Spectral State Compression of Markov Processes

Anru Zhang⬛ and Mengdi Wang

Abstract— **Model reduction of Markov processes is a basic problem in modeling state-transition systems. Motivated by the state aggregation approach rooted in control theory, we study the statistical state compression of a discrete-state Markov chain from empirical trajectories. Through the lens of spectral decomposition, we study the rank and features of Markov processes, as well as properties like representability, aggregability, and lumpability. We develop spectral methods for estimating the transition matrix of a low-rank Markov model, estimating the leading subspace spanned by Markov features, and recovering latent structures like state aggregation and lumpable partition of the state space. We prove statistical upper bounds for the estimation errors and nearly matching minimax lower bounds. Numerical studies are performed on synthetic data and a dataset of New York City taxi trips.**

Index Terms— **Computational complexity, maximum likelihood estimation, minimax techniques, signal denoising, tensor SVD.**

I. INTRODUCTION

MODEL reduction is a central problem in scientific studies, system engineering, and data science. In many situations one needs to learn about a complex system from trajectories of noisy observations. When data is limited, the unknown system becomes difficult to model, analyze, infer and let alone optimize.

In this paper, we study the dimension reduction of a Markov chain $\{X_0, X_1, \ldots, X_n\}$ where the state space is discrete and finite but very large. There are two goals: The first goal is data compression and recovery of a reduced-order Markov model. The second goal is to extract features for state representation, which can be further used to find state aggregation or lumpable clusters. These two goals are closely tied to each other - achieving either one would trivialize the other one. We refer to the combination of these two goals as the problem of *state compression*.

State compression of discrete Markov chains finds wide applications. For an example of network analysis, records of taxi trips can be viewed as a fragmented sample path realized from a city-wide Markov chain [1], [2], and experiments suggest that one can estimate latent traffic network from sample paths [3]. Similar needs for analyzing Markov transition data

also arise from ranking problems in e-commerce [4], [5], where clickstreams can be viewed as a random walk on the space of all possible clicks.

Our work is inspired by the state aggregation approach that is commonly used to reduce the complexity of reinforcement learning and control systems. State aggregation means to aggregate "similar" states into a small number of "meta states," which are typically handpicked based on domain-specific knowledge [6], [7] or based on given similarity metrics or feature functions [8]. In the context of discrete-state Markov chains, the goal of state aggregation is to find a partition mapping E such that $\mathbb{P}(X_{t+1} \mid X_t) \approx \mathbb{P}(X_{t+1} \mid E(X_t))$. In fact, the state aggregation structure corresponds to a particular low-rank decomposition of the system's transition kernel (see Proposition 3). Another inspiring example is the use of membership models for modeling large Markov decision processes, where each observed state is mapped into a mixture over meta states [7], [9]. This membership model, also known as soft state aggregation, corresponds to a low-rank decomposition structure of the transition kernel (see Proposition 2). These existing approaches for dimension reduction of control and reinforcement learning mainly rely on *priorly known* meta-states or membership models. In contrast, we aim to learn the state aggregation structure from trajectorial data in an unsupervised manner.

Let us investigate the spectral decomposition of the Markov chain, of the form

$$\mathbb{P}(X_{t+1} \mid X_t) \approx \sum_{k=1}^{r} f_k(X_t) g_k(X_{t+1}),$$

where $f_1, \ldots, f_r, g_1, \ldots, g_r$ are some feature functions and r is the rank. The spectral decomposition of the transition kernel provides a natural venue towards state compression, where $f_1, \ldots, f_r, g_1, \ldots, g_r$ can be used as basis functions to represent the state space using a small set of parameters. There are many open fundamental questions: How to estimate the feature functions and the leading feature space? How to estimate the Markov model under a low-rank assumption? What are the statistical limits for state compression? In this paper, we plan to take a substantial step towards answering these questions.

We propose a class of *spectral state compression* methods for finite-state nonreversible Markov process with provably sharp statistical guarantees. Our main results are summarized as follows.

1) *Spectral properties of Markov chains, aggregability, and lumpability.* We study the spectral decomposition of Markov chains, and we show it is closely related to aggregability and lumpability of the process.

Aggregability means that the states can be aggregated into blocks while preserving the transition probability distributions, while lumpability means that states can be clustered while preserving the strong Markov property.

2) *Sharp statistical guarantees for estimating low-rank Markov models.* For Markov chains with a *known* small rank, we provide a spectral method for estimating the transition matrices and establish upper bounds on the finite-sample total variation error. We also establish a nearly matching minimax lower bound. These results also extend to the estimation of general low-rank stochastic matrices that are not necessarily square.

3) *Sharp statistical guarantees for state space compression of general Markov chains.* For general Markov chains that is not low-rank, we show that the spectral method recovers the leading Markov feature space with high accuracy. Upper bounds and minimax lower bounds for the subspace recovery errors are established. In special cases of aggregable or lumpable processes, we show that one can further recover the state aggregation or lumpable partition with statistical guarantees.

In numerical experiments, we apply state compression to analyze the New York City Yellow Cab data. By modeling taxi trips as sample transitions realized from a citywide random walk, our spectral state aggregation method indeed reveals latent traffic patterns and meaningful partition of NYC.

A. Outline

Section II surveys related literature. Section III studies the spectral decomposition of the Markov chains and properties such as the representability, aggregability and lumpability. Section IV proposes a spectral method for estimating low-rank Markov models and provides theoretical guarantees. Section V proposes state compression methods for estimating the leading feature space and recovery of the state aggregation structure or lumpable partition. Section VI gives numerical experiments. Proofs are given in the supplement.

B. Notations

We use lowercase letters such as x, y, z to denote scalars and vectors, and use boldface uppercase letters like $\mathbf{X}, \mathbf{F}, \mathbf{P}$ to represent matrices. For $x, y \in \mathbb{R}$, we denote $x \wedge y = \max\{x, y\}$, $x \vee y = \min\{x, y\}$ and $(x)_+ = \max\{x, 0\}$. For a vector $v \in \mathbb{R}^p$, we denote $\|u\|_q = \left(\sum_{i=1}^p |v_i|^q\right)^{1/q}$ for all $q > 0$ and $\|u\|_\infty = \max_{1 \le i \le p} |u_i|$. For a matrix $\mathbf{X} \in \mathbb{R}^{p_1 \times p_2}$, we denote by $\sigma_k(\mathbf{X})$ its k-th largest singular value, and denote $\|\mathbf{X}\|_F = \left(\sum_{i,j} \mathbf{X}_{ij}^2\right)^{1/2}$, $\|\mathbf{X}\| = \|\mathbf{X}\|_2 = \sup_{\|u\|_2 \le 1} \|\mathbf{X}u\|_2$, and $\|\mathbf{X}\|_1 = \sum_{i,j} |\mathbf{X}_{ij}|$. For two sequences $\{a_n\}, \{b_n\}$, we say $a_n \asymp b_n$ if there exists $c_1 > c_2 > 0$ such that $c_2 b_n \le a_n \le c_1 b_n$ for all n sufficiently large.

II. RELATED LITERATURE

This work relates to a broad range of model reduction methods from dynamical systems, control theory, and reinforcement learning. For instance in studies of fluid dynamics and molecular dynamics, various spectral methods were developed for approximating the transfer operators, their eigenvalues, eigenfunctions and eigenmodes, including time-lagged independent component analysis (e.g., [10], [11]) and dynamic mode decomposition (e.g. [12], [13]). See [14] for a review of data-driven dimension reduction methods for dynamical systems. In control theory and reinforcement learning, state aggregation is a long known approach for reducing the complexity of the state space and thus reducing computational costs for approximating the optimal value function or policy; see e.g., [7]–[9], [15], [16]. Beyond the state aggregation approach, a related direction of research, known as representation learning, is to construct basis functions for representing high-dimensional value functions. Methods have been developed based on diagonalization or dilation of some Laplacian operator that is used as a surrogate of the exact transition operator; see for examples [17]–[20]. Reference [21] gave a comprehensive review of representation learning for Markov decision problems and an extension to continuous-state control problems. The aforementioned methods typically require prior knowledge about structures of the problem or transition function of the system, lacking statistical guarantees.

Our methods and analyses developed in this paper use ideas and proof techniques that can be traced back to discrete distribution estimation, matrix completion, principal component analysis and spectral clustering. In what follows, we review related results in these areas.

Our first main results are the minimax upper and lower bounds for estimating low-rank Markov models (Section IV). These results are related to the problem of discrete distribution estimation, which is a basic problem that has been considered in both the classic and recent literature [22]–[26]. These works established minimax-optimal estimation results or various losses (e.g., total variation distance and Kullback-Leibler divergence) and specific discrete distributions when the observations are generated independently from the target distribution.

Another related topic is matrix completion, where the goal is to recover a low-rank matrix from a limited number of randomly observable entries. Various methods, such as nuclear norm minimization [27], [28], projected gradient descent [29], [30], singular value thresholding [31], [32], max norm minimization [33], [34], etc, were introduced and extensively studied in the past decade. Similar to [31], [32], our proposed estimators involve a singular value thresholding step. In contrast to matrix completion, the input data considered in this paper are transitions from a sample path of a random walk - they never reveal any exact entry of the unseen transition matrix and the data are highly dependent. In addition, the transition matrix to be estimated is known to be a stochastic matrix, making the problem distinct from matrix completion.

Recovery of a low-rank probability transition matrix has been considered by [35]–[37].[1] Reference [35] studied a spectral method for estimating hidden Markov models and proved sample complexity for the Kullback-Leibler divergence that depends on spectral properties of the model. A subroutine

---

[1] [37] was completed after the initial arxiv version of the current paper was released, therefore [37] is not a prior work.

of the method conducts spectral decomposition of a multi-step empirical transition matrix for identifying the hidden states. Reference [36] recently studied the estimation of a rank-two probabilistic matrix from observations of independent samples and provided error upper bounds. Reference [37] studied a rank-constrained likelihood estimator for Markov chains and provide upper and lower bounds for the Kullback-Leibler divergence. In comparison to these works, we focus on the Markov processes, and we provide explicit upper bounds and minimax lower bounds for the total variation distance and the subspace angle.

Our results for spectral state aggregation and spectral lumpable partition can be viewed as variants of spectral clustering. Spectral clustering is a powerful tool in unsupervised machine learning for analyzing high-dimensional data [38], [39]. It is widely used in community detection [40]–[42], high-dimensional feature clustering [43], [44], imaging segmentation [45], [46], matrix completion [31], [47]. In most of these works, the input data are independent and clusters are computed based on some similarity metric or symmetric covariance matrices. In comparison, the proposed methods of spectral state aggregation and spectral lumpable partition are not based on any similarity metric or symmetric matrix. The two methods are developed to exploit linear algebraic structures that are particular to aggregatability and lumpability, respectively. In particular, the spectral state aggregation method aims to cluster states while maximally preserving the outgoing distributions, while the spectral lumpable partition method focuses on preserving the strong Markov property of the random walk. More specifically, state aggregation is based on the left Markov features, while lumpable partition relates to both the left and right features. A related work by [48] studied the lumpable network partition problem by analyzing the eigen-structures when the network is exactly given. Our spectral method for estimating the lumpable partition is based on singular value decomposition rather than eigendecomposition. Following this work, the paper [49] later studied nonnegative factorization for estimating the soft state aggregation model and the paper [50] developed a kernelized state compression method for representation learning of multivariate time series data.

III. MARKOV RANK, AGGREGABILITY, AND LUMPABILITY

Let $\{X_0, \ldots, X_n\}$ be a Markov chain on the space Ω. When Ω is a finite set $\Omega = \{1, \ldots, p\}$, let the transition matrix be $\mathbf{P} \in \mathbb{R}^{p \times p}$ where $\mathbf{P}_{ij} = \mathbb{P}(X_k = j | X_{k-1} = i, X_{k-2}, \ldots, X_0)$ for all $k \geq 1$, $1 \leq i, j \leq p$. Throughout this paper, we assume $\{X_0, \ldots, X_n\}$ is ergodic so there exists an invariant distribution $\pi \in \mathbb{R}^p$, i.e., $\pi_i = \lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^n 1_{\{X_k = i\}}$. Furthermore, π is an invariant distribution if and only if $\pi^\top \mathbf{P} = \pi^\top$, $\pi_i \geq 0$, and $\sum_{i=1}^p \pi_i = 1$. Let $\pi_{\min} = \min_{1 \leq i \leq p} \pi_i$, $\pi_{\max} = \max_{1 \leq i \leq p} \pi_i$. Let $\mathbf{F} \in \mathbb{R}^{p \times p}$ be the long-run frequency matrix $\mathbf{F}_{ij} = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^n 1_{\{X_k = i, X_{k+1} = j\}}$, so that $\mathbf{F} = \text{diag}(\pi)\mathbf{P}$. For any $\varepsilon > 0$, the ε-*mixing time* of the Markov chain is defined as

$$\tau(\varepsilon) = \min \left\{ k : \max_{1 \leq i \leq p} \frac{1}{2} \left\| (\mathbf{P}^k)_{[i,:]} - \pi^\top \right\|_1 \leq \varepsilon \right\}. \quad (1)$$

We call $\tau_* = \tau(1/4)$ the mixing time for short. Please refer to [51], [52] for comprehensive discussions on the theory of Markov chain and mixing times.

Let us consider Markov chains with a small rank. This notion was introduced for Markov processes with a general state space by [53] as an example of "dependence that is close to independence". For more examples and properties of the finite-rank Markov chain, please refer to [54] and [55].

Definition 1 (Markov Rank, Kernel and Features). *The rank of a Markov chain $X_0, \ldots, X_n$ is the smallest integer r such that its transition kernel can be written in the form of*

$$\mathbb{P}(X_{t+1} \mid X_t) = \sum_{k=1}^r f_k(X_t) g_k(X_{t+1}), \quad (2)$$

where $f_1, \ldots, f_r$ are real-valued functions and $g_1, \ldots, g_r$ are probability mass functions. The non-degenerate $r \times r$ matrix $\mathbf{C}$ such that $\mathbf{C}_{ij} = f_j^\top g_i$ is referred to as the Markov kernel. We refer to $f_1, \ldots, f_r$ as left Markov features and $g_1, \ldots, g_r$ as right Markov features. If the Markov process has p discrete states, $f_1, \ldots, f_r, g_1, \ldots, g_r$ are p-dimensional vectors and $\mathbf{C}_{ij} = \sum_{k=1}^p f_j(k) g_i(k)$.

A low-rank Markov chain admits infinitely many decompositions of the form (2), therefore the kernel $\mathbf{C}$ and feature functions $f_1, \ldots, f_r, g_1, \ldots, g_r$ are not uniquely identifiable. In this paper, we will mainly focus what are identifiable, i.e., the transition kernel $\mathbf{P}$ and the feature spaces spanned by $f_1, \ldots, f_r$ and $g_1, \ldots, g_r$ respectively.

Proposition 1 shows that Markov features are sufficient to represent the multi-step Markov transition and the stationary distribution.

Proposition 1 (Representability of Markov Features; [54]). *Suppose that the Markov chain $X_0, \ldots, X_n$ has a rank r taking the form of* (2), *then*

1) *If the state space is finite, the transition matrix $\mathbf{P}$ satisfies* rank$(\mathbf{P}) = r$.
2) $\mathbb{P}(X_{t+n} \mid X_t) = \sum_{i=1}^r \sum_{j=1}^r f_i(X_t)(\mathbf{C}^{n-1})_{ij} g_j(X_{t+n})$.
3) *There exists $\gamma \in \mathbb{R}^r$ such that $\pi(\cdot) = \sum_{k=1}^r \gamma_k g_k(\cdot)$ and $\gamma^\top \mathbf{C} = \gamma^\top$.*

In addition, Markov features can be used as basis functions in the context of control and reinforcement learning for representing *value functions*. For example consider the reward process $h(X_0), \ldots, h(X_n)$, where $h : \Omega \mapsto \mathbb{R}$ is a reward functio. In control and reinforcement learning, a central quantity for evaluating the current state of the system is the discounted cumulative *value function* $\upsilon : \Omega \mapsto \mathbb{R}$, given by $\upsilon(x) = \mathbb{E}\left[\sum_{n=0}^\infty \alpha^n h(X_n) \mid X_0 = x\right]$, where $\alpha \in (0, 1)$ is a discount factor. Now if the Markov chain admits a decomposition of the form (2), we have $\upsilon(\cdot) = r(\cdot) + \sum_{k=1}^r w_k f_k(\cdot)$ for some scalars $w_1, \ldots, w_r$. In other words, the value function can be represented as a linear combination of left Markov features.

Next we introduce a notion of Markov non-negative rank, which is slightly more restrictive than the Markov rank.

Definition 2 (Markov Non-negative Rank). *The non-negative rank of a Markov chain is the smallest r such that its transition*
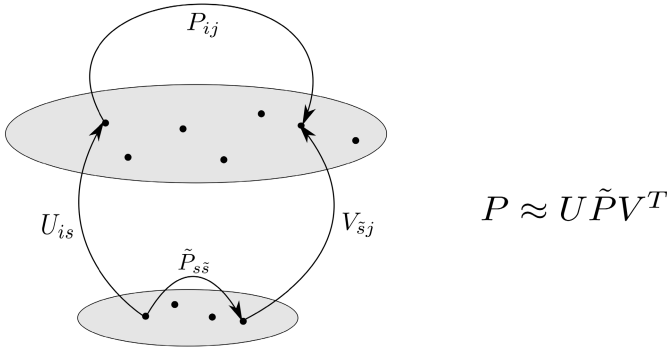
Fig. 1. Soft state aggregation of Markov chain with a small nonnegative rank. Raw states can be mapped to "meta-states" through a factorization model of the transition matrix.

kernel can be written in the form of

$$\mathbb{P}(X_{t+1} \mid X_t) = \sum_{k=1}^{r} f_k(X_t) g_k(X_{t+1})$$

for some nonnegative functions $f_1, \ldots, f_r, g_1, \ldots, g_r$.

This definition of nonnegative rank remains the same even if we restrict $g_1, \ldots, g_r$ are probability mass functions and for each $x \in \Omega$, $i \mapsto f_i(x)$ is a probability mass function. Denote by $\text{rank}_+(\mathbf{P})$ the nonnegative rank of $\mathbf{P}$. It is easy to verify that $\text{rank}_+(\mathbf{P}) = r$ if and only if there exist nonnegative matrices $\mathbf{U}, \mathbf{V} \in \mathbb{R}_+^{p \times r}$ and $\tilde{\mathbf{P}} \in \mathbb{R}_+^{r \times r}$ such that $\mathbf{P} = \mathbf{U}\tilde{\mathbf{P}}\mathbf{V}^\top$, where $\mathbf{U1} = \mathbf{1}, \mathbf{V}^\top \mathbf{1} = \mathbf{1}, \tilde{\mathbf{P}}\mathbf{1} = \mathbf{1}$. This decomposition means that one can map the states into meta-states while preserving most of the system dynamics (see Figure 1). In the context of control and dynamic programming, rows of $\mathbf{U}$ are referred to as *aggregation distributions* and columns of $\mathbf{V}$ are referred to as *disaggregation distributions* (see [56] Secion 6.3.7). It always holds that $\text{rank}(\mathbf{P}) \leq \text{rank}_+(\mathbf{P})$.

Low-rank decomposition of the Markov chain is related to several reduced-order models. For example, the Markov chain with a small nonnegative rank is equivalent to a membership model.

