_{s(i)}) = p_n^{(K)}(\vec{x}, \vec{y}) \leq p_{n+K-1}(x_K, y_K) \leq t(n + K - 1, (l'_i, l'_{i+1})).$$

Thus, $\overline{\lim}(1/n) \log t^{(K)}(n, (l_{s(i)-1}, l_{s(i)})) \leq \overline{\lim}(1/n) \log t(n, (l'_i, l'_{i+1}))$ and hence

$$\begin{aligned} \mathcal{U}_0^{(K)}(i, j) &= \sum_{m=0}^k \overline{\lim} \frac{1}{n} \log t^{(K)}(n, (l_m, l_{m+1})) \\ &\leq \sum_{m=0}^{k'} \overline{\lim} \frac{1}{n} \log t^{(K)}(n, (l_{s(m)-1}, l_{s(m)})) \quad \text{using nonpositivity of the limits} \\ &\leq \sum_{m=0}^{k'} \overline{\lim} \frac{1}{n} \log t(n, (l'_m, l'_{m+1})) \leq \mathcal{U}_0(\phi(i), \phi(j)) \end{aligned}$$

after noting definition (3). □

Proof of Proposition 2.3 (II). When $K = 1$, the bound follows as $\mathcal{T}_1 \geq \mathcal{T}_2$ (cf. step 3, section 2) and ϕ is the identity. Also, as $M \geq 2$ by assumption, it will be useful to note necessarily the number of states $\mathfrak{r} \geq 2$. Suppose now $K \geq 2$, and let $p_{\min} = \min\{p(s, t) : p(s, t) > 0, s, t \in C_l, l \in \mathcal{G}\}$ be the minimum positive transition probability within nondegenerate P -sets. Let $\vec{y} \in C_i^{(K)} = \bar{C}_{\phi(i)}^{(K)}$ and $\vec{z} \in C_j^{(K)} = \bar{C}_{\phi(j)}^{(K)}$. As $C_{\phi(i)}, C_{\phi(j)}$ are irreducible, we can find $y \in C_{\phi(i)}$ and $z \in C_{\phi(j)}$ where $p(y_K, y) > 0$ and $p(z, z_1) > 0$. It will be enough to show for all large n that

$$\underline{\gamma}^{(K),1}(n, \vec{y}, \vec{z}) \geq (p_{\min}/2)^{K+1} \underline{\gamma}^2(n + K, y, z)$$

as then $\underline{\gamma}^{(K),1}(n, \vec{y}, \vec{z}) \geq (p_{\min}/2)^{K+1} \underline{\gamma}^2(n + K, (\phi(i), \phi(j)))$, and the result would follow by taking an appropriate infimum and limit.

Applying the definition in section 2, let $0 \leq k \leq N + M - 2$, $L_k = \langle \phi(i) = l_0, l_1, \dots, l_k, l_{k+1} = \phi(j) \rangle$ be distinct indices in $\{1, \dots, N + M\}$, $Q_k = \langle q_0, \dots, q_{k+1} \rangle$ satisfy (5) with $K = 1$, and $V_k = \langle \vec{x}^0, \vec{x}^1, \dots, \vec{x}^{k+1} \rangle$ with vectors \vec{x}^i of distinct elements $x_1^i, \dots, x_{q_i}^i \in C_{l_i}$ such that

$$\underline{\gamma}^2(n + K, y, z) = \mathbb{P}_{(n+K-1, y)}(\bar{X}_{n+K}^{n+K+r(k+1)} = \langle \vec{x}^0, \vec{x}^1, \dots, \vec{x}^{k+1}, z \rangle)$$

where $\bar{X}_l^m = \bar{X}_{l,m} = \langle X_l, \dots, X_m \rangle$ for $l \leq m$ makes the notation compact.

We now form a path from \vec{y} to \vec{z} , with respect to the K -word chain, denoted by V' :

$$\begin{aligned} \vec{x}^{(K),0} &= \langle y_2, \dots, y_K, y \rangle, \quad \vec{x}^{(K),1} = \langle y_3, \dots, y_K, y, x_1^0 \rangle, \dots \\ &\dots, \quad \vec{x}^{(K),k''-1} = \langle x_{q_{k+1}}^{k+1}, z, z_1, \dots, z_{K-2} \rangle, \quad \vec{x}^{(K),k''} = \langle z, z_1, \dots, z_{K-1} \rangle \end{aligned}$$

where the path has $k'' + 1 = K + 1 + r(k + 1)$ words. Note, by construction, we ensure $\vec{x}^{(K),0} \in \bar{C}_{\phi(i)}^{(K)} = C_i^{(K)}$, as $p(y_K, y) \prod_{l=2}^{K-1} p(y_l, y_{l+1}) > 0$, and similarly $\vec{x}^{(K),k''} \in \bar{C}_{\phi(j)}^{(K)} = C_j^{(K)}$.