Proposition 2 (Nonnegative Markov Rank and Membership Model). *The Markov chain with transition probability matrix* $\mathbf{P}$ *has a nonnegative rank* $\text{rank}_+(\mathbf{P}) \leq r$ *if and only if there exists a stochastic process* $\{Z_t\} \subset \{1, \ldots, r\}$ *such that*

$$\mathbb{P}(Z_t \mid X_t) = \mathbb{P}(Z_t \mid X_1, \ldots, X_t),$$
$$\mathbb{P}(X_{t+1} \mid Z_t) = \mathbb{P}(X_{t+1} \mid X_1, \ldots, X_t, Z_t). \quad (3)$$

Next we consider an important special case of low-rank Markov processes that is amenable to state aggregation. State aggregation is a basic approach for describing complicated systems [7], [56] and is particularly useful for approximating value functions in optimization, control theory, and reinforcement learning [57]. The idea is to partition the state space into disjoint blocks and treat each block as a single new state.

Definition 3 (Aggregability of Markov Chains). *A Markov chain is* r-*state aggregatable if there exists a partition*

$\Omega_1, \ldots, \Omega_r$ *of* Ω *such that*

$$\mathbb{P}(X_{t+1} \mid X_t = i) = \mathbb{P}(X_{t+1} \mid X_t = j)$$

for any $i, j \in \Omega_k$, $k \in \{1, \ldots, r\}$.

It is easy to show that aggregability corresponds to a particular non-negative decomposition.

Proposition 3 (Decomposition of Aggregatable Markov Chains). *If a Markov chain is state-aggregatable with respect to a partition* $\Omega_1, \ldots, \Omega_r$, *its nonnegative rank is at most* r *and*

$$\mathbb{P}(X_{t+1} \mid X_t) = \sum_{k=1}^{r} 1_{\Omega_k}(X_t) g_k(X_{t+1})$$

for nonnegative functions $g_1, \ldots, g_r : \Omega \mapsto \mathbb{R}_+$, *where* 1_S *denotes the indicator function of a set* S.

Proposition 3 implies, the Markov chain with transition matrix $\mathbf{P}$ is r-state aggregatable if and only if $\text{rank}_+(\mathbf{P}) = r$ and there exist $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{p \times r}$ such that $\mathbf{P} = \mathbf{UV}^\top$, where $\mathbf{V}$ is nonnegative and $\mathbf{U} = [\mathbf{1}_{\Omega_1}, \ldots, \mathbf{1}_{\Omega_r}]$ indicates the membership.

A Markov process is called lumpable if the state space can be partitioned into blocks while still preserving the strong Markov property [58], [59].

Definition 4 (Lumpability of Markov Chains [58]). *A Markov process* $X_1, \ldots, X_n$ *is lumpable with respect to a partition* $\Omega_1, \ldots, \Omega_r$, *if for any* $k, \ell \in \{1, \ldots, r\}$, $x, x' \in \Omega_k$,

$$\mathbb{P}(X_{t+1} \in \Omega_\ell \mid X_t = x) = \mathbb{P}(X_{t+1} \in \Omega_\ell \mid X_t = x'). \quad (4)$$

If the Markov chain is lumpable, it has eigenvectors that are block structured and equal to indicators functions of the subsets [48]. However, a lumpable Markov chain is not necessarily low-rank. There may exist other eigenvectors corresponding to local dynamics within a subset. See Figure 2 for an example of Markov chain that is lumpable but not exactly low-rank. We show that the lumpable Markov chain has the following decomposition.

Proposition 4 (Decomposition of Lumpable Markov Chains). *Let the Markov chain with transition matrix* $\mathbf{P} \in \mathbb{R}^{p \times p}$ *be lumpable with respect to a partition* $\Omega_1 \ldots, \Omega_r$. *Then there exist* $\mathbf{P}_1, \mathbf{P}_2$ *such that* $\mathbf{P} = \mathbf{P}_1 + \mathbf{P}_2$ *and* $\mathbf{P}_1\mathbf{P}_2^\top = 0$, *where* $\mathbf{P}_1$ *can be written as*

$$\mathbf{P}_1 = \mathbf{Z} \cdot \bar{\mathbf{P}} \cdot \text{diag}(|\Omega_1|^{-1}, \ldots, |\Omega_r|^{-1}) \cdot \mathbf{Z}^\top,$$

where $\mathbf{Z} = [\mathbf{1}_{\Omega_1}, \ldots, \mathbf{1}_{\Omega_r}] \in \mathbb{R}^{p \times r}$, $\bar{\mathbf{P}} \in \mathbb{R}^{r \times r}$ *is the stochastic matrix such that* $\bar{\mathbf{P}}_{kl} = \mathbb{P}(X_{t+1} \in \Omega_l \mid X_t \in \Omega_k)$. *Let the SVD of* $\mathbf{P}_1$ *be* $\mathbf{P}_1 = \mathbf{U}_{P_1} \boldsymbol{\Sigma}_{P_1} \mathbf{V}_{P_1}^\top$. *Then* $(\mathbf{U}_{P_1})_{[i,:]} = (\mathbf{U}_{P_1})_{[i',:]}$ *and* $(\mathbf{V}_{P_1})_{[i,:]} = (\mathbf{V}_{P_1})_{[i',:]}$ *for any* $i, i' \in \Omega_k$ *and* $k \in \{1, \ldots, r\}$.

Lumpability is a more general concept and it contains aggregability as a special case. According to Prop. 3, aggregability is closely related to blockwise structures of the left Markov features, while according to Prop. 4, lumpability is related to structures of the Markov features of $\mathbf{P}_1$ instead of the full transition matrix.
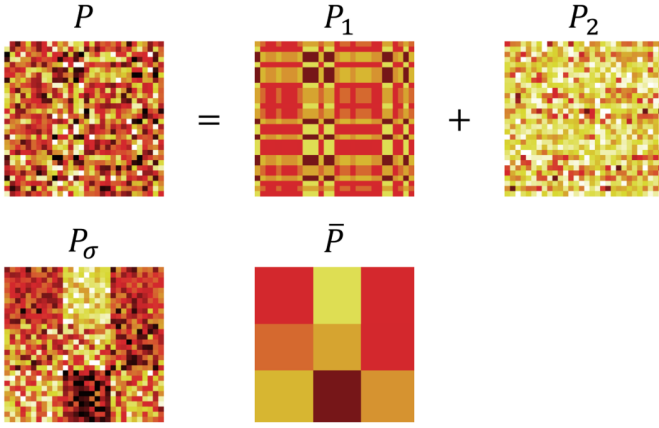
Fig. 2. Illustration of a lumpable Markov chain that is not exactly low-rank. The lumpable partition corresponds to a block-structured transition matrix after permutation. Here $\mathbf{P}_\sigma$ is the transition matrix after permutation and $\bar{\mathbf{P}}$ is the law of transition on the lumpable blocks.

Part of the results stated in Props. 1-4 are known in various works cited above. For completeness, we provide their proofs in Section VIII of the supplementary materials. In summary, the spectral decomposition of Markov processes plays a central role in many reduced-order models. Therefore the estimation of low-rank Markov models provides a natural venue towards state compression.

IV. MINIMAX ESTIMATION OF LOW-RANK MARKOV MODELS

In this section we focus on the p-state Markov chain $\{X_0, \ldots, X_n\}$ which has a priorly known rank r. Under the low-rank assumption, we aim to estimate the transition probability matrix $\mathbf{P}$ based on a sample path of n empirical state transitions. To this end, we propose a spectral estimation method and analyze the total variance distance between the estimator and the truth. A nearly matching minimax lower bound is also provided.

A. A Spectral Method for Markov Chain Estimation

Consider a Markov chain with transition matrix $\mathbf{P} \in \mathbb{R}^{p \times p}$ and frequency matrix $\mathbf{F} \in \mathbb{R}^{p \times p}$. Suppose that $\text{rank}(\mathbf{P}) = \text{rank}(\mathbf{F}) = r$ and we are given a $(n+1)$-long trajectory $\{X_0, \ldots, X_n\}$ starting at an arbitrary initial state. It is natural to estimate $\mathbf{P}$ and $\mathbf{F}$ via the *empirical frequency matrix* and *empirical transition matrix*, given by

$$\tilde{\mathbf{F}} = \left(\tilde{\mathbf{F}}_{ij}\right)_{1 \le i,j \le p}, \quad \tilde{\mathbf{F}}_{ij} = \frac{1}{n}\sum_{k=1}^{n} 1_{\{X_{k-1}=i, X_k=j\}}; \quad (5)$$

$$\tilde{\mathbf{P}} = \left(\tilde{\mathbf{P}}_{ij}\right)_{1 \le i,j \le p},$$

$$\tilde{\mathbf{P}}_{ij} = \begin{cases} \frac{\sum_{k=1}^{n} 1_{\{X_{k-1}=i, X_k=j\}}}{\sum_{k=1}^{n} 1_{\{X_{k-1}=i\}}}, & \text{if } \sum_{k=1}^{n} 1_{\{X_{k-1}=i\}} \ge 1; \\ \frac{1}{p}, & \text{if } \sum_{k=1}^{n} 1_{\{X_{k-1}=i\}} = 0. \end{cases}$$

$$(6)$$

Here, $1_{\{\cdot\}}$ is the indicator function and 1_p is the p-dimensional vector with all ones. Note that $\tilde{\mathbf{F}}, \tilde{\mathbf{P}}$ are in fact the maximum

---

Algorithm 1 Spectral Estimation of Low-Rank Markov Models

Input: $X_1, \ldots, X_n, r$

1) Construct $\tilde{\mathbf{F}}$ and $\tilde{\mathbf{P}}$ using (5)-(6).
2) Let the singular value decomposition (SVD) of $\tilde{\mathbf{F}}$ be $\tilde{\mathbf{F}} = \tilde{\mathbf{U}}_F \tilde{\boldsymbol{\Sigma}}_F \tilde{\mathbf{V}}_F^\top$, where $\tilde{\mathbf{U}}_F, \tilde{\mathbf{V}}_F$ are p-by-p orthogonal matrices and $\tilde{\boldsymbol{\Sigma}}_F$ is a p-by-p diagonal matrix.
3) Denoting $(x)_+ = \max\{x, 0\}$, let the frequency estimator $\hat{\mathbf{F}}$ be

$$\hat{\mathbf{F}} = (\hat{\mathbf{F}}_0)_+ / \|(\hat{\mathbf{F}}_0)_+\|_1,$$
$$\text{where} \quad \hat{\mathbf{F}}_0 = \tilde{\mathbf{U}}_{F,[:,1:r]} \tilde{\boldsymbol{\Sigma}}_{F,[1:r,1:r]} (\tilde{\mathbf{V}}_{F,[:,1:r]})^\top. \quad (7)$$

4) Let the transition matrix estimator $\hat{\mathbf{P}} \in \mathbb{R}^{p \times p}$ be

$$\hat{\mathbf{P}}_{[i,:]} = \begin{cases} \hat{\mathbf{F}}_{[i,:]} / \sum_{j=1}^{p} \hat{\mathbf{F}}_{ij}, & \text{if } \sum_{j=1}^{p} \hat{\mathbf{F}}_{ij} > 0, \\ \frac{1}{p} 1_p, & \text{if } \sum_{j=1}^{p} \hat{\mathbf{F}}_{ij} = 0, \end{cases} \quad (8)$$

$$i = 1, \ldots, p.$$

Output: $\hat{\mathbf{P}}, \hat{\mathbf{F}}$.

---

likelihood estimators and strongly consistent [60]. However, they do not account for the knowledge of a small Markov rank. Consider the special case of $r = 1$, where the Markov chain reduces to a sequence of i.i.d. random variables. Knowledge of $r = 1$ reduces the matrix estimation problem into estimation of a p-state discrete distribution. In contrast, the empirical estimators essentially look for a p^2-state distributions and will incur larger estimation errors.

We propose the following spectral method for estimating low-rank Markov chains.

Note that $\hat{\mathbf{F}}, \hat{\mathbf{P}}$ are not necessarily low-rank, due to the nonnegativity-preserving step $(\cdot)_+$ in (7). However, Algorithm 1 still enables data compression, because $\hat{\mathbf{F}}, \hat{\mathbf{P}}$ can be easily constructed based on the low-rank matrix $\hat{\mathbf{F}}_0$. As an alternative to $\hat{\mathbf{F}}$, we can also apply the algorithm by [61] to project $\hat{\mathbf{F}}_0$ onto the probability simplex to obtain an estimation of $\mathbf{F}$, and obtain

$$\hat{\mathbf{F}}^1 = \arg\min_{\hat{\mathbf{F}}^1} \|\hat{\mathbf{F}}^1 - \hat{\mathbf{F}}_0\|_F^2 \quad \text{subject to} \quad \hat{\mathbf{F}}_{ij}^1 \ge 0, \sum_{ij} \hat{\mathbf{F}}_{ij}^1 = 1.$$

The proposed estimators $\hat{\mathbf{F}}, \hat{\mathbf{P}}$ are related to the hard singular value thresholding estimators (HSVT), which were previously studied in matrix denoising [62], [63] and matrix completion [32]. Our method and its subsequent analysis differ from that of HSVT in two aspects. First, our estimation problem requires $\mathbf{P}, \mathbf{F}$ to be stochastic matrices that belong to particular simplexes (see Lemma 1 in the supplement). This is achieved by normalizing rows of the matrices and truncating negative values, which complicates the analysis of the estimation errors. Second, the analysis needs to account for the Markov dependency of the data, while the data are typically independent in matrix denoising and matrix completion.

Algorithm 1 requires r be selected in advance, which is needed by many other spectral-based methods. In practice, this value can be chosen empirically, for example one can (1) draw a scree plot for the SVD of $\tilde{\mathbf{F}}$, i.e., the cumulative ratio of total

variance as explained by the leading principal components, then select r as the location of the "elbow" in the scree plot; (2) evaluate the smallest r such that the first r principal components explain a certain level of total variation criterion, and (3) test by cross-validation. The readers are referred to [64] for a comprehensive discussion for rank selection.

B. Total Variation Upper Bound

Our first main result establishes the total variation distance upper bound between the proposed estimator and the truth.

Theorem 1 (Upper Bound). *Suppose* $\{X_0, \ldots, X_n\}$ *is generated by an ergodic Markov chain with transition probability matrix* $\mathbf{P} \in \mathbb{R}^{p \times p}$, *invariant distribution* $\pi \in \mathbb{R}^p$, *and mixing time* τ_*. *Let* $\hat{\mathbf{F}}$, $\hat{\mathbf{P}}$ *be the estimators given by* (7)-(8). *If* $\text{rank}(\mathbf{P}) = r$, *we have*

$$\mathbb{E}\|\hat{\mathbf{F}} - \mathbf{F}\|_1 \leq \sqrt{\frac{Crp}{n} \cdot \pi_{\max} p \cdot \tau_* \log^2(n)} \wedge 2, \quad (9)$$

$$\mathbb{E}\|\hat{\mathbf{F}}^1 - \mathbf{F}\|_1 \leq \sqrt{\frac{Crp}{n} \cdot \pi_{\max} p \cdot \tau_* \log^2(n)} \wedge 2,$$

$$\mathbb{E}\frac{1}{p}\sum_{i=1}^{p}\|\hat{\mathbf{P}}_{[i,:]} - \mathbf{P}_{[i,:]}\|_1 \leq \sqrt{\frac{Cr}{n} \cdot \frac{\pi_{\max}}{\pi_{\min}^2} \cdot \tau_* \log^2(n)} \wedge 2. \quad (10)$$

Let $\tilde{r} = \|\mathbf{F}\|_F^2/\sigma_r^2(\mathbf{F})$, $\kappa = p^2 \max_{ij} \mathbf{F}_{ij}$. *Then*

$$\mathbb{E}\max_{1 \leq i \leq p}\|\hat{\mathbf{P}}_{[i,:]} - \mathbf{P}_{[i,:]}\|_1 \leq \sqrt{\frac{C\tilde{r}}{n} \cdot \frac{\kappa^3}{p\pi_{\min}^2} \cdot \tau_* \log^2(n)} \wedge 2, \quad (11)$$

where $\sigma_r(\mathbf{F})$ *is the* r-*th singular value of* $\mathbf{F}$, C *is a universal constant.*

The proof of (9) and (1) relies on novel matrix Markov chain concentration inequalities with mixing time (Lemma 7), which characterizes the 2-norm distance between $\tilde{\mathbf{F}}$ and $\mathbf{F}$. Then based on the low-rank assumption of $\mathbf{F}$ and $\mathbf{P}$, a careful spectral analysis (Lemma 3) is performed to obtain the average error bound for $\hat{\mathbf{F}}$ and $\hat{\mathbf{P}}$. The proof of (11) is more involved. By using similar arguments, we can prove that $\hat{\mathbf{F}}^1$ also achieves the ℓ_2 risk upper bound in (9). Particularly, we derived concentration inequalities for projected Markov chains (Lemma 8), performed a more careful algebraic analysis, and obtained the uniform upper bound of total deviation for $\hat{\mathbf{P}}$. In what follows we make a few technical remarks.

Remark 1 (Spectral estimators $\hat{\mathbf{P}}, \hat{\mathbf{F}}$ vs. Empirical estimators $\tilde{\mathbf{P}}, \tilde{\mathbf{F}}$). Theorem 1 shows that $\mathbb{E}\|\hat{\mathbf{F}} - \mathbf{F}\|_1 \asymp \mathbb{E}\frac{1}{p}\|\hat{\mathbf{P}} - \mathbf{P}\|_1 \asymp \sqrt{pr/n}$, assuming all other parameters are fixed. In comparison, we can show that $\mathbb{E}\|\tilde{\mathbf{F}} - \mathbf{F}\|_1 \asymp \mathbb{E}\frac{1}{p}\|\tilde{\mathbf{P}} - \mathbf{P}\|_1 \asymp \sqrt{p^2/n}$, based on minimax error bounds for discrete distribution estimation [25], [26]. Therefore the spectral estimators are much more efficient because they utilize the low-rank structure. Numerical comparisons between the spectral and empirical estimators are given in Section VI.

Remark 2 (Dependence on the stationary distribution). The error bounds of Theorem 1 rely on $\pi_{\max}$ and $\pi_{\min}$, and they

take smaller values if π does not deviate much from the uniform distribution. When $\pi_{\min}$ is small, one has to pay a higher price for those states appearing the least frequently in the sample path. The dependence on π_{max} is due to a technical argument used in the proof to establish spectral norm concentration inequalities for asymmetric matrices (Lemma 7), which may be improvable under additional assumptions like reversibility.

Remark 3 (Dependence on the mixing time). The error bounds of $\hat{\mathbf{F}}, \hat{\mathbf{P}}$ involve a key quantity of Markov mixing time τ_*, whose actual value could be difficult to evaluate in practice [65]. We further show that similar error bounds like those in Theorem 1 hold if the mixing time is replaced with some eigengap. Please see Section IX-B Corollary for a generalization of Theorem 1 using an eigengap condition. Further improvement of the error bounds will require novel Markov concentration inequalities that have been developed in recent literature (see [66], [67]).

Remark 4 (About the row-wise uniform bounds). We introduce the entry-wise upper bound condition of κ for establishing the row-wise uniform upper bound (11). The dependence on κ suggests that estimating "overly-spiky" matrices is typically more difficult. Similar conditions were also used in the literature of low-rank matrix estimation (e.g. [27], [28]).

C. Minimax Lower Bound for Estimating Low-Rank Markov Chains

Now we investigate the information-theoretic limits of recovering low-rank Markov models. Consider the following class of low-rank transition matrices

$$\mathcal{P}_{p,r} = \left\{\mathbf{P} \in \mathcal{P}_p, \text{rank}(\mathbf{P}) \leq r\right\}, \quad (12)$$

where $\mathcal{P}_p$ is the class of all p-by-p transition matrices (see its definition in Eq. (54) in the supplementary material). Furthermore, we consider a more restricted class of low-rank Markov models with bounded mixing time and uniform ergodic distributions, given by

$$\mathcal{P}_{p,r}^* = \left\{\mathbf{P} \in \mathcal{P}_{p,r} : \tau_* = 1, \pi_{\max} = \pi_{\min} = 1/p\right\}. \quad (13)$$

We provide error lower bounds for recovering transition matrices within the aforementioned classes from finite trajectories.

Theorem 2 (Lower Bound). *Suppose we observe* $(n + 1)$ *consecutive transition states* $\{X_0, \ldots, X_n\}$, *where the starting point* X_0 *is randomly generated from the invariant distribution. Then*

$$\inf_{\hat{\mathbf{P}}} \sup_{\mathbf{P} \in \mathcal{P}_{p,r}} \mathbb{E}\frac{1}{p}\sum_{i=1}^{p}\left\|\hat{\mathbf{P}}_{[i,:]} - \mathbf{P}_{[i,:]}\right\|_1$$

$$\geq \inf_{\hat{\mathbf{P}}} \sup_{\mathbf{P} \in \mathcal{P}_{p,r}^*} \mathbb{E}\frac{1}{p}\sum_{i=1}^{p}\left\|\hat{\mathbf{P}}_{[i,:]} - \mathbf{P}_{[i,:]}\right\|_1 \geq c\left(\sqrt{\frac{rp}{n}} \wedge 1\right),$$

$$\inf_{\hat{\mathbf{F}}} \sup_{\substack{\mathbf{F}=\text{diag}(\pi)\mathbf{P}; \\ \mathbf{P} \in \mathcal{P}_{p,r}}} \mathbb{E}\sum_{i=1}^{p}\left\|\hat{\mathbf{F}}_{[i,:]} - \mathbf{F}_{[i,:]}\right\|_1$$

$$\geq \inf_{\hat{\mathbf{F}}} \sup_{\substack{\mathbf{F}=\mathrm{diag}(\pi)\mathbf{P}; \\ \mathbf{P}\in\mathcal{P}^*_{p,r}}} \mathbb{E} \sum_{i=1}^{p} \left\| \hat{\mathbf{F}}_{[i,:]} - \mathbf{F}_{[i,:]} \right\|_1 \geq c \left(\sqrt{\frac{rp}{n}} \wedge 1 \right),$$

where $c > 0$ is a universal constant, $\inf_{\hat{\mathbf{F}}}$ and $\inf_{\hat{\mathbf{P}}}$ are taken infimum over arbitrary estimators $\hat{\mathbf{P}}$ and $\hat{\mathbf{F}}$, respectively.

The proof is by constructing a series of instances of low-rank Markov chains with uniform stationary distributions and constant mixing times. We show that these instances are not distinguishable based on $(n + 1)$ sample transitions, by using the generalized Fano's lemma. See Section IX-C for the full proof.

Let us compare Theorems 1 and 2. The error upper bounds achieved by the spectral estimators $\hat{\mathbf{F}}, \hat{\mathbf{P}}$ are nearly minimax-optimal in their dependence on r, p, n (up to polylogarithmic terms), as long as parameters of the ergodic distribution $\pi_{\max}/\pi_{\min}$ and the mixing time τ_* are bounded by constants. This suggests that our spectral estimators are statistically efficient for fast mixing Markov processes as long as the ergodic distribution is balanced. It is not yet known whether the dependence on $\pi_{\max}/\pi_{\min}$ and τ_* is optimal.

D. Extension to Rectangular Probability Matrix

The proposed spectral method can be extended to estimating a broader class of probability matrices - not limited to transition matrices of Markov chains. An example of such an estimation problem arises from policy imitation in reinforcement learning, where one observes a sequence of state-action pairs generated by an expert policy that is applied in a Markov decision process. In this case, the expert policy can be represented using a transition probability matrix where each entry assigns the probability of choosing an action at a given state. The policy matrix is typically low-rank, as long as the Markov decision process admits state aggregation structures or can be represented using membership models.

Specifically, suppose we are given a stochastic process $\{(X_0, Y_0), (X_1, Y_1), \ldots, (X_n, Y_n)\}$. We assume that $\{X_0, X_1, \ldots, X_n\}$ is an ergodic Markov process on p states with invariant distribution π and Markov mixing time τ_*. We are interested in estimating the transition matrix $\mathbf{Q} \in \mathbb{R}^{p\times q}$ such that

$$\mathbf{Q}_{ij} = \mathbb{P}(Y_k = j \mid X_k = i),$$

for all k, i, j. Analogous to Section IV-A, we propose a spectral estimator for $\mathbf{Q}$ assuming that it has a priorly known rank r.

Theorem 3. *Let* $\{(X_0, Y_0), (X_1, Y_1), \ldots, (X_n, Y_n)\}$ *be a stochastic process as described previously,* $r = rank(\mathbf{Q})$. *Let* π, τ_* *be the stationary distribution and mixing time of* $\{X_0, \ldots, X_n\}$ *respectively. Let* $\mathbf{G} = \mathrm{diag}(\pi)\mathbf{Q}$ *and* $\kappa = pq \max_{ij} \mathbf{G}_{ij}$. *Then*

$$\mathbb{E}\frac{1}{p} \sum_{i=1}^{p} \|\hat{\mathbf{Q}}_{[i,:]} - \mathbf{Q}_{[i,:]}\|_1$$

$$\leq C\sqrt{\frac{(p \vee q)r}{n} \cdot \frac{\kappa}{(p\pi_{\min})^2} \cdot \tau_* \log^2(n)} \wedge 2. \quad (15)$$

Algorithm 2 Spectral Estimation of Rectangular Probability Matrix

Input: $\{(X_0, Y_0), (X_1, Y_1), \ldots, (X_n, Y_n)\}$, r.

1) Let $\tilde{\mathbf{G}}$ be the empirical estimate of the frequency matrix $\mathbf{G} = \mathrm{diag}(\pi)\mathbf{Q}$ such that

$$\tilde{\mathbf{G}} \in \mathbb{R}^{p\times q}, \quad \tilde{\mathbf{G}}_{ij} = \frac{1}{n} \sum_{k=1}^{n} 1_{\{(X_k, Y_k)=(i,j)\}}.$$

2) Calculate the SVD $\tilde{\mathbf{G}} = \tilde{\mathbf{U}}_G \tilde{\mathbf{\Sigma}}_G \tilde{\mathbf{V}}_G^\top$ and let $\hat{\mathbf{G}}_0 = \tilde{\mathbf{U}}_{G,[:,1:r]} \tilde{\mathbf{\Sigma}}_{G,[1:r,1:r]} \tilde{\mathbf{V}}_{G,[:,1:r]}^\top$.

3) Let the frequency estimator $\hat{\mathbf{G}}$ be

$$\hat{\mathbf{G}} = (\hat{\mathbf{G}}_0)_+ / \|(\hat{\mathbf{G}}_0)_+\|_1,$$

where $\hat{\mathbf{G}}_0 = \tilde{\mathbf{U}}_{G,[:,1:r]} \tilde{\mathbf{\Sigma}}_{G,[1:r,1:r]} (\tilde{\mathbf{V}}_{G,[:,1:r]})^\top$. (14)

4) Let the estimator $\hat{\mathbf{Q}} \in \mathbb{R}^{p\times q}$ be

$$\hat{\mathbf{Q}}_{[i,:]} = \begin{cases} \hat{\mathbf{G}}_{[i,:]} / \sum_{j=1}^{q} \hat{\mathbf{G}}_{ij}, & \text{if } \sum_{j=1}^{q} \hat{\mathbf{G}}_{ij} > 0; \\ \frac{1}{q}, & \text{if } \sum_{j=1}^{q} \hat{\mathbf{G}}_{ij} = 0. \end{cases}$$

Output: $\hat{\mathbf{Q}}, \hat{\mathbf{G}}$.

Let $\tilde{r} = \|\mathbf{G}\|_F^2 / \sigma_r^2(\mathbf{G})$. Then,

$$\mathbb{E} \max_{1\leq i\leq p} \|\hat{\mathbf{Q}}_{[i,:]} - \mathbf{Q}_{[i,:]}\|_1$$

$$\leq C\sqrt{\frac{(p \vee q)\tilde{r}}{n} \cdot \frac{\kappa^3}{(p\pi_{\min})^2} \cdot \tau_* \log^2(n)} \wedge 2. \quad (16)$$

Note that estimating a square probability matrix is a special case of estimating rectangular matrices. So our lower bound result given by Theorem 2 is also relevant in the setting of general transition matrices. It suggests that the total variation bounds given in Theorem 3 are sharp in their dependence on p, r and n, provided that other parameters are bounded by constant factors.

V. SPECTRAL STATE COMPRESSION OF NEARLY LOW-RANK MARKOV CHAINS

In this section we consider general Markov processes with full rank. Our aim is to recover the principal subspace associated with $\mathbf{P}$ that is spanned by the leading Markov features. We also provide two state clustering methods that are able to partition the state space into disjoint blocks in accordance with the aggregability or lumpability.

A. Estimating the Leading Markov Feature Subspace

As noted in Section III, spectral decomposition of Markov chains provides feature functions that can be used to represent operators and functions on the state space. The Markov features also correspond to the block-partition membership when aggregability holds. Now we aim to estimate the space spanned by leading Markov features.

Let the singular value decomposition of $\mathbf{P}$ and $\mathbf{F}$ be

$$
\begin{aligned}
\mathbf{P} &= [\mathbf{U}_P \ \mathbf{U}_{P\perp}] \begin{bmatrix} \mathbf{\Sigma}_{P1} & 0 \\ 0 & \mathbf{\Sigma}_{P2} \end{bmatrix} \cdot \begin{bmatrix} \mathbf{V}_P^\top \\ \mathbf{V}_{P\perp}^\top \end{bmatrix}, \\
\mathbf{F} &= [\mathbf{U}_F \ \mathbf{U}_{F\perp}] \begin{bmatrix} \mathbf{\Sigma}_{F1} & 0 \\ 0 & \mathbf{\Sigma}_{F2} \end{bmatrix} \cdot \begin{bmatrix} \mathbf{V}_F^\top \\ \mathbf{V}_{F\perp}^\top \end{bmatrix},
\end{aligned} \tag{17}
$$

where $\mathbf{U}_P, \mathbf{V}_P, \mathbf{U}_F, \mathbf{V}_F \in \mathbb{O}_{p,r}$, $\mathbf{U}_{P\perp}, \mathbf{V}_{P\perp}, \mathbf{U}_{F\perp}, \mathbf{V}_{F\perp} \in \mathbb{O}_{p,p-r}$, $\mathbf{\Sigma}_{P1}, \mathbf{\Sigma}_{P2}, \mathbf{\Sigma}_{F1}, \mathbf{\Sigma}_{F2}$ are diagonal matrices with non-increasing order of diagonal entries. It is noteworthy that $\mathbf{V}_P$ and $\mathbf{V}_F$ represent the same subspace when $\mathbf{P}$ or $\mathbf{F}$ is of exactly rank-r, since $\mathbf{F} = \mathrm{diag}(\pi) \cdot \mathbf{P}$. We use the following estimators for the leading singular vectors of $\mathbf{F}$ and $\mathbf{P}$,

$$
\begin{aligned}
\hat{\mathbf{U}}_F &= \mathrm{SVD}_r\left(\tilde{\mathbf{F}}\right) = \text{leading } r \text{ left singular vectors of } \tilde{\mathbf{F}}; \\
\hat{\mathbf{V}}_F &= \mathrm{SVD}_r\left(\tilde{\mathbf{F}}^\top\right) = \text{leading } r \text{ right singular vectors of } \tilde{\mathbf{F}}; \\
\hat{\mathbf{U}}_P &= \mathrm{SVD}_r\left(\tilde{\mathbf{P}}\right) = \text{leading } r \text{ left singular vectors of } \tilde{\mathbf{P}}; \\
\hat{\mathbf{V}}_P &= \mathrm{SVD}_r\left(\tilde{\mathbf{P}}^\top\right) = \text{leading } r \text{ right singular vectors of } \tilde{\mathbf{P}},
\end{aligned} \tag{18}
$$

where $\tilde{\mathbf{F}}$ and $\tilde{\mathbf{P}}$ are given by (5) and (6), respectively. By using matrix norm concentration inequalities for $\tilde{\mathbf{F}}, \tilde{\pi}$ and singular value perturbation analysis, we prove the following angular error bounds for the subspace estimators.

Theorem 4 (Feature Space Recovery Bounds). *Let the assumptions of Theorem 1 hold, let $n \geq C\tau_* p \log^2(n)$. Then the estimators $\hat{\mathbf{U}}_F, \hat{\mathbf{V}}_F, \hat{\mathbf{U}}_P, \hat{\mathbf{V}}_P$ given by (18) satisfy*

$$
\mathbb{E}\left(\|\sin\Theta(\hat{\mathbf{U}}_F, \mathbf{U}_F)\| \vee \|\sin\Theta(\hat{\mathbf{V}}_F, \mathbf{V}_F)\|\right)
$$
$$
\leq \frac{C\sqrt{1/n \cdot \pi_{\max} \cdot \tau_* \log^2(n)}}{\sigma_r(\mathbf{F}) - \sigma_{r+1}(\mathbf{F})} \wedge 1, \tag{19}
$$

$$
\mathbb{E}\left(\|\sin\Theta(\hat{\mathbf{U}}_P, \mathbf{U}_P)\| \vee \|\sin\Theta(\hat{\mathbf{V}}_P, \mathbf{V}_P)\|\right)
$$
$$
\leq \frac{C\|\mathbf{P}\|\sqrt{1/n \cdot \pi_{\max}/\pi_{\min}^2 \cdot \tau_* \log^2(n)}}{\sigma_r(\mathbf{P}) - \sigma_{r+1}(\mathbf{P})} \wedge 1, \tag{20}
$$

where C is a universal constant.

In parallel, we study the theoretical error lower bounds for estimating the leading Markov feature spaces. Let the class of approximately low-rank stochastic matrices be

$$
\begin{aligned}
\mathcal{F}_{p,r,\delta} &= \left\{\mathbf{F} \in \mathcal{F}_p : \sigma_r(\mathbf{F}) - \sigma_{r+1}(\mathbf{F}) \geq \delta\right\}, \\
\mathcal{F}_{p,r,\delta}^* &= \left\{\mathbf{F} \in \mathcal{F}_{p,r,\delta} : \tau_* = 1, \pi_{\max} = \pi_{\min} = 1/p\right\}, \\
\mathcal{P}_{p,r,\delta} &= \left\{\mathbf{P} \in \mathcal{P}_p : (\sigma_r(\mathbf{P}) - \sigma_{r+1}(\mathbf{P}))/\|\mathbf{P}\| \geq \delta\right\}, \\
\mathcal{P}_{p,r,\delta}^* &= \left\{\mathbf{P} \in \mathcal{P}_{p,r,\delta} : \tau_* = 1, \pi_{\max} = \pi_{\min} = 1/p\right\}.
\end{aligned}
$$

Here, $\mathcal{P}_p$ and $\mathcal{F}_p$ represent the p-by-p transition and frequency matrix classes respectively, whose rigorous definitions are given in (54) and (55) in the supplementary materials.

Theorem 5 (Lower Bound for estimating the leading subspace). *Suppose that $2 \leq r \leq p/2$, $\delta \leq 1/(4p\sqrt{2})$ and*

Algorithm 3 Spectral State Aggregation

Input: $X_1, \ldots, X_n, r$

1) Construct the empirical frequency matrix $\tilde{\mathbf{F}}$ using (5).

2) Estimate the *left* Markov features $\hat{\mathbf{U}}_P$ using (18).
3) Solve the optimization problem

$$
\hat{\mathbf{\Omega}}_1, \ldots, \hat{\mathbf{\Omega}}_r = \arg\min_{\hat{\mathbf{\Omega}}_1, \ldots, \hat{\mathbf{\Omega}}_r} \min_{\bar{v}_1, \ldots, \bar{v}_r \in \mathbb{R}^r} \sum_{s=1}^{r} \sum_{i \in \hat{\mathbf{\Omega}}_s} \|(\hat{\mathbf{U}}_P)_{[i,:]} - \bar{v}_s\|_2^2.
$$

Output: Blocks $\hat{\mathbf{\Omega}}_1, \ldots, \hat{\mathbf{\Omega}}_r$

$\delta' \leq 1/(4\sqrt{2})$. *Then for sufficiently large p we have*

$$
\inf_{\hat{\mathbf{U}}_F, \hat{\mathbf{V}}_F} \sup_{\mathbf{F} \in \mathcal{F}_{p,r,\delta}^*} \mathbb{E}\left(\|\sin\Theta(\hat{\mathbf{U}}_F, \mathbf{U}_F)\| \wedge \|\sin\Theta(\hat{\mathbf{V}}_F, \mathbf{V}_F)\|\right)
$$
$$
\geq c\left(\frac{\sqrt{1/(np)}}{\delta} \wedge 1\right),
$$

$$
\inf_{\hat{\mathbf{U}}_P, \hat{\mathbf{V}}_P} \sup_{\mathbf{P} \in \mathcal{P}_{p,r,\delta'}^*} \mathbb{E}\left(\|\sin\Theta(\hat{\mathbf{U}}_P, \mathbf{U}_P)\| \wedge \|\sin\Theta(\hat{\mathbf{V}}_P, \mathbf{V}_P)\|\right)
$$
$$
\geq c\left(\frac{\sqrt{p/n}}{\delta'} \wedge 1\right),
$$

where $\hat{\mathbf{U}}_P, \hat{\mathbf{V}}_P, \hat{\mathbf{U}}_F, \hat{\mathbf{V}}_F$ are arbitrary estimators, c is a universal constant. The same relations also hold for $\mathcal{F}_{p,r,\delta}$ and $\mathcal{P}_{p,r,\delta}$.

The proofs of Theorems 4, 5 traced back to the analysis of classic PCA in multivariate analysis [68]. Our method is similar to PCA in the sense that they are both based on the factorization of some matrix that is estimated from data. It differs from PCA and aims to exact the Markov features that capture the *mean* transition kernel of dependent data.

B. Spectral State Compression for Aggregable Markov Chain

Next we develop an unsupervised state compression method based on the state aggregation model. According to Definition 3, a Markov chain is aggregable if the states can be partitioned into a few groups such that the states from the same group possess the identical transition distribution. In this case, $\mathbf{P}$ is low-rank and the leading left singular subspace of $\mathbf{U}_P$ exhibits piecewise constant structure in accordance with the group partition (Prop. 3). To estimate the group partition from empirical transitions, we propose the following method. In Step 2, the optimization problem is a combinatorial one. In practice, we can use discrete optimization solvers like k-means to find an approximate solution.

We evaluate the state aggregation method using the following misclassification rate

$$
M(\hat{\mathbf{\Omega}}_1, \ldots, \hat{\mathbf{\Omega}}_r) = \min_{\rho} \sum_{j=1}^{r} \frac{|\{i : i \in \mathbf{\Omega}_j, \text{ but } i \notin \hat{\mathbf{\Omega}}_{\rho(j)}\}|}{|\mathbf{\Omega}_j|}, \tag{21}
$$

where ρ is any permutation among the r group. We prove the following misclassification rate upper bound. The proof is given in the supplementary materials.