Let $L' = \langle i = l_0^{(K)}, l_1^{(K)}, \dots, l_{k'}^{(K)}, l_{k'+1}^{(K)} = j \rangle$ be the say $k' + 2$ indices of sets $\{C_j^{(K)}\}$ visited by the K -word chain. That is, the path V' begins in $C_i^{(K)}$, then eventually enters a different set $C_{l_1^{(K)}}^{(K)}$ from which it enters $C_{l_2^{(K)}}^{(K)}$ where $l_2^{(K)} \neq l_1^{(K)}$, and so on until it enters $C_j^{(K)}$ from $C_{l_{k'}^{(K)}}^{(K)}$ with $l_{k'}^{(K)} \neq j$. Let $Q' = \langle q_0^{(K)}, \dots, q_{k'+1}^{(K)} \rangle$ list the associated numbers of steps in each set along the path V' .

1. L' consists of distinct indices in $\{1, \dots, R_0^{(K)}\}$, and so $k' \leq R_0^{(K)} - 2$. Indeed, as L_k is composed of distinct indices, and vectors \vec{x}^i composed of distinct states, the elements in concatenated vector $\langle \vec{x}^0, \dots, \vec{x}^{k+1} \rangle$ are all distinct, and so words $\vec{x}^{(K),1}, \dots, \vec{x}^{(K),k''-1}$ are distinct. Hence, degenerate transient singleton sets are not repeated in the path V' . Also, as L_k has distinct indices, nondegenerate sets are not repeated as well.

2. Q' satisfies (5). Indeed, by following the ‘‘last letter evolution,’’ the last letter of $\vec{x}^{(K),q_0+1}$ is $x_1^1 \notin C_{l_0} = C_{\phi(i)}$, and so $\vec{x}^{(K),q_0+1} \notin C_i^{(K)}$; hence as $q_0 \leq \mathfrak{r}$, we have $1 \leq q_0^{(K)} \leq \mathfrak{r} + 1$, and as $\mathfrak{r}, K \geq 2$, $\mathfrak{r} + 1 \leq \mathfrak{r}^K$. Similarly, $1 \leq q_{k'+1}^{(K)} \leq 1 + \mathfrak{r} \leq \mathfrak{r}^K$, and as $q_l \leq \mathfrak{r} + 1$ for $1 \leq l \leq k$, $1 \leq q_s^{(K)} \leq \mathfrak{r} + 1 \leq \mathfrak{r}^K + 1$ for $1 \leq s \leq k'$. Then,

$$\begin{aligned} \underline{\gamma}^{(K),1}(n, \vec{y}, \vec{z}) &\geq \mathbb{P}_{(n-1, \vec{y})}^{(K)}(\vec{X}_{n, n+k'+1}^{(K)} = \langle \vec{x}^{(K),0}, \vec{x}^{(K),1}, \dots, \vec{x}^{(K),k''} \rangle, \vec{z}) \\ &= p_{n+K-1}(y_K, y) \mathbb{P}_{(n+K-1, y)}(\vec{X}_{n+K}^{n+K+r(k+1)} = \langle \vec{x}^0, \vec{x}^1, \dots, \vec{x}^{k+1} \rangle, z) \\ &\quad \cdot \mathbb{P}_{(n+K+r(k+1), z)}(\vec{X}_{n+K+1+r(k+1)}^{n+2K+r(k+1)} = \vec{z}) \\ &\geq (p_{\min}/2)^{K+1} \underline{\gamma}^2(n + K, y, z) \end{aligned}$$

for all n large since $p_{n+K-1}(y_K, y) \rightarrow p(y_K, y) \geq p_{\min}$ and

$$\mathbb{P}_{(n+K+r(k+1), z)}(\vec{X}_{n+K+1+r(k+1)}^{n+2K+r(k+1)} = \vec{z}) \rightarrow p(z, z_1) \prod_{l=1}^{K-1} p(z_l, z_{l+1}) \geq (p_{\min})^K. \quad \square$$

References

- [1] Dembo, A., Zeitouni, O. (1998) *Large Deviations Techniques and Applications*. Second Edition. Applications of Mathematics **38**, Springer-Verlag, New York.
- [2] Dietz, Z., and Sethuraman, S. (2005) Large deviations for a class of nonhomogeneous Markov chains. *Ann. Appl. Probab.* **15** 421-486.
- [3] Seneta, E. (1981) *Non-negative Matrices and Markov Chains*. Springer-Verlag, New York.