Theorem 6 (Misclassification Rate of Spectral State Aggregation). *Suppose the assumptions in Theorem 1 hold and the Markov chain is aggregable with respect to groups* $\{\Omega_1, \ldots, \Omega_r\}$. *Assume* $n \geq C\tau_* p \log^2(n)$. *The estimated partition* $\hat{\Omega}_1, \ldots, \hat{\Omega}_r$ *given by Alg. 3 satisfies*

$$\mathbb{E}M(\hat{\Omega}_1, \ldots, \hat{\Omega}_r) \leq \frac{C\|\mathbf{P}\|^2 pr \cdot \tau_* \log^2(n) \cdot \pi_{\max}/(\pi_{\min}^2 p)}{n\sigma_r^2(\mathbf{P})} \wedge r.$$

We remark that the state aggregation structure can only be uncovered from the *left* Markov features. As suggested by Prop. 3, the left Markov features of a state-aggregable Markov process exhibit a block structure that corresponds to the latent partition. The right features do not carry such information.

The proposed method of spectral state aggregation can be viewed as a special variant of clustering. It provides an unsupervised approach to identify partition/patterns from random walk data. Similar to many known clustering methods, spectral state aggregation is based on spectral decomposition of some kernel matrix, and it is related to latent-variable models. Yet there is a critical distinction. While standard clustering methods are typically based on some similarity metric, spectral state aggregation is based on the notion of preserving the state-to-state transition dynamics of the time series. One may view that spectral state clustering yields a partition mapping E from the state space into a smaller alphabet such that $\mathbb{P}(X_{t+1} \mid X_t) \approx \mathbb{P}(X_{t+1} \mid E(X_t))$. It can be interpreted as a form of state compression while preserving the predictability of the state variables.

C. Spectral State Compression for Lumpable Markov Chain

Finally we develop the state compression method for lumpable Markov process. Recall the discussion in Section III, the Markov chain is lumpable with respect to partition $\Omega_1, \ldots, \Omega_r \subseteq \{1, \ldots, p\}$, if the original p states can be compressed into r groups, where the law of walkers on $\{\Omega_1, \ldots, \Omega_r\}$ remains a Markov chain. Our goal is to identify the partition according to lumpability. Recall from Proposition 4 and additional discussions in its proof (see Section VIII), the transition matrix $\mathbf{P}$ and frequency matrix $\mathbf{F}$ do not have to be low-rank. Instead, they admit the decompositions of the form $\mathbf{P} = \mathbf{P}_1 + \mathbf{P}_2$ and $\mathbf{F} = \mathbf{F}_1 + \mathbf{F}_2$, where $\mathbf{P}_1$ and $\mathbf{F}_1$ are rank-r, and $\mathbf{U}_{\mathbf{P}_1}, \mathbf{V}_{\mathbf{P}_1}$ and $\mathbf{V}_{\mathbf{F}_1}$ have piece-wise constant columns that correspond to the block partition structure. Thus we propose the following spectral method for recovering the lumpable partition.

We remark that the spectral lumpable partition method is based on analyzing the *right* Markov features, i.e., the matrix of leading singular vectors $\hat{\mathbf{V}}_{F_1}$. This is because that empirically we find that the right Markov features can be typically estimated more accurately.

Theorem 7 (Misclassification Rate of Spectral Lumpable Partition). *Under the setting of Theorem 1, assume the Markov process is lumpable with respect to* $\Omega_1, \ldots, \Omega_r$, *and* $n \geq C\tau_* p \log^2(n)$. *Suppose the partitions* $\hat{\Omega}_1, \ldots, \hat{\Omega}_r$ *are obtained*

Algorithm 4 Spectral Lumpable Partition

Input: $X_1, \ldots, X_n, r$.

1. Evaluate the leading r *right* singular vectors for the empirical frequency matrix $\tilde{\mathbf{F}}$,

$$\hat{\mathbf{V}}_F = \text{SVD}_r\left(\tilde{\mathbf{F}}^\top\right),$$

where $\tilde{\mathbf{F}} = \left(\tilde{\mathbf{F}}_{ij}\right)_{1 \leq i, j \leq p}, \tilde{\mathbf{F}}_{ij} = \frac{1}{n}\sum_{k=1}^{n} 1_{\{X_{k-1}=i, X_k=j\}}.$

2. Solve the optimization problem

$$\hat{\Omega}_1, \ldots, \hat{\Omega}_r = \underset{\hat{\Omega}_1, \ldots, \hat{\Omega}_r}{\arg\min} \min_{\bar{v}_1, \ldots, \bar{v}_r \in \mathbb{R}^r} \sum_{s=1}^{r} \sum_{i \in \hat{\Omega}_s} \|(\hat{\mathbf{V}}_F)_{[i,:]} - \bar{v}_s\|_2^2. \quad (22)$$

Output: Block partition $\hat{\Omega}_1, \ldots, \hat{\Omega}_r$

---

by (22). *Then*

$$\mathbb{E}M(\hat{\Omega}_1, \ldots, \hat{\Omega}_r)$$
$$\leq \frac{C\pi_{\max}r\tau_* \log^2(n)/n + (r\|\mathbf{F}_2\|^2) \wedge \|\mathbf{F}_2\|_F^2}{\sigma_r^2(\mathbf{F}_1)} \wedge r.$$

Note that if the singular vectors of $\mathbf{F}_1$ are not the leading ones for the full matrix $\mathbf{R}$, it means $\sigma_r(\mathbf{F}_1) \leq \|\mathbf{F}_2\|$, and the error bound above becomes large. Therefore one can only recover the lumpable partition accurately if the random walk on the groups correspond to leading dynamics of the process.

VI. NUMERICAL STUDIES

A. Simulation Analysis

We simulate random walk trajectories to test the state compression procedures against naive empirical estimators. Let $\mathbf{P}_0 = \mathbf{U}_0\mathbf{V}_0^\top$, where $\mathbf{U}_0$ and $\mathbf{V}_0$ are two $p \times r$ matrices with i.i.d. standard normal entries in absolute values. Then we normalize each row to obtain a rank-r stochastic matrix $\mathbf{P}$, i.e., $\mathbf{P}_{[i,:]} = (\mathbf{P}_0)_{[i,:]}/\sum_{j=1}^{p}(\mathbf{P}_0)_{ij}$. Let $p = 200, r = 3, n = \text{round}(kpr\log^2(p))$, where k is a tuning integer. For each parameter setting, we conduct experiment for 100 independent trials and plot the mean estimation errors in Figure 3. We also conduct the experiments where both n, p vary. The results are plotted in Figure 4, where we let $r = 3, n = \text{round}(kpr\log^2(p))$, $p \in [100, 1000]$ and $k \in [2, 12]$. Figures 3-4 suggests that the spectral estimators $\hat{\mathbf{F}}, \hat{\mathbf{P}}$ significantly outperform the empirical estimators $\tilde{\mathbf{F}}, \tilde{\mathbf{P}}$ in all parameter settings. They also show that the subspaces spanned by leading Markov features can be estimated efficiently. We observe that $\hat{\mathbf{V}}_P$ tends to have smaller estimation error than $\hat{\mathbf{U}}_P$, although they enjoy the same error bounds (Theorems 4 and 5). This is because the theoretical results are mainly focused the errors' dependence on p, r, n. It remains open how do the estimation errors of $\|\sin\Theta(\hat{\mathbf{V}}_P, \mathbf{U}_P)\|$ and $\|\sin\Theta(\hat{\mathbf{U}}_P, \mathbf{U}_P)\|$ depend on the stationary distribution π. Our observations suggest that $\hat{\mathbf{U}}_P$ is more sensitive to the stationary distribution, especially when $\pi_{\min}$ is small.
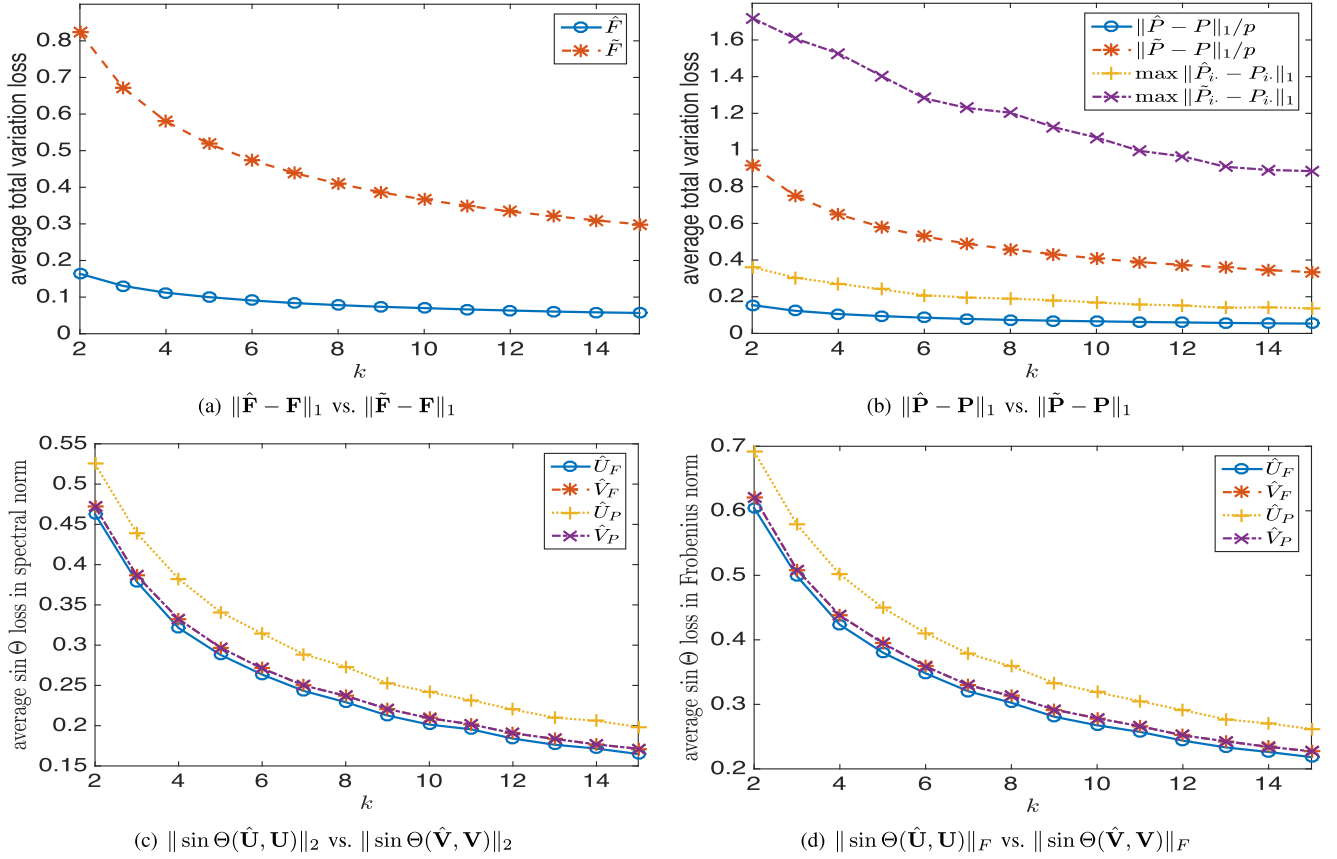
Fig. 3. Spectral estimators obtained by state compression based on sample paths of length $n = \text{round}(kpr \log^2(p))$. Plots (a) and (b) suggest that the spectral low-rank estimators are substantially more accurate than the empirical estimators, validating the bounds given by Theorem 1. Plots (c) and (d) suggest that one can estimate the principal subspace spanned by the leading Markov features efficiently, validating the bounds given by Theorem 4. In particular, $\hat{\mathbf{U}}_P$ is noisier than the other three subspace estimators because it is the most sensitive to states that are rarely visited, validating the error bound (20).

Next, we investigate the scenario that the invariant distribution π is "imbalanced", in the sense that $\pi_{\min}$ is small and some states appear much less frequently than the others. We generate the random walk data as follows. Let $\mathbf{P}_0 = \mathbf{U}_0 \mathbf{V}_0^\top$, where $\mathbf{U}_0, \mathbf{V}_0$ are generated similarly as the previous settings. Then we randomly generate I as a subset of $\{1, \ldots, p\}$ with cardinality $(p/2)$ and rescale transition probabilities from I^c to I by $1/\delta$. In this way, those states in I are visited less frequently in the long run when δ gets larger, corresponding to decreasing values of π_{min}. The numerical results in Figure 5 show that the estimation errors of $\hat{\mathbf{F}}$ stays roughly steady as δ varies. However the estimation errors of $\hat{\mathbf{P}}$ increases as the invariant distribution becomes more imbalanced. The reason is that those rows of $\hat{\mathbf{P}}$ corresponding to infrequent states in I become harder to estimate. This does not affect $\hat{\mathbf{F}}$ much, because the corresponding rows have smaller absolute values so they play a smaller role in the overall ℓ_1 error.

Lastly we test the spectral lumpable partition method for recovering the latent partition of a lumpable and full-rank Markov process. Let $\mathbf{P} = \mathbf{P}_1 + \mathbf{P}_2$ and let $\mathbf{P}_1 = \mathbf{Z}\bar{\mathbf{P}}\mathbf{Z}^\top$. Here, $\mathbf{Z} \in \mathbb{R}^{p \times r}$ is a randomly generated membership matrix where each row has one entry equality 1 and all other entries equaling 0s; $\bar{\mathbf{P}}$ is a randomly generated stochastic matrix given by $(\bar{\mathbf{P}})_{[i,:]} = (\mathbf{I}_r + \mathbf{B})_{[i,:]} / \|(\mathbf{I}_r + \mathbf{B})_{[i,:]}\|_1$, where $\mathbf{B} \overset{iid}{\sim} \text{Unif}[0, 1/2]$.

Let $\mathbf{P}_2$ be randomly generated as a low-rank matrix in a way to ensure the lumpability of the overall Markov chain. It can be verified that the Markov chain with transition matrix $\mathbf{P}$ generated from above is lumpable with respect to a partition of r groups. For various values of r, p, k, we test the spectral lumpable partition method (Algorithm 4) on sample paths of length $n = \text{round}(kpr \log^2(n))$. For each parameter setting, we repeat the experiment for 1000 independent trials and compute the averaged misclassification rates. The results are plotted in Figure 6, and they are consistent with the theoretical results in Section V-C.

B. Analysis of Manhattan Taxi Trips

We apply the state compression method to study the New York City Yellow Cab data.[2] The dataset contains 1.1×10^7 taxi trip records from 2016. Each record contains the information of one trip, including coordinates of pick-up/drop-off locations, starting/ending times of the trip, distance, length of trip, payment type and itemized fares. We view each trip as a transition from the pickup location to the dropoff location, so that the data is a collection of fragmented sample paths of a

[2] Data source: https://s3.amazonaws.com/nyc-tlc/trip+data/yellow_tripdata_2016-01.csv

(a) $\|\hat{\mathbf{F}} - \mathbf{F}\|_1$

(b) $\|\tilde{\mathbf{F}} - \mathbf{F}\|_1$

(c) $\frac{1}{p}\|\hat{\mathbf{P}} - \mathbf{P}\|_1$ (lower) vs $\frac{1}{p}\|\tilde{\mathbf{P}} - \mathbf{P}\|_1$ (upper)

(d) $\max_i \|\hat{\mathbf{P}}_{i\cdot} - \mathbf{P}_{i\cdot}\|_1$ (lower) vs $\max_i \|\tilde{\mathbf{P}}_{i\cdot} - \mathbf{P}_{i\cdot}\|_1$ (upper)
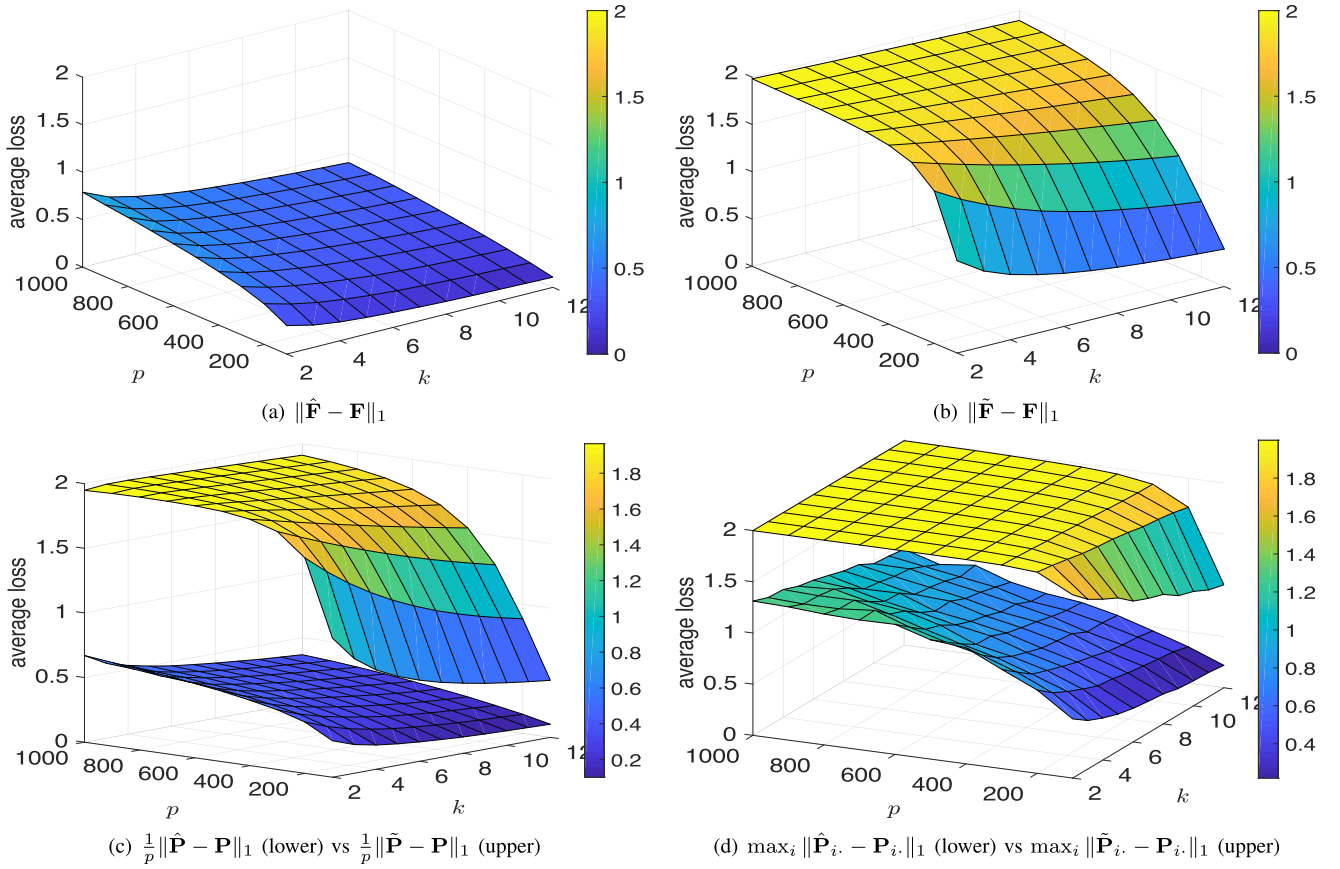
Fig. 4. Total variation errors of the estimators $\hat{\mathbf{F}}, \hat{\mathbf{P}}$ with growing dimension p and sample size n. Here $n = \mathrm{round}(kpr\log^2(p))$ and k is a tuning integer. The spectral estimators $\hat{\mathbf{F}}, \hat{\mathbf{P}}$ consistently outperform the empirical estimators $\tilde{\mathbf{F}}, \tilde{\mathbf{P}}$ in all parameter settings.
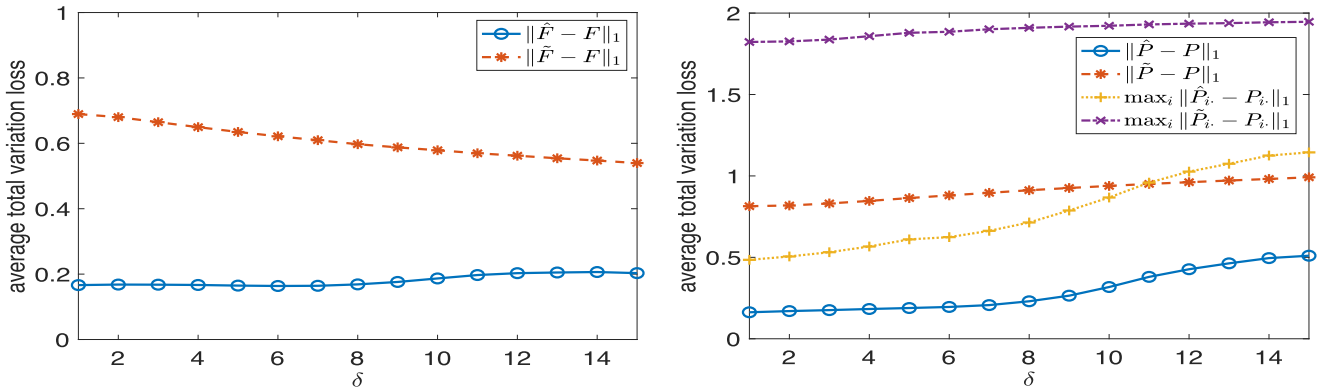

Fig. 5. Average estimation losses of $\hat{\mathbf{F}}, \tilde{\mathbf{F}}$ and $\hat{\mathbf{P}}, \tilde{\mathbf{P}}$ for Markov processes with imbalanced invariant distribution. Here larger values of δ indicates more severe imbalance in the invariant distribution (e.g., smaller values of π_{min}).

city-wide Markov process. For more analysis on such taxi-trip data, see for examples [1], [2].

We apply state compression to analyze the NYC taxi-trip dynamics. The first step is to preprocess the data by discretizing the map of Manhattan into a fine grid and treat each taxi trip as a single transition between the two cells that contain the pick-up and drop-off locations respectively. We remove those states (aka cells) with less than 200 total visits in a year (i.e., total number of pick-ups anddrop-offs), yielding approximate 5000 states. Then we compute the empirical transition matrix

$\tilde{\mathbf{P}}$ from the taxi trips. See Figure 7 for the singular values of $\tilde{\mathbf{P}}$.

In order to estimate the left Markov features and the citywide state aggregation structure, we apply the spectral state aggregation methods given by Eq. (18) and Alg. 3. Figure 8 plots the top four estimated Markov features, in comparison with the empirical frequency of visits. Figure 9 plots the citywide partition identified using Algorithm 3 with various values of r. In the case where $r = 5$, we obtain five clusters as shown in the first panel of Figure 9. The five
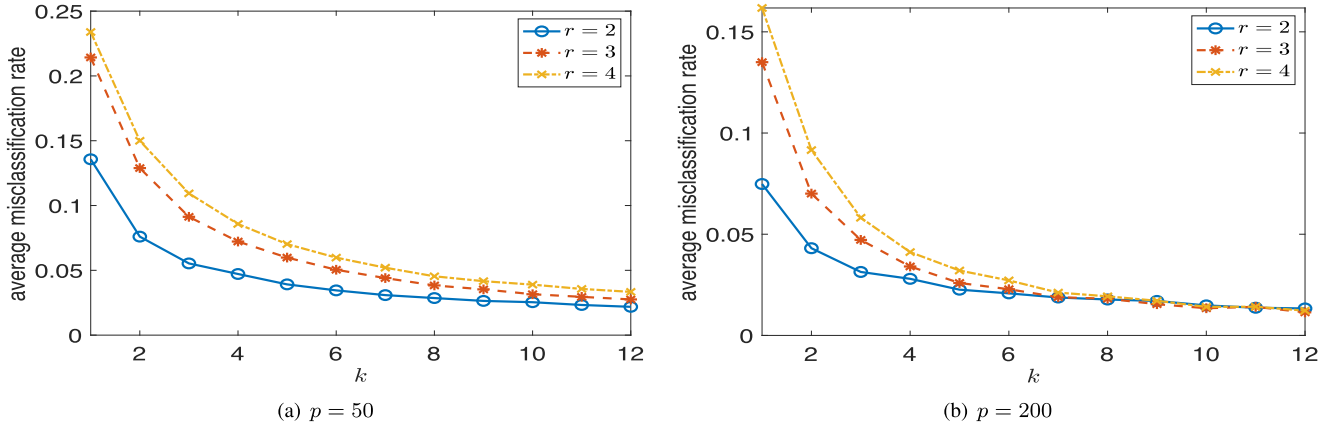
Fig. 6. Misclassification rate for recovery the lumpable partition. The input data are sample paths of length $n = \text{round}(kpr \log^2(n))$, where k is a tuning parameter.
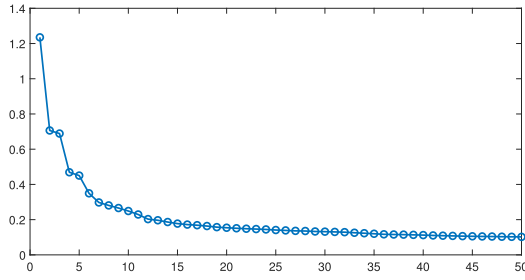

Fig. 7. Singular values of the empirical transition matrix $\tilde{\mathbf{P}}$ from the NYC taxi data.

clusters roughly correspond to: (1) Upper west side: residential areas (red); (2) Upper east side: residential areas (yellow); (3) Middleton: central business area (blue); (4) Lower west Manhattan (pink); (5) lower east Manhattan (green). When r is further increased to 9 and 12, the state aggregation method uncovers a finer citywide partition according to transition patterns of the taxi trips. For comparison, we implement the k-means clustering method directly on rows of $\tilde{\mathbf{F}}$ and $\tilde{\mathbf{P}}$ and plot the results in Figure 9 (b), which yield less interpretable results.

It is worth noting that our state compression method does not use any information about the geospatial proximity between locations. The partition is obtained to maximally preserve the transition dynamics of taxi trips. The experiment reveals an informative partition of the Manhattan city, which suggest that passengers who depart from the same zone share similar distributions of their destinations.

Finally we analyze the taxi trips by taking into consideration the time of the trips. We consider three time segments: morning 6:00-11:59am, afternoon 12:00-17:59pm and evening 18:00-23:59pm. We stratify the data according to these segments and apply the state compression methods to analyze trips within each segment separately. The results are presented in Figure 10. Indeed, the traffic pattern varies throughout the day. In particular, the morning-time state aggregation result differs significantly from the partition structure learned from trips in the afternoons and evenings.

VII. SUMMARY

Markov process is the most basic stochastic systems. Therefore we believe that state compression of the Markov process is naturally the first topic to investigate before moving on to more complicated problems. In this article, we studied the spectral decomposition of Markov processes and its connections to latent-variable process, aggregability and lumpability. We proposed a class of spectral state compression methods for analyzing Markov state trajectories, and established minimax upper and lower bounds for the estimation errors. For special cases where the Markov process is state-aggregatable or lumpable, we show that one can recover the underlying partition structure with theoretical guarantees. The numerical studies on both synthetic and real datasets illustrate the merits of the proposed methods. We hope that establishing the spectral state compression theory for Markov process would shed light on the estimation and system identification of higher-order processes that are not necessarily Markovian.

VIII. PROOFS FOR PROPERTIES OF LOW-RANK MARKOV CHAINS

A. Proof of Proposition 1

Let $\mathbf{F}, \mathbf{G} \in \mathbb{R}^{p \times r}$, $\mathbf{F}_{[:,k]} = f_k, \mathbf{G}_{[:,k]} = g_k, 1 \leq k \leq r$. By definition, $\mathbf{C} = \mathbf{G}^\top \mathbf{F}$ is non-degenerate. This implies $\mathbf{F}$ and $\mathbf{G}$ are both non-singular, and $\text{rank}(\mathbf{P}) = \text{rank}(\mathbf{FG}^\top) = r$.

Next,

$$\mathbf{P}^n = \overbrace{\mathbf{FG}^\top \mathbf{FG}^\top \cdots \mathbf{FG}^\top}^{n} = \mathbf{F}(\mathbf{G}^\top \mathbf{F})^{n-1} \mathbf{G}^\top = \mathbf{FC}^{n-1} \mathbf{G}^\top.$$

If π is the invariant distribution, π satisfies $\pi^\top \mathbf{P} = \pi^\top$. Let $\gamma = \mathbf{F}^\top \pi$. Then, γ satisfies

$$\mathbf{G}\gamma = \mathbf{GF}^\top \pi = \mathbf{P}^\top \pi = \pi, \quad \Rightarrow \quad \pi(x) = \sum_{k=1}^{r} \gamma_k g_k(x),$$

$$\gamma^\top \mathbf{CG}^\top = \pi^\top \mathbf{FCG}^\top = \pi^\top \mathbf{P}^2 = \pi^\top \mathbf{P} = \pi^\top \mathbf{FG}^\top = \gamma^\top \mathbf{G}^\top.$$

Since $\mathbf{G}^\top$ is non-singular, the previous equality implies $\gamma^\top \mathbf{C} = \gamma^\top$. $\square$

(a) Top Markov features estimated via state compression (quantile heat map)   (b) Distribution $\hat{\pi}$
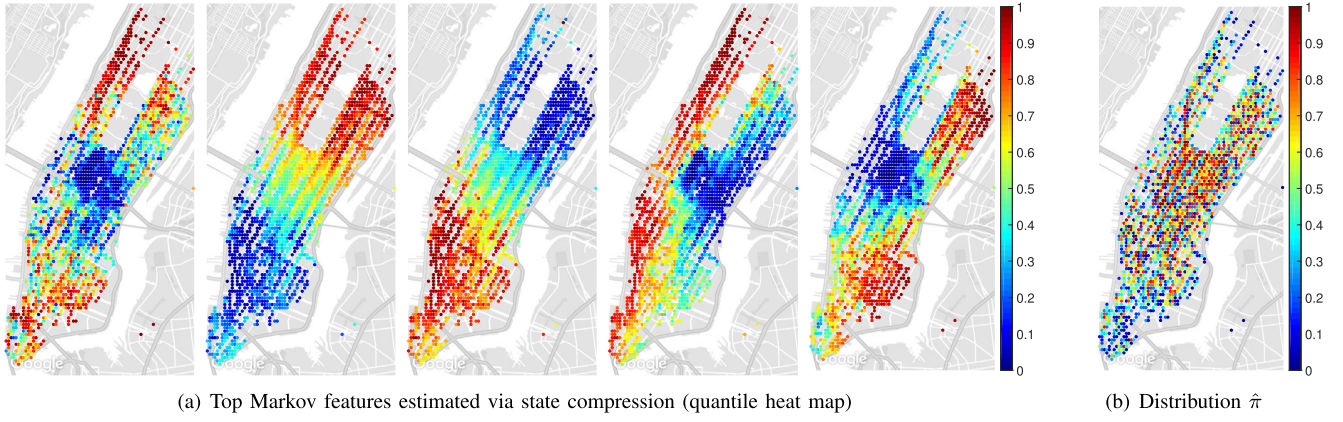
Fig. 8.   The leading Markov features yielded by spectral state compression reveal transition patterns across the city of Manhattan. Comparing (a) and (b), the transition patterns appearing in the leading Markov features cannot be learned from the empirical stationary distribution.

(a) Spectral state aggregation ($r = 5, 9, 12$)   (b) Clustering based on rows of $\tilde{\mathbf{F}}$, $\tilde{\mathbf{P}}$
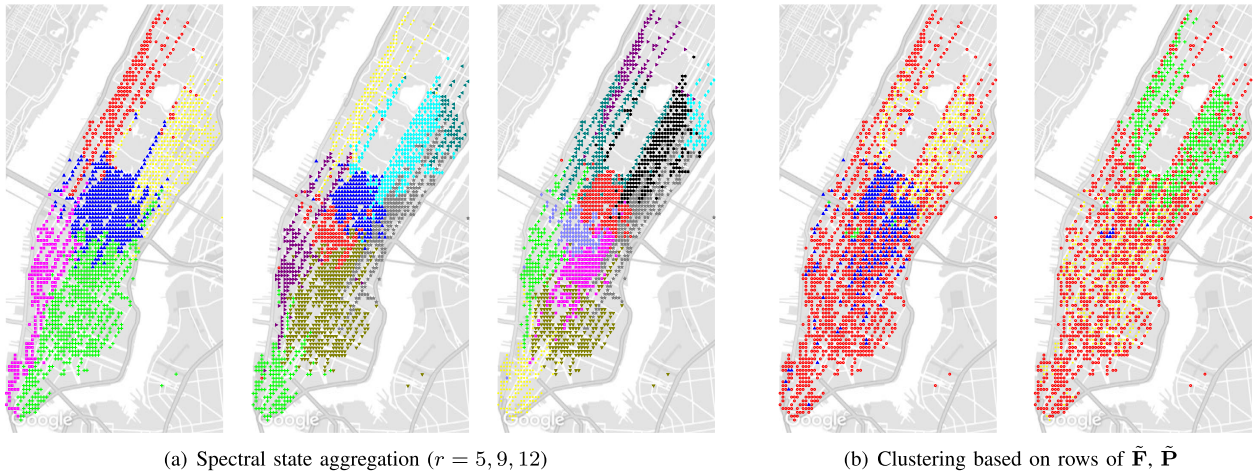
Fig. 9.   Spectral state aggregation applied to NYC taxi trips finds a citywide partition. In (a), each colored zone corresponds to an area from which taxi passengers share similar distributions over their destinations. In (b), clustering the row distributions does not yield a meaningful partition.

B. Proof of Proposition 2

If there exists a latent process $\{Z_t\} \subset [r]$ that satisfies (3), we have

$$\mathbf{P}_{ij} = \mathbb{P}\left(X_{t+1} = j \mid X_t = i\right)$$

$$= \sum_{l=1}^{r} \mathbb{P}\left(X_{t+1} = j \mid X_t = i, Z_t = l\right) \mathbb{P}\left(Z_t = l \mid X_t = i\right)$$

$$= \sum_{l=1}^{r} \mathbb{P}\left(X_{t+1} = j \mid Z_t = l\right) \mathbb{P}\left(Z_t = l \mid X_t = i\right)$$

$$:= \sum_{l=1}^{r} f_l(i) g_l(j),$$

where $f_1, \ldots, f_r$ and $g_1, \ldots, g_r$ are set as

$$f_l(i) = \mathbb{P}\left(Z_t = l \mid X_t = i\right), \quad g_l(j) = \mathbb{P}\left(X_{t+1} = j \mid Z_t = l\right)$$

for any $i, j, 1 \le l \le r$. Then both f_i and g_i are non-negative and g_i is a probability mass function

$$\sum_j g_l(j) = \sum_j \mathbb{P}\left(X_{t+1} = j \mid Z_t = l\right) = 1.$$

On the other hand, if the Markov process has non-negative rank r, based on Definition 2, we have

$$\mathbb{P}(X_{t+1} = j \mid X_t = i) = \sum_{l=1}^{r} f_l(i) g_l(j).$$

We introduce another process $Z_t \subset [r]$ based on $X_0, X_1, \ldots$ as follows: for $k = 1, \ldots, r$,

$$\mathbb{P}\left(Z_t = k \mid X_{t+1} = j, X_t = i\right) = \frac{f_k(i) g_k(j)}{\sum_{l=1}^{r} f_l(i) g_l(j)}.$$

Based on the Markovian property of $\{X_t\}$ and the definition of Z_t, we have

$$\mathbb{P}\left(Z_t \mid X_t\right) = \mathbb{P}\left(Z_t \mid X_1, \ldots, X_t\right)$$

since Z_t only relies on X_t and X_{t+1}.

$$\mathbb{P}\left(X_{t+1} = j \mid Z_t = k, X_t = i\right)$$

$$= \frac{\mathbb{P}\left(Z_t = k \mid X_{t+1} = j, X_t = i\right) \cdot \mathbb{P}\left(X_{t+1} = j \mid X_t = i\right)}{\sum_{j'} \mathbb{P}\left(Z_t = k \mid X_{t+1} = j', X_t = i\right) \mathbb{P}\left(X_{t+1} = j' \mid X_t = i\right)}$$

$$= \frac{\frac{f_k(i) g_k(j)}{\sum_{l=1}^{r} f_l(i) g_l(j)} \cdot \sum_{l=1}^{r} f_l(i) g_l(j)}{\sum_{j'} \left(\frac{f_k(i) g_k(j')}{\sum_{l=1}^{r} f_l(i) g_l(j')} \cdot \sum_{l=1}^{r} f_l(i) g_l(j') \right)}$$

(a) Top four Markov features estimated via spectral state compression   (b)   5-cluster-Aggregation
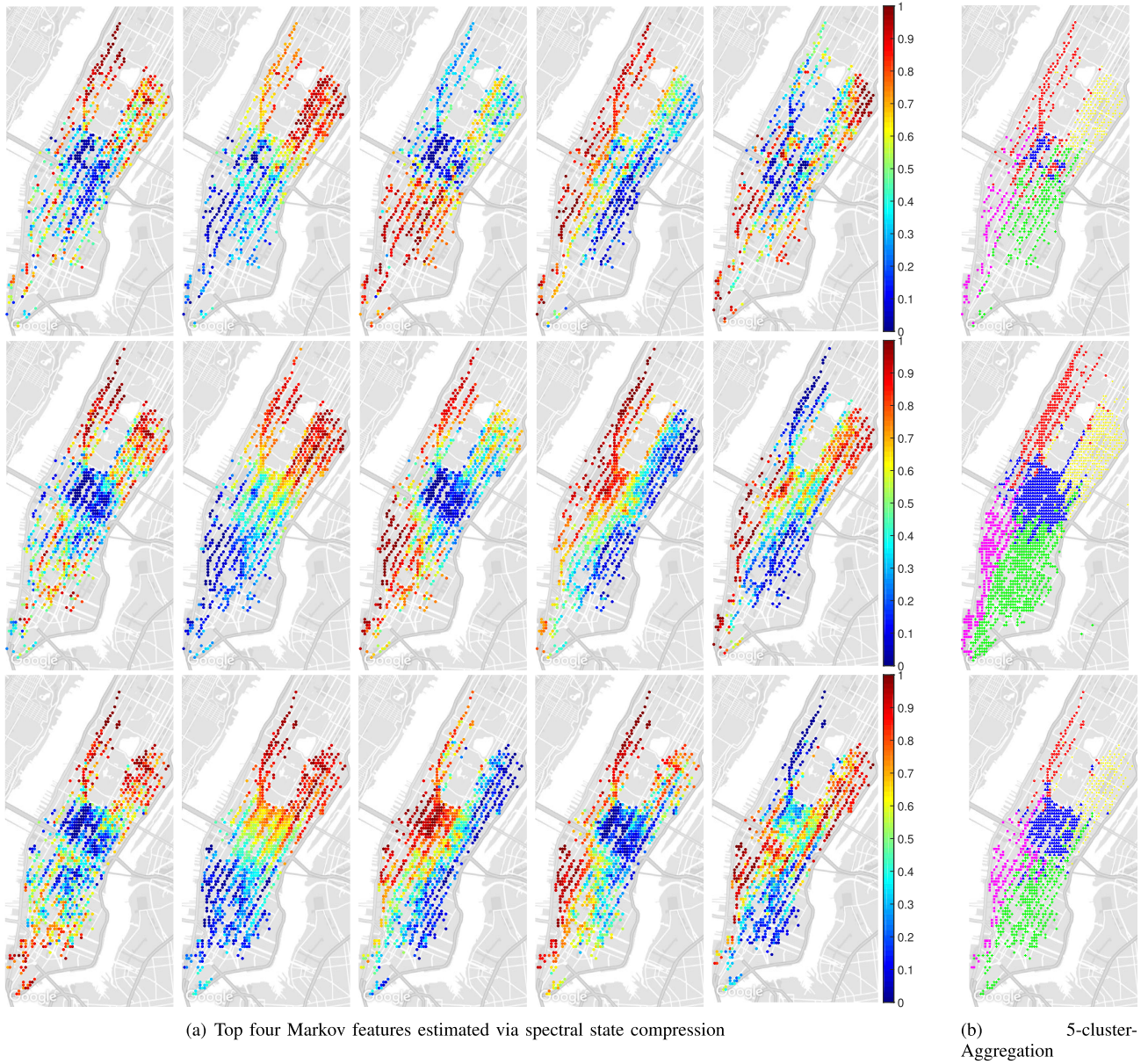
Fig. 10.   Per-time-segment results from applying spectral state compression to NYC taxi-trip data: mornings (upper row), afternoons (middle row) and evenings (lower row). One can see the leading Markov features vary throughout the day. The day-time state aggregation results differ significantly from that of the evening time.

$$= \frac{f_k(i) g_k(j)}{\left(\sum_{j'} f_k(i) g_k(j') \right)}$$

$$= \frac{f_k(i) g_k(j)}{f_k(i)} = g_k(j).$$

Here, we used the fact that g_l is a probability mass function so that $\sum_{j'} g_l(j') = 1$. Based on the previous calculation, we can see $\mathbb{P}(X_{t+1} = j \mid Z_t = l, X_t = i)$ is free of i, which means $\mathbb{P}(X_{t+1} \mid Z_t) = \mathbb{P}(X_{t+1} \mid Z_t, X_t) = \mathbb{P}(X_{t+1} \mid Z_t, X_0, \ldots, X_t)$. □

C. Proof of Proposition 3

We construct $\mathbf{G} = [g_1, \ldots, g_r] \in \mathbb{R}^{p \times r}$ as

$$\forall 1 \leq j \leq p, 1 \leq k \leq r,$$

$$\text{if } i \in \Omega_k, \quad \mathbf{G}_{jk} = g_k(j) = \mathbb{P}(X_{t+1} = j \mid X_t = i).$$

Then, $\mathbf{G}$ is non-negative and well-defined since $\mathbf{P}$ is state-aggretagable. Next, for any states $1 \leq i, j \leq p$, if $i \in \Omega_k$, we have

$$\mathbf{P}_{ij} = \mathbb{P}(X_{t+1} = j \mid X_t = i) = \mathbf{G}_{jk}$$

$$= \sum_{l=1}^{r} \mathbf{1}_{\Omega_l}(i) \mathbf{G}_{jl} = \sum_{l=1}^{r} \mathbf{1}_{\Omega_l}(i) g_l(j).$$

□

D. Proof of Proposition 4

Let $\bar{\mathbf{P}} \in \mathbb{R}^{r \times r}$, $\mathbf{P}_1 \in \mathbb{R}^{p \times p}$ and $\mathbf{P}_2 \in \mathbb{R}^{p \times p}$ be constructed as follows

$$\bar{\mathbf{P}}_{kl} = \sum_{b \in \Omega_l} \mathbf{P}_{ib} \quad \forall i \in \Omega_k; \forall k, l \in [r]$$

$$\mathbf{P}_1 = \mathbf{Z} \bar{\mathbf{P}} \text{diag}(|\Omega_1|^{-1}, \ldots, |\Omega_r|^{-1}) \mathbf{Z}^\top, \qquad \mathbf{P}_2 = \mathbf{P} - \mathbf{P}_1,$$

where $\mathbf{Z} = [\mathbf{1}_{\Omega_1}, \ldots, \mathbf{1}_{\Omega_r}] \in \mathbb{R}^{p \times r}$. Here $\bar{\mathbf{P}}$ is well-defined because of the lumpability, it is transition matrix of the random walk on the blocks $\Omega_1, \ldots, \Omega_r$.

For any $k, l \in [r], i \in \Omega_k$, we have

$$\mathbf{1}_i^\top \mathbf{P}_2 \mathbf{1}_{\Omega_l} = \mathbf{1}_i^\top (\mathbf{P} - \mathbf{P}_1) \mathbf{1}_{\Omega_l}$$
$$= \sum_{j \in \Omega_l} \mathbf{P}_{ij} - \sum_{j \in \Omega_l} \mathbf{P}_{1,ij} = \sum_{j \in \Omega_l} \mathbf{P}_{ij} - \sum_{i \in \Omega_l} \frac{1}{|\Omega_l|} \sum_{b \in \Omega_l} \mathbf{P}_{ib} = 0.$$

Since i, l can be arbitrary, we have $\mathbf{P}_2 \mathbf{1}_{\Omega_l} = 0$ therefore $\mathbf{P}_2 \mathbf{Z} = 0$. It follows that $\mathbf{P}_1 \mathbf{P}_2^\top = 0$.

Finally, let $\mathbf{P}_1 = \mathbf{U}_{P_1} \mathbf{\Sigma}_{P_1} \mathbf{V}_{P_1}^\top$ be the economic-size SVD. Then for any $k = 1, \ldots, r, i, i' \in \Omega_k$, i.e., i, i' belonging to the same block, by definition of $\mathbf{P}_1$ and $\mathbf{Z}$,

$$(\mathbf{P}_1)_{[i,:]} = \mathbf{Z}_{[i,:]} \bar{\mathbf{P}} \mathrm{diag}(|\Omega_1|^{-1}, \ldots, |\Omega_r|^{-1}) \mathbf{Z}^\top$$
$$= \mathbf{Z}_{[i',:]} \bar{\mathbf{P}} \mathrm{diag}(|\Omega_1|^{-1}, \ldots, |\Omega_r|^{-1}) \mathbf{Z}^\top = (\mathbf{P}_1)_{[i',:]}.$$

Then,

$$(\mathbf{U}_{P_1})_{[i,:]} = (\mathbf{P}_1)_{[i,:]} \mathbf{V}_{P_1} \mathbf{\Sigma}_{P_1}^{-1} = (\mathbf{P}_1)_{[i',:]} \mathbf{V}_{P_1} \mathbf{\Sigma}_{P_1}^{-1} = (\mathbf{U}_{P_1})_{[i',:]}.$$

By the same argument, we can also show $(\mathbf{V}_{P_1})_{[i,:]} = (\mathbf{V}_{P_1})_{[i',:]}$.

In fact, the frequency matrix $\mathbf{F}$ also has the similar decomposition since $\mathbf{F} = \mathrm{diag}(\pi)\mathbf{P}$,

$$\mathbf{F} = \mathbf{F}_1 + \mathbf{F}_2, \quad \text{where} \quad \mathbf{F}_1 = \mathrm{diag}(\pi)\mathbf{P}_1, \mathbf{F}_2 = \mathrm{diag}(\pi)\mathbf{P}_2.$$

For this decomposition, we also have

$$\mathbf{F}_1 \mathbf{F}_2^\top = \mathrm{diag}(\pi)\mathbf{F}_1 \mathbf{F}_2^\top \mathrm{diag}(\pi) = 0.$$

Although $\mathbf{F}_1$ is not necessarily symmetric and the columns of $\mathbf{F}_1$ may not have piece-wise constant structure, the rows of $\mathbf{F}_1$ is still piece-wise constant according to partition, i.e., $\mathbf{F}_{1,ij} = \mathbf{F}_{1,ij'}$ if j, j' belong to the same group. By the similar argument as the one for $\mathbf{P}_1$, we can show the right singular vectors $\mathbf{V}_{F_1}$ is also piece-wise constant, i.e., $(\mathbf{V}_{F_1})_{[i,:]} = (\mathbf{V}_{F_1})_{[i',:]}$ if i, i' belong to the same group.

IX. PROOFS FOR RESULTS OF SECTION IV

A. Proof of Theorem 1

(a) First we prove (9). Since $\hat{\mathbf{F}}, \mathbf{F} \geq 0$ and $\sum_{i,j} \hat{\mathbf{F}}_{ij} = \sum_{i,j} \mathbf{F}_{ij} = 1$, the trivial bound

$$\|\hat{\mathbf{F}} - \mathbf{F}\|_1 \leq 2 \tag{23}$$

holds, we only need focus on the case with additional assumption that

$$n \geq Cpr \cdot (\pi_{\max} p) \cdot \tau_* \log^2(n). \tag{24}$$

Given the previous assumption, Lemma 7 implies that there exists constants $C > 0$ and $c > 1$ such that

$$\mathbb{P}(\mathcal{A}) \geq 1 - n^{-c}, \mathcal{A} = \left\{ \max \left\{ \|\tilde{\mathbf{F}} - \mathbf{F}\|, \|\tilde{\pi} - \pi\|_\infty \right\} \right.$$
$$\left. \leq C \sqrt{\frac{\pi_{\max} \tau_* \log^2(n)}{n}} \right\}. \tag{25}$$

Assume that the probabilistic event $\mathcal{A}$ holds. Recall $\hat{\mathbf{F}}_0$ is the leading r principal components of $\tilde{\mathbf{F}}$ (Algorithm 1), Lemma 3 implies

$$\|\hat{\mathbf{F}}_0 - \mathbf{F}\|_F \leq 2\sqrt{2r} \|\tilde{\mathbf{F}} - \mathbf{F}\| \leq C \sqrt{\frac{r \pi_{\max} \tau_* \log^2(n)}{n}}.$$

Since $\hat{\mathbf{F}} = (\mathbf{F}_0)_+ / \|(\mathbf{F}_0)_+\|_1 \geq 0$ and $\|\mathbf{F}\|_1 = 1$, we have

$$\|\hat{\mathbf{F}} - \mathbf{F}\|_1 = \left\| \frac{(\hat{\mathbf{F}}_0)_+}{\|(\hat{\mathbf{F}}_0)_+\|_1} - \frac{\mathbf{F}}{\|\mathbf{F}\|_1} \right\|_1$$

$$\overset{\text{Lemma 2}}{\leq} 2\|(\hat{\mathbf{F}}_0)_+ - \mathbf{F}\|_1$$

$$= 2 \sum_{i=1}^p \sum_{j=1}^p |(\hat{\mathbf{F}}_{0,ij})_+ - \mathbf{F}_{ij}| \leq 2 \sum_{i=1}^p \sum_{j=1}^p |\hat{\mathbf{F}}_{0,ij} - \mathbf{F}_{ij}|$$

$$\leq 2p \left(\sum_{i=1}^p \sum_{j=1}^p |\hat{\mathbf{F}}_{0,ij} - \mathbf{F}_{ij}|^2 \right)^{1/2} = 2p \|\hat{\mathbf{F}}_0 - \mathbf{F}\|_F$$

$$\leq Cp \sqrt{\frac{r \pi_{\max} \tau_* \log^2(n)}{n}}$$

$$= C \sqrt{\frac{rp}{n} \cdot p\pi_{\max} \cdot \tau_* \log^2(n)}. \tag{26}$$

with probability at least $1 - n^{-c}$, because of (25). We finally have

$$\mathbb{E}\|\hat{\mathbf{F}} - \mathbf{F}\|_1 = \mathbb{E}\left[\|\hat{\mathbf{F}} - \mathbf{F}\|_1 \mathbf{1}_\mathcal{A}\right] + \mathbb{E}\left[\|\hat{\mathbf{F}} - \mathbf{F}\|_1 \mathbf{1}_{\mathcal{A}^c}\right]$$

$$\overset{(26)(23)}{\leq} C \sqrt{\frac{rp}{n} \cdot p\pi_{\max} \cdot \tau_* \log^2(n)} + 2 \cdot \mathbb{P}(\mathcal{A}^c)$$

$$\leq C \sqrt{\frac{rp}{n} \cdot p\pi_{\max} \cdot \tau_* \log^2(n)} + 2n^{-c}. \tag{27}$$

When $c > 1$, we obtain the error bound for $\hat{\mathbf{F}}$. In addition, if the probabilistic event $\mathcal{A}$ holds, we have

$$\|\hat{\mathbf{F}}^1 - \mathbf{F}\|_F \leq \|\hat{\mathbf{F}}^1 - \hat{\mathbf{F}}_0\|_F + \|\hat{\mathbf{F}}_0 - \mathbf{F}\|_F$$

$$\overset{(*)}{\leq} \|\mathbf{F} - \hat{\mathbf{F}}_0\|_F + \|\hat{\mathbf{F}}_0 - \mathbf{F}\|_F \leq C \sqrt{\frac{r \pi_{\max} \tau_* \log^2(n)}{n}}.$$

Here, (*) is due to the definition of $\hat{\mathbf{F}}^1$ and the fact $\mathbf{F}$ belongs to the probability simplex. Applying the previous argument again, we can show the same error bound holds for $\hat{\mathbf{F}}^1$.

(b) Next, we consider the average total variation error bound for $\hat{\mathbf{P}}$. Since $\hat{\mathbf{P}}_{[i,:]} = \frac{\hat{\mathbf{F}}_{[i,:]}}{\|\hat{\mathbf{F}}_{[i,:]}\|_1}$, $\mathbf{P}_{[i,:]} = \frac{\mathbf{F}_{[i,:]}}{\|\mathbf{F}_{[i,:]}\|_1}$, and $\|\mathbf{F}_{[i,:]}\|_1 = \pi_i \geq \pi_{\min}$, we have

$$\mathbb{E}\|\hat{\mathbf{P}} - \mathbf{P}\|_1 = \sum_{i=1}^p \mathbb{E}\|\hat{\mathbf{P}}_{[i,:]} - \mathbf{P}_{[i,:]}\|_1$$

$$\overset{\text{Lemma 2}}{\leq} \sum_{i=1}^p \mathbb{E}\frac{2\|\hat{\mathbf{F}}_{[i,:]} - \mathbf{F}_{[i,:]}\|_1}{\pi_{\min}}$$

$$\leq C \sqrt{\frac{r}{n} \cdot \frac{\pi_{\max}}{\pi_{\min}^2} \cdot \tau_* \log^2(n)}.$$

(c) Then, we consider the uniform row-wise bound (11). Recall $\mathbf{F} = \mathbf{U}_F \mathbf{\Sigma}_F \mathbf{V}_F^\top$, $\tilde{\mathbf{F}} = \tilde{\mathbf{U}}_F \tilde{\mathbf{\Sigma}}_F \tilde{\mathbf{V}}_F^\top$, $\hat{\mathbf{F}}_0 = \tilde{\mathbf{U}}_{F,[:,1:r]} \tilde{\mathbf{\Sigma}}_{F,[1:r,1:r]} \tilde{\mathbf{V}}_{F,[:,1:r]}^\top$. Without ambiguity, we simply note $\mathbf{U}_F, \mathbf{V}_F, \tilde{\mathbf{U}}_F, \tilde{\mathbf{V}}_F$, etc as $\mathbf{U}, \mathbf{V}, \tilde{\mathbf{U}}, \tilde{\mathbf{V}}$, etc. Then

$$\mathbf{F} = \mathbf{F}\Pi_{\mathbf{V}}, \quad \hat{\mathbf{F}}_0 = \hat{\mathbf{F}}_0 \Pi_{\tilde{\mathbf{V}}_{[:,1:r]}}, \tag{28}$$

$$\hat{\mathbf{P}}_{[i,:]} = (\hat{\mathbf{F}}_{0,[i,:]})_+ / \sum_{j=1}^p (\hat{\mathbf{F}}_{0,[i,j]})_+. \tag{29}$$

Here, $\Pi_{\mathbf{V}} = \mathbf{V}\mathbf{V}^\top$ and $\Pi_{\tilde{\mathbf{V}}_{[:,1:r]}} = \tilde{\mathbf{V}}_{[:,1:r]} \tilde{\mathbf{V}}_{[:,1:r]}^\top$ are the projection matrices on to the column space of $\mathbf{V}$ and $\tilde{\mathbf{V}}_{[:,1:r]}$, respectively. Since the trivial bound $\max_i \|\hat{\mathbf{P}}_{[i,:]} - \mathbf{P}_{[i,:]}\|_1 \le 2$ always hold, we can assume

$$n \ge C p \tilde{r} \frac{\kappa^3}{(p\pi_{\min})^2} \cdot \tau_* \log^2(n) \tag{30}$$

in the rest of the proof without loss of generality. Let $\delta = \sqrt{p/r} \max_j \|\mathbf{V}^\top e_j\|_2$ be the incoherence constant. We aim to develop a bound for δ. Since $\mathbf{F} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, we have $\|\mathbf{F}_{[:,j]}\|_2 = \|\mathbf{U}\mathbf{\Sigma}(\mathbf{V}_{[j,:]})^\top\|_2$ for any $1 \le j \le p$. On one hand,

$$\|\mathbf{F}_{[:,j]}\|_2 = \sqrt{\sum_{i=1}^p \mathbf{F}_{ij}^2} \le \frac{\kappa\sqrt{p}}{p^2} = \frac{\kappa}{p^{3/2}};$$

on the other hand,

$$\|\mathbf{U}\mathbf{\Sigma}(\mathbf{V}_{[j,:]})^\top\|_2 = \|\mathbf{\Sigma}(\mathbf{V}_{[j,:]})^\top\|_2 \ge \sigma_r(\mathbf{\Sigma})\|\mathbf{V}_{[j,:]}\|_2$$
$$= \frac{\|\mathbf{\Sigma}\|_F \|\mathbf{V}_{[j,:]}\|_2}{\sqrt{\tilde{r}}} = \frac{\|\mathbf{F}\|_F \|\mathbf{V}_{[j,:]}\|_2}{\sqrt{\tilde{r}}} \ge \frac{\|\mathbf{F}\|_1 \|\mathbf{V}_{[j,:]}\|_2}{p\sqrt{\tilde{r}}}$$
$$= \frac{\|\mathbf{V}_{[j,:]}\|_2}{p\sqrt{\tilde{r}}},$$

which means

$$\frac{\delta\sqrt{r/p}}{p\sqrt{\tilde{r}}} \le \frac{\max_j \|\mathbf{V}_{[j,:]}\|_2}{p\sqrt{\tilde{r}}} \le \frac{\kappa}{p^{3/2}}, \quad \Rightarrow \quad \tilde{r} \ge \delta^2 \, r/\kappa^2.$$

Additionally,

$$\pi_i = (\pi^\top \mathbf{P})_i = \pi^\top \mathbf{P} e_i = 1_p^\top \text{diag}(\pi)\mathbf{P}e_i = \sum_{j=1}^p \mathbf{F}_{ij} \le \kappa/p$$

$$\Rightarrow \quad \pi_{\max} \le \kappa/p.$$

Therefore, we can apply Lemma 8 and obtain

$$\max_{1 \le i \le p} \|(\tilde{\mathbf{F}}_{[i,:]} - \mathbf{F}_{[i,:]})\mathbf{V}\|_2 \le C\left(\frac{\pi_{\max}\delta^2 r \tau_* \log^2(n)}{pn}\right)^{1/2}$$
$$\le C\left(\frac{\kappa^3 \tilde{r}\tau_* \log^2(n)}{p^2 n}\right)^{1/2} \tag{31}$$

with probability at least $1 - n^{-c}$.
By Lemma 7,

$$\|\tilde{\mathbf{F}} - \mathbf{F}\| \le C\left(\frac{\pi_{\max}\tau_* \log^2(n)}{n}\right)^{1/2} \tag{32}$$

with probability at least $1 - n^{-c}$. Given (31) and (32) hold, we have

$$\|\hat{\mathbf{F}}_{0,[i,:]} - \mathbf{F}_{[i,:]}\|_2$$
$$\le \|\hat{\mathbf{F}}_{0,[i,:]}\Pi_{\mathbf{V}} - \mathbf{F}_{[i,:]}\|_2 + \|\hat{\mathbf{F}}_{0,[i,:]} - \hat{\mathbf{F}}_{0,[i,:]}\Pi_{\mathbf{V}}\|_2$$
$$= \|(\hat{\mathbf{F}}_{0,[i,:]} - \mathbf{F}_{[i,:]})\Pi_{\mathbf{V}}\|_2$$
$$\quad + \|\hat{\mathbf{F}}_{0,[i,:]}(\Pi_{\tilde{\mathbf{V}}_{[:,1:r]}} - \Pi_{\mathbf{V}})\|_2$$
$$\le \|(\hat{\mathbf{F}}_{0,[i,:]} - \mathbf{F}_{[i,:]})\mathbf{V}\|_2$$
$$\quad + \left(\|\hat{\mathbf{F}}_{0,[i,:]} - \mathbf{F}_{[i,:]}\|_2 + \|\mathbf{F}_{[i,:]}\|_2\right) \cdot 2\|\sin\Theta(\tilde{\mathbf{V}}_{[:,1:r]}, \mathbf{V})\|$$
$$\le \|(\hat{\mathbf{F}}_{0,[i,:]} - \mathbf{F}_{[i,:]})\mathbf{V}\|_2$$
$$\quad + \left(\|\hat{\mathbf{F}}_{0,[i,:]} - \mathbf{F}_{[i,:]}\|_2 + \|\mathbf{F}_{[i,:]}\|_2\right) \frac{C\|\tilde{\mathbf{F}} - \mathbf{F}\|}{\sigma_r(\mathbf{F})}$$
$$\le C\left(\frac{\tilde{r}\kappa^3 \tau_* \log^2(n)}{np^2}\right)^{1/2}$$
$$\quad + \frac{C\left(\|\hat{\mathbf{F}}_{0,[i,:]} - \mathbf{F}_{[i,:]}\|_2 + \|\mathbf{F}_{[i,:]}\|_2\right)}{\sigma_r(\mathbf{F})}$$
$$\quad \times \left(\frac{\pi_{\max}\tau_* \log^2(n)}{n}\right)^{1/2}$$
$$\le C\left(\frac{\tilde{r}\kappa^3 \tau_* \log^2(n)}{np^2}\right)^{1/2}$$
$$\quad + \frac{C\left(\|\hat{\mathbf{F}}_{0,[i,:]} - \mathbf{F}_{[i,:]}\|_2 + \|\mathbf{F}_{[i,:]}\|_2\right)}{\|\mathbf{F}\|_F}$$
$$\quad \times \left(\frac{\pi_{\max}\tilde{r}\tau_* \log^2(n)}{n}\right)^{1/2}. \tag{33}$$

Here, the second line is due to (28); the third line is due to the property of $\sin\Theta$ distance (see Lemma 1 in [44]); the fourth line is due to $\mathbf{V}$ and $\tilde{\mathbf{V}}_{[:,1:r]}$ are the leading singular vectors of $\mathbf{F}$ and $\tilde{\mathbf{F}}$ and Wedin's perturbation theorem; the fifth line is due to (31) and (32); the sixth line is due to $\tilde{r} = \|\mathbf{F}\|_F^2/\sigma_r^2(\mathbf{F})$ by definition. Thus,

$$\|\hat{\mathbf{F}}_{0,[i,:]} - \mathbf{F}_{[i,:]}\|_2$$
$$\le \frac{C\left(\frac{\tilde{r}\kappa^3 \tau_* \log^2(n)}{np^2}\right)^{1/2} + \frac{C\|\mathbf{F}_{[i,:]}\|_2}{\|\mathbf{F}\|_F}\left(\frac{\pi_{\max}\tilde{r}\tau_* \log^2(n)}{n}\right)^{1/2}}{\left(1 - \frac{C}{\|\mathbf{F}\|_F}\left(\frac{\pi_{\max}\tilde{r}\tau_* \log^2(n)}{n}\right)^{1/2}\right)_+}. \tag{34}$$

In addition, by the Cauchy-Schwarz inequality,

$$\|\mathbf{F}\|_F = \left(\sum_{i,j=1}^p \mathbf{F}_{ij}^2\right)^{1/2} \ge \frac{1}{p}\sum_{i,j=1}^p |\mathbf{F}_{ij}| = \frac{1}{p};$$

We also have

$$\pi_{\max} \le \kappa/p,$$
$$\frac{\|\mathbf{F}_{[i,:]}\|_2}{\|\mathbf{F}\|_F} \le \frac{\sqrt{\sum_{j=1}^p \mathbf{F}_{ij}^2}}{1/p} \le \frac{\sqrt{(\kappa/p^2)\sum_{j=1}^p \mathbf{F}_{ij}}}{1/p}$$
$$\le \sqrt{\kappa\pi_{\max}}.$$

Thus, the denominator of (34) satisfies

$$1 - \frac{C}{\|\mathbf{F}\|_F} \left(\frac{\pi_{\max} \tilde{r} \tau_* \log^2(n)}{n} \right)^{1/2}$$

$$\geq 1 - C \left(\frac{\kappa p \tilde{r} \tau_* \log^2(n)}{n} \right)^{1/2} \geq \frac{1}{2}$$

provided (30) holds with a large constant $C > 0$ on the right hand side of (30). Combining these inequalities with (34), one has for any $1 \leq i \leq p$,

$$\mathbb{P}\left(\|\hat{\mathbf{F}}_{0,[i,:]} - \mathbf{F}_{[i,:]}\|_2 \leq C \left(\frac{\tilde{r}\kappa^3 \tau_* \log^2(n)}{np^2} \right)^{1/2} \right)$$

$$\geq 1 - C n^{-c}. \tag{35}$$

Finally, by Lemma 2 and the definition of $\hat{\mathbf{P}}$,

$$\|\hat{\mathbf{P}}_{[i,:]} - \mathbf{P}_{[i,:]}\|_1 = \left\| \frac{(\hat{\mathbf{F}}_{0,[i,:]})_+}{\|(\hat{\mathbf{F}}_{0,[i,:]})_+\|_1} - \frac{\mathbf{F}_{[i,:]}}{\|\mathbf{F}_{[i,:]}\|_1} \right\|_1$$

$$\leq \frac{2\|(\hat{\mathbf{F}}_{0,[i,:]})_+ - \mathbf{F}_{[i,:]}\|_1}{\|\mathbf{F}_{[i,:]}\|_1} \leq \frac{2\|\hat{\mathbf{F}}_{[i,:]} - \mathbf{F}_{[i,:]}\|_1}{\pi_i}$$

$$\leq \frac{C\sqrt{p}\|\hat{\mathbf{F}}_{[i,:]} - \mathbf{F}_{[i,:]}\|_2}{\pi_i}$$

for any $1 \leq i \leq p$. By (35) and the previous inequality, we have the following high-probability upper bound

$$\mathbb{P}\left(\max_{1 \leq i \leq p} \|\hat{\mathbf{P}}_{[i,:]} - \mathbf{P}_{[i,:]}\|_1 \right.$$

$$\left. \leq C \left(\frac{p\tilde{r}}{n} \frac{\kappa^3}{(p\pi_{\min})^2} \tau_* \log^2(n) \right)^{1/2} \right)$$

$$\geq 1 - Cpn^{-c} \geq 1 - Cn^{-c+1},$$

since $n \geq Cp\tilde{r}\frac{\kappa^3}{(p\pi_{\min})^2} \cdot \tau_* \log^2(n)$. We can additionally develop the expectation upper bound similarly as the argument of (27). $\square$

B. Eigen-gap Condition

Eigengap Condition: When $\mathbf{P}$ satisfies the *detailed balance condition*, i.e., $\pi_i \mathbf{P}_{ij} = \pi_j \mathbf{P}_{ji}$ for any $1 \leq i, j \leq p$, or equivalently $\mathbf{F}$ is symmetric, the corresponding Markov process is referred to as being *reversible*. The reversibility is an important and widely considered condition in stochastic process literature. When the Markov process is reversible, it is well-known that all eigenvalues of $\mathbf{P}$ must be real and between -1 and 1; the largest eigenvalue of a reversible Markov transition matrix is always 1 [52, Chapter 12]. Suppose the second largest eigenvalue of $\mathbf{P}$ is $\lambda_2 < 1$, then $1 - \lambda_2$ plays an important role in regulating the connectivity of the Markov chain: the more close λ_2 is to 1, the more likely the Markov chain is congested. Moreover, the eigengap of reversible Markov processes can be estimated from the observable states via a plug-in estimator [65].

The following results hold as an extension of Theorem 1 based on eigengap assumption.

Corollary 1. *Under the assumption of Theorem 1, if* $\mathbf{P}$ *is reversible and with second largest eigenvalue* $\lambda_2 < 1$,

then (9), (1), *and* (11) *hold if one replace* $\tau_* \log^2(n)$ *by* $\log(n/\pi_{\min}) \log(n)/(1 - \lambda_2)$.

Proof of Corollary 1. If the Markov process is reversible and $1 - \lambda_2$ is the eigengap (see Section IX-B), by Lemma 7, one has

$$\mathbb{P}\left(\max\{\|\tilde{\mathbf{F}} - \mathbf{F}\|, \|\tilde{\pi} - \pi\|_\infty\} \right.$$

$$\left. \geq C\sqrt{\frac{\pi_{\max} \log(n/\pi_{\min}) \log(n)}{n(1 - \lambda_2)}} \right) \leq n^{-c_0}.$$

By replacing $\tau_* \log^2(n)$ by $\log(n/\pi_{\min})/(1 - \lambda_2)$, the proof for Corollary 1 immediately follows from the arguments in Theorem 1. $\square$

C. Proof of Theorem 2

First, we study the Kullback-Leibler divergence between two Markov processes with same the same state space $\{1, \dots, p\}$ but different transition matrices $\mathbf{P}$ and $\mathbf{Q}$. Suppose π is the invariant distribution of both $\mathbf{P}$ and $\mathbf{Q}$, $X^{(1)} = \{x_0^{(1)}, \dots, x_n^{(1)}\}$ and $X^{(2)} = \{x_0^{(2)}, \dots, x_n^{(2)}\}$ are two Markov chains generated from $\mathbf{P}$ and $\mathbf{Q}$, and $x_0^{(1)} \sim \pi$, i.e. the starting point of $X^{(1)}$ is from its invariant distribution. Then, clearly $x_0^{(1)}, \dots, x_n^{(1)}$ identically satisfy the distribution of π (though they are dependent). Recall the KL divergence between two discrete random distributions p and q is defined as $D_{KL}(p\|q) = \sum_x p(x) \log(p(x)/q(x))$. Thus,

$$D_{KL}\left(X^{(1)}\|X^{(2)}\right) := \sum_{X \in [p]^{n+1}} p_{X^{(1)}}(X) \log\left(\frac{p_{X^{(1)}}(X)}{p_{X^{(2)}}(X)} \right)$$

$$= \sum_{\substack{i_0,\dots,i_n \\ \in [p]^{n+1}}} \mathbb{P}\left(X^{(1)} = (i_0, \dots, i_n)\right) \log\left(\frac{\mathbb{P}\left(X^{(1)} = (i_0, \dots, i_n)\right)}{\mathbb{P}\left(X^{(2)} = (i_0, \dots, i_n)\right)} \right)$$

$$= \sum_{i_0,\dots,i_n \in [p]^{n+1}} \pi_{i_0} \mathbf{P}_{i_0,i_1} \cdots \mathbf{P}_{i_{n-1},i_n} \log\left(\frac{\pi_{i_0} \mathbf{P}_{i_0,i_1} \cdots \mathbf{P}_{i_{n-1},i_n}}{\pi_{i_0} \mathbf{Q}_{i_0,i_1} \cdots \mathbf{Q}_{i_{n-1},i_n}} \right)$$

$$= \sum_{i_0,\dots,i_{n-1} \in [p]^n} \sum_{i_n \in [p]} \pi_{i_0} \mathbf{P}_{i_0,i_1} \cdots \mathbf{P}_{i_{n-1},i_n}$$

$$\left\{ \log\left(\frac{\pi_{i_0} \mathbf{P}_{i_0,i_1} \cdots \mathbf{P}_{i_{n-2},i_{n-1}}}{\pi_{i_0} \mathbf{Q}_{i_0,i_1} \cdots \mathbf{Q}_{i_{n-2},i_{n-1}}} \right) + \log\left(\frac{\mathbf{P}_{i_{n-1},i_n}}{\mathbf{Q}_{i_{n-1},i_n}} \right) \right\}$$

$$= D_{KL}\left(\{x_0^{(1)}, \dots, x_{n-1}^{(1)}\} \| \{x_0^{(2)}, \dots, x_{n-1}^{(2)}\} \right)$$

$$+ \sum_{i_{n-1} \in [p]} \pi_{i_{n-1}} \sum_{i_n \in [p]} \mathbf{P}_{i_{n-1},i_n} \log\left(\frac{\mathbf{P}_{i_{n-1},i_n}}{\mathbf{Q}_{i_{n-1},i_n}} \right)$$

$$= D_{KL}\left(\{x_0^{(1)}, \dots, x_{n-1}^{(1)}\} \| \{x_0^{(2)}, \dots, x_{n-1}^{(2)}\} \right)$$

$$+ \sum_{i \in [p]} \pi_i D_{KL}\left(\mathbf{P}_{[i,:]} \| \mathbf{Q}_{[i,:]} \right).$$

Then it is easy to use induction to show that

$$D_{KL}\left(X^{(1)}\|X^{(2)}\right)$$

$$= D_{KL}\left(\{x_0^{(1)}, \dots, x_{n-1}^{(1)}\} \| \{x_0^{(2)}, \dots, x_{n-1}^{(2)}\} \right)$$

$$+ \sum_{i \in [p]} \pi_i D_{KL}\left(\mathbf{P}_{[i,:]} \| \mathbf{Q}_{[i,:]} \right)$$

$$= \cdots = D_{KL}\left(x_0^{(1)}\middle\|x_0^{(2)}\right) + n\sum_{i\in[p]} \pi_i D_{KL}\left(\mathbf{P}_{[i,:]}\|\mathbf{Q}_{[i,:]}\right).$$

$$\tag{36}$$

Next, we prove the lower bound for estimating $\mathbf{F}$. Let $p_0 = \lfloor p/2 \rfloor$, $l_0 = \lfloor p_0/\{2(r-1)\}\rfloor$. We construct a sequence of instances of rank-r Markov chains, with transition matrices $\mathbf{P}^{(1)}, \ldots, \mathbf{P}^{(m)}$ (m to be specified later)

Here $\{\mathbf{R}^{(k)}\}_{k=1}^m$ are i.i.d. Bernoulli p_0-by-$(r-1)$ random matrices, $\mathbf{0}_{a\times b}$ is the a-by-b zero matrix, and $0 < \eta \leq 1/2$ is some constant to be determined later. Then clearly, $\mathbf{P}^{(k)}$ is a transition matrix, and $\frac{1}{p}\mathbf{1}_p$ is the invariant distribution, then the corresponding frequency matrix is $\mathbf{F}^{(k)} = \frac{1}{p}\mathbf{P}^{(k)}$ and $\pi_{\max} = \pi_{\min} = 1/p$. Since $\mathrm{rank}(\mathbf{R}^{(k)}) \leq r - 1$, we also have $\mathrm{rank}(\mathbf{P}^{(k)}) \leq r$. Additionally, it is easy to see that each entry of $\mathbf{P}^{(k)}$ is between $(1/p - \eta/2p)$ and $(1/p + \eta/(2p))$. Thus for any $1 \leq i \leq p$,

$$\left\|e_i^\top \mathbf{P}^{(k)} - \pi\right\|_1 = \sum_{j=1}^p |\mathbf{P}_{ij}^{(k)} - \pi_j| \leq p \cdot \eta/(2p) \leq 1/4,$$

which means $\tau_* := \tau(1/4) \leq 1$. By definitions of $\mathcal{P}_{p,r}$ and $\mathcal{P}_{p,r}^*$, we have for any k that

$$\mathbf{P}^{(k)} \in \mathcal{P}_{p,r}^* \subseteq \mathcal{P}_{p,r}. \tag{38}$$

Now for any $k \neq l$,

$$\|\mathbf{F}^{(k)} - \mathbf{F}^{(l)}\|_1 = \frac{1}{p}\|\mathbf{P}^{(k)} - \mathbf{P}^{(l)}\|_1 = \frac{2l_0\eta}{p^2}\|\mathbf{R}^{(k)} - \mathbf{R}^{(l)}\|_1$$

$$= \frac{2l_0\eta}{p^2}\sum_{i=1}^{p_0}\sum_{j=1}^{r-1}\left|\mathbf{R}_{ij}^{(k)} - \mathbf{R}_{ij}^{(l)}\right|.$$

It is easy to see that $\left\{\left|\mathbf{R}_{ij}^{(k)} - \mathbf{R}_{ij}^{(l)}\right|\right\}$ are i.i.d. uniformly distributed on $\{0, 2\}$. These random variables also satisfy

$$\mathbb{E}\left|\mathbf{R}_{ij}^{(k)} - \mathbf{R}_{ij}^{(l)}\right| = 1, \quad \mathrm{Var}\left(\left|\mathbf{R}_{ij}^{(k)} - \mathbf{R}_{ij}^{(l)}\right|\right) = 1,$$
$$\left|\left|\mathbf{R}_{ij}^{(k)} - \mathbf{R}_{ij}^{(l)}\right| - 1\right| = 1.$$

By Bernstein's inequality, for any $\varepsilon > 0$ we have

$$\mathbb{P}\left(\left|\left\|\mathbf{F}^{(k)} - \mathbf{F}^{(l)}\right\|_1 - \frac{2l_0\eta p_0(r-1)}{p^2}\right| \geq \frac{2l_0\eta}{p^2}\varepsilon\right)$$
$$\leq 2\exp\left(\frac{-\varepsilon^2/2}{p_0(r-1) + \varepsilon/3}\right).$$

Set $\varepsilon = p_0(r-1)/2$, $m = \sqrt{\lfloor\exp(p_0(r-1)/28)\rfloor}$, then we further have

$$\mathbb{P}\Bigg(\forall 1 \leq k < l \leq m,$$
$$\frac{l_0\eta p_0(r-1)}{p^2} \leq \left\|\mathbf{F}^{(k)} - \mathbf{F}^{(l)}\right\|_1 \leq \frac{3l_0\eta p_0(r-1)}{p^2}\Bigg)$$
$$\geq 1 - m(m-1)\exp\left(\frac{-p_0(r-1)}{28}\right)$$
$$> 1 - m^2\exp\left(\frac{-p_0(r-1)}{28}\right) > 0.$$

By such an argument, we can see there exists $\{\mathbf{R}^{(1)}, \ldots, \mathbf{R}^{(m)}\} \subseteq \{-1, 1\}^{p_0 \times (r-1)}$ such that

$$\forall 1 \leq k < l \leq m,$$
$$\frac{l_0\eta p_0(r-1)}{p^2} \leq \left\|\mathbf{F}^{(k)} - \mathbf{F}^{(l)}\right\|_1 \leq \frac{3l_0\eta p_0(r-1)}{p^2}. \tag{39}$$

We thus assume (39) is satisfied.

Next, we construct m Markov chains of length $(n+1)$: $\{X^{(1)}, \ldots, X^{(m)}\}$. For each $k \in \{1, \ldots, m\}$, $x_0^{(k)} \sim \frac{1_p}{p}$, and the rest of the states are generated according to $\mathbf{P}^{(k)}$ and $\mathbf{F}^{(k)}$. Based on the calculation in (36),

$$D_{KL}\left(X^{(k)}\middle\|X^{(l)}\right) = \frac{n}{p}\sum_{i=1}^p D_{KL}\left(\mathbf{P}_{[i,:]}^{(k)}\middle\|\mathbf{P}_{[i,:]}^{(l)}\right)$$

Based on Lemma 4 and $1/(2p) \leq \mathbf{P}_{ij}^{(k)} \leq 3/(2p)$, we further have $D_{KL}\left(\mathbf{P}_{[i,:]}^{(k)}\middle\|\mathbf{P}_{[i,:]}^{(l)}\right) \leq 3p\|\mathbf{P}_{[i,:]}^{(k)} - \mathbf{P}_{[i,:]}^{(l)}\|_2^2$. Thus, for any $1 \leq k < l \leq m$,

$$D_{KL}\left(X^{(k)}\middle\|X^{(l)}\right) \leq 3n\sum_{i=1}^p \|\mathbf{P}_{[i,:]}^{(k)} - \mathbf{P}_{[i,:]}^{(l)}\|_2^2$$

$$= 3n\sum_{i,j=1}^p \left(\mathbf{P}_{ij}^{(k)} - \mathbf{P}_{ij}^{(l)}\right)^2 \leq \frac{6n\eta}{p}\sum_{i,j=1}^p \left|\mathbf{P}_{ij}^{(k)} - \mathbf{P}_{ij}^{(l)}\right|$$

$$\leq 6n\eta \cdot \|\mathbf{F}^{(k)} - \mathbf{F}^{(l)}\|_1 \leq \frac{18n\eta^2\, l_0 p_0(r-1)}{p^2}.$$

Now, by the generalized Fano's lemma (see, e.g., [69], [70]), we have

$$\inf_{\hat{\mathbf{F}}} \sup_{\mathbf{F}\in\{\mathbf{F}^{(1)},\ldots,\mathbf{F}^{(m)}\}} \mathbb{E}\left\|\hat{\mathbf{F}} - \mathbf{F}\right\|_1$$
$$\geq \frac{l_0\eta p_0(r-1)}{p^2}\left(1 - \frac{18n\eta^2 l_0 p_0(r-1)/p^2 + \log 2}{\log m}\right).$$

$$\mathbf{P}^{(k)} = \frac{1}{p}\mathbf{1}_p\mathbf{1}_p^\top + \frac{\eta}{2p}\begin{bmatrix} \overbrace{\begin{matrix}\mathbf{R}^{(k)} & \cdots & \mathbf{R}^{(k)}\end{matrix}}^{l_0} & \overbrace{\begin{matrix}-\mathbf{R}^{(k)} & \cdots & -\mathbf{R}^{(k)}\end{matrix}}^{l_0} & \mathbf{0}_{p_0\times(p-2l_0(r-1))} \\ -\mathbf{R}^{(k)} & \cdots & -\mathbf{R}^{(k)} & \mathbf{R}^{(k)} & \cdots & \mathbf{R}^{(k)} & \mathbf{0}_{p_0\times(p-2l_0(r-1))} \\ \mathbf{0}_{(p-2p_0)\times(l_0(r-1))} & & & \mathbf{0}_{(p-2p_0)\times(l_0(r-1))} & & & \mathbf{0}_{(p-2p_0)\times(p-2l_0(r-1))} \end{bmatrix} \tag{37}$$

Finally, we set $\eta^2 = \left\{ \frac{p^2}{18nl_0p_0(r-1)} \left(\frac{1}{2} \log(m) - \log(2) \right) \right\} \wedge 1/2$ and apply (38),

$$\inf_{\substack{\hat{\mathbf{F}} \\ \mathbf{F}=\mathrm{diag}(\pi)\mathbf{P}}} \sup_{\mathbf{P}\in\mathcal{P}_{p,r}} \mathbb{E}\left\|\hat{\mathbf{F}} - \mathbf{F}\right\|_1 \geq \inf_{\substack{\hat{\mathbf{F}} \\ \mathbf{F}=\mathrm{diag}(\pi)\mathbf{P}}} \sup_{\mathbf{P}\in\mathcal{P}_{p,r}} \mathbb{E}\left\|\hat{\mathbf{F}} - \mathbf{F}\right\|_1$$

$$\geq \inf_{\hat{\mathbf{F}}} \sup_{\mathbf{F}\in\{\mathbf{F}^{(1)},\dots,\mathbf{F}^{(m)}\}} \mathbb{E}\left\|\hat{\mathbf{F}} - \mathbf{F}\right\|_1$$

$$\geq \frac{p_0l_0(r-1)}{2p^2} \cdot \sqrt{\frac{p^2 \cdot \left(\frac{1}{2}\log(m) - \log(2)\right)}{18np_0l_0(r-1)}} \geq c\sqrt{\frac{pr}{n}} \wedge 1.$$

Finally, since $\mathbf{P}^{(k)} = p\mathbf{F}^{(k)}$ based on the set-up,

$$\inf_{\hat{\mathbf{P}}} \sup_{\mathbf{P}\in\mathcal{P}_{p,r}} \mathbb{E}\frac{1}{p}\left\|\hat{\mathbf{P}} - \mathbf{P}\right\|_1 \geq \inf_{\hat{\mathbf{P}}} \sup_{\mathbf{P}\in\mathcal{P}^*_{p,r}} \frac{1}{p}\mathbb{E}\left\|\hat{\mathbf{P}} - \mathbf{P}\right\|_1$$

$$\geq \inf_{\hat{\mathbf{P}}} \sup_{\mathbf{P}\in\{\mathbf{P}^{(1)},\dots,\mathbf{P}^{(m)}\}} \frac{1}{p}\mathbb{E}\left\|\hat{\mathbf{P}} - \mathbf{P}\right\|_1$$

$$= p \inf_{\hat{\mathbf{F}}} \sup_{\mathbf{F}\in\{\mathbf{F}^{(1)},\dots,\mathbf{F}^{(m)}\}} \frac{1}{p}\mathbb{E}\left\|\hat{\mathbf{F}} - \mathbf{F}\right\|_1 \geq c\sqrt{\frac{pr}{n}} \wedge 1.$$

$\square$

X. PROOFS FOR RESULTS OF SECTION V

A. Proof of Theorem 3

Let $\mathbf{G} = \mathrm{diag}(\pi)\mathbf{Q}$ be the frequency matrix of transition x to y. Suppose $\mathbf{G} = \mathbf{U}_G\boldsymbol{\Sigma}_G\mathbf{V}_G^\top$ is the SVD, where $\mathbf{U}_G \in \mathbb{O}_{p,r}, \mathbf{V}_G \in \mathbb{O}_{q,r}$. Define $g_{\max} = \max_j \sum_{i=1}^p \mathbf{G}_{ij}$. Recall $\kappa/pq = \max_{ij} \mathbf{G}_{ij}$. Then $g_{\max} \leq \sum_{j=1}^q \kappa/(pq) = \kappa/p$, $\pi_{\max} \leq \sum_{i=1}^p \kappa/(pq) \leq \kappa/q$. Similar to Lemma 7, one can show that

$$\|\tilde{\mathbf{G}} - \mathbf{G}\| \leq C\left(\frac{(\pi_{\max} \vee g_{\max})\tau_* \log^2(n)}{n}\right)^{1/2}$$

$$\leq C\left(\frac{(p \vee q)\kappa\tau_* \log^2(n)}{npq}\right)^{1/2} \quad (40)$$

with probability at least $1 - Cn^{-c}$. Based on the above concentration inequality, the rest of the proof of (15) is similar to the average upper bound result in Theorem 1.

Note that the trivial bound

$$\mathbb{E}\max_i \|\hat{\mathbf{Q}}_{[i,:]} - \mathbf{Q}_{[i,:]}\|_1 \leq 2$$

always holds. In order to prove (16), we only need to show under the assumption that $n \geq C_0\tau_* \log^2(n)(p \vee q)\left(\tilde{r} \cdot \frac{\kappa^3}{p^2\pi_{\min}^2}\right)$. Similarly as the proof of Lemma 8 and Theorem 1, we have

$$\|(\tilde{\mathbf{G}}_{[i,:]} - \mathbf{G}_{[i,:]})\mathbf{V}_G\|_2 \leq C\left(\frac{\tilde{r}\kappa^3\tau_* \log^2(n)}{npq}\right)^{1/2}. \quad (41)$$

for any $1 \leq i \leq p$ with probability at least $1 - Cn^{-c}$. Then the rest of the proof is essentially the same as the one in the uniform upper bound of $\hat{\mathbf{P}}$ in Theorem 1. $\square$

B. Proof of Theorem 4

By Lemma 7, one has

$$\mathbb{P}\left(\|\tilde{\mathbf{F}} - \mathbf{F}\| \geq C\sqrt{\pi_{\max}\tau_* \log^2(n)/n}\right) \leq n^{-c_0}.$$

Wedin's lemma [71] implies

$$\mathbb{P}\left(\max\left\{\|\sin\Theta(\hat{\mathbf{U}}_F, \mathbf{U}_F)\|, \|\sin\Theta(\hat{\mathbf{V}}_F, \mathbf{V}_F)\|\right\}\right.$$

$$\left. \leq \frac{C\sqrt{\pi_{\max}\tau_* \log^2(n)/n}}{\sigma_r(\mathbf{F}) - \sigma_{r+1}(\mathbf{F})}\right) \geq 1 - n^{-c_0}.$$

Let Q be the event that the above inequality holds. Since the trivial bound

$$\max\{\|\sin\Theta(\hat{\mathbf{U}}_F, \mathbf{U}_F)\|, \|\sin\Theta(\hat{\mathbf{V}}_F, \mathbf{V}_F)\|\} \leq 1$$

holds, we must have

$$\mathbb{E}\max\left\{\|\sin\Theta(\hat{\mathbf{U}}_F, \mathbf{U}_F)\|, \|\sin\Theta(\hat{\mathbf{V}}_F, \mathbf{V}_F)\|\right\}$$

$$\leq \mathbb{E}\max\left\{\|\sin\Theta(\hat{\mathbf{U}}_F, \mathbf{U}_F)\|, \|\sin\Theta(\hat{\mathbf{V}}_F, \mathbf{V}_F)\|\right\} 1_Q$$

$$+ \mathbb{E}\max\left\{\|\sin\Theta(\hat{\mathbf{U}}_F, \mathbf{U}_F)\|, \|\sin\Theta(\hat{\mathbf{V}}_F, \mathbf{V}_F)\|\right\} 1_{Q^c}$$

$$\leq \frac{C\sqrt{\pi_{\max}\tau_* \log^2(n)/n}}{\sigma_r(\mathbf{F}) - \sigma_{r+1}(\mathbf{F})} + 1 \cdot \mathbb{P}(Q^c)$$

$$\leq \frac{C\sqrt{\pi_{\max}\tau_* \log^2(n)/n}}{\sigma_r(\mathbf{F}) - \sigma_{r+1}(\mathbf{F})} + \frac{1}{n^{c_0}}.$$

Since $\sum_{i,j} \mathbf{F}_{ij} = 1$ and $0 \leq \mathbf{F}_{ij} \leq 1$, we must have

$$0 \leq \sigma_r(\mathbf{F}) - \sigma_{r+1}(\mathbf{F}) \leq \left(\sum_i \sigma_i^2(\mathbf{F})\right)^{1/2}$$

$$= \|\mathbf{F}\|_F = \left(\sum_{ij} \mathbf{F}_{ij}^2\right)^{1/2} \leq \|\mathbf{F}\|_1^{1/2} \leq 1, \quad \pi_{\max} \geq 1/p.$$

Thus, if $c_0 > 1$, one has $1/n^{c_0} \leq \frac{C\sqrt{\pi_{\max}\tau_* \log^2(n)/n}}{\sigma_r(\mathbf{F})-\sigma_{r+1}(\mathbf{F})}$ and

$$\mathbb{E}\max\left\{\|\sin\Theta(\hat{\mathbf{U}}_F, \mathbf{U}_F)\|, \|\sin\Theta(\hat{\mathbf{V}}_F, \mathbf{V}_F)\|\right\}$$

$$\leq \frac{C\sqrt{\pi_{\max}\tau_* \log^2(n)/n}}{\sigma_r(\mathbf{F}) - \sigma_{r+1}(\mathbf{F})} \wedge 1,$$

which implies (20).

Next we consider $\hat{\mathbf{U}}_P$, and $\hat{\mathbf{V}}_P$. Note that $\|\mathbf{P}\|/(\sigma_r(\mathbf{P}) - \sigma_{r+1}(\mathbf{P})) \geq 1$. If $n \leq Cp\pi_{\max}/(\pi_{\min}^2 p)\tau_* \log^2(n)$, the trivial bound $\mathbb{E}\left(\|\sin\Theta(\hat{\mathbf{U}}_P, \mathbf{U}_P)\| \vee \|\sin\Theta(\hat{\mathbf{V}}_P, \mathbf{V})\|\right) \leq 1$ has already provided sharp enough result for proving (20). Thus for the rest of proof, we assume $n \geq Cp\pi_{\max}/(\pi_{\min}^2 p)\tau_* \log^2(n)$ for large enough constant C. Let $\tilde{\pi}$ be the empirical distribution of π,

$$\tilde{\pi} \in \mathbb{R}^p, \quad \tilde{\pi}_i = \frac{1}{n}\sum_{k=1}^n 1_{\{X_{k-1}=i\}}.$$

Provided that $n \geq C\frac{\pi_{\max}\tau_* \log^2(n)}{\pi_{\min}^2}$ for large enough constant $C > 0$, we have

$$\|\tilde{\pi} - \pi\|_\infty \leq C\sqrt{\frac{\pi_{\max}\tau_* \log^2(n)}{n}} \leq \frac{1}{2}\pi_{\min}.$$

Then

$$\min_i \tilde\pi_i \geq \min_i \pi_i - \|\tilde\pi - \pi\|_\infty \geq \frac{1}{2}\pi_{\min}, \qquad (42)$$

and

$$|\pi_i/\tilde\pi_i - 1| = \frac{|\pi_i - \tilde\pi_i|}{\tilde\pi_i} \leq 2\pi_{\min}^{-1} \cdot C\sqrt{\frac{\pi_{\max}\tau_* \log^2(n)}{n}}. \quad (43)$$

Since $\tilde{\mathbf{P}} = \tilde\pi^{-1}\tilde{\mathbf{F}}$, we have

$$\left\|\tilde{\mathbf{P}} - \mathbf{P}\right\| = \left\|\mathrm{diag}(\tilde\pi)^{-1}\tilde{\mathbf{F}} - \mathrm{diag}(\pi)^{-1}\mathbf{F}\right\|$$

$$\leq \left\|\tilde\pi^{-1}(\tilde{\mathbf{F}} - \mathbf{F})\right\| + \left\|\left(\mathrm{diag}(\pi)^{-1} - \mathrm{diag}(\tilde\pi)^{-1}\right)\mathbf{F}\right\|$$

$$\leq \left\|\tilde\pi^{-1}\right\| \cdot \|\tilde{\mathbf{F}} - \mathbf{F}\| + \|\mathbf{I} - \mathrm{diag}(\pi/\tilde\pi)\| \cdot \|\mathrm{diag}(\pi)^{-1}\mathbf{F}\|$$

$$\leq \left(\min_i \tilde\pi_i\right)^{-1} \cdot \|\tilde{\mathbf{F}} - \mathbf{F}\| + \max_i |\pi_i/\tilde\pi_i - 1| \cdot \|\mathbf{P}\|$$

$$\overset{(42)(43)}{\leq} C\pi_{\min}^{-1}\sqrt{\frac{\pi_{\max}\tau_* \log^2(n)}{n}}$$

$$+ C\pi_{\min}^{-1}\sqrt{\frac{\pi_{\max}\tau_* \log^2(n)}{n}}\|\mathbf{P}\|.$$

Since $\|\mathbf{P}\| \geq \|\frac{1}{\sqrt{p}}1_p^\top\mathbf{P}\|_2 = 1$, the inequality above further yields

$$\left\|\tilde{\mathbf{P}} - \mathbf{P}\right\| \leq C\pi_{\min}^{-1}\sqrt{\frac{\pi_{\max}\tau_* \log^2(n)}{n}}\|\mathbf{P}\|.$$

Finally, by Wedin's perturbation bound, we have

$$\max\left\{\|\sin\Theta(\hat{\mathbf{U}}_P, \mathbf{U}_P)\|, \|\sin\Theta(\hat{\mathbf{V}}_P, \mathbf{V}_P)\|\right\}$$

$$\leq \frac{C\|\mathbf{P}\| \cdot \sqrt{(p/n) \cdot \pi_{\max}/(p\pi_{\min}^2) \cdot \tau_* \log^2(n)}}{\sigma_r(\mathbf{P}) - \sigma_{r+1}(\mathbf{P})}$$

with probability at least $1 - n^{-c_0}$. By similar argument as the one in Theorem 1, one can finally show (20). $\qquad\square$

C. Proof of Theorem 5

We focus on the proof for $\mathbf{U}_P$ and $\mathbf{U}_F$ and $r = 2$, as the proof for $\mathbf{V}_P$ and $\mathbf{V}_F$ or $r \geq 3$ essentially follows. Without loss of generality we also assume p is a multiple of 4. First, we construct a series of rank-2 Markov chain transition matrices, which are all in $\mathcal{P}^*_{p,r,\delta_P}$. To be specific, let

$$\mathbf{P}^{(k)} = \frac{1}{p}1_p 1_p^\top + \frac{\sqrt{2}\delta_P}{p} \cdot$$

$$\left[\begin{array}{cccc|cccc}
\overbrace{1_{p/4} \quad \cdots \quad 1_{p/4}}^{p/2} & \overbrace{-1_{p/4} \quad \cdots \quad -1_{p/4}}^{p/2} \\
-1_{p/4} \quad \cdots \quad -1_{p/4} & 1_{p/4} \quad \cdots \quad 1_{p/4} \\
\zeta\beta^{(k)} \quad \cdots \quad \zeta\beta^{(k)} & -\zeta\beta^{(k)} \quad \cdots \quad -\zeta\beta^{(k)} \\
-\zeta\beta^{(k)} \quad \cdots \quad -\zeta\beta^{(k)} & \zeta\beta^{(k)} \quad \cdots \quad \zeta\beta^{(k)}
\end{array}\right].$$

$$(44)$$

Here $\{\beta^{(k)}\}_{k=1}^m$ are m copies of i.i.d. Rademacher $(p/4)$-dimensional random vectors, $0 < \zeta \leq 1$ and m are fixed values to be determined later. It is not hard to check that the

invariant distribution $\pi = \frac{1}{p}1_p$ and the SVD of $\mathbf{P}^{(k)}$ can be written as

$$\mathbf{P}^{(k)} = \left(\frac{1}{\sqrt{p}}1_p\right)\left(\frac{1}{\sqrt{p}}1_p\right)^\top + \sigma^{(k)}u^{(k)}(v^{(k)})^\top, \qquad (45)$$

where

$$\sigma^{(k)} = \frac{\sqrt{2}\zeta}{p}\sqrt{\frac{p^2}{2}(1+\zeta^2)} \geq \zeta,$$

$$u^{(k)} = \frac{1}{\sqrt{\frac{p}{2}(1+\zeta^2)}}\begin{bmatrix}1_{p/4}\\-1_{p/4}\\\zeta\beta^{(k)}\\-\zeta\beta^{(k)}\end{bmatrix}, \quad v^{(k)} = \frac{1}{\sqrt{p}}\begin{bmatrix}1_{p/2}\\-1_{p/2}\end{bmatrix}.$$

Thus, $\|\mathbf{P}^{(k)}\| = 1$ and $(\sigma_2(\mathbf{P}^{(k)}) - \sigma_3(\mathbf{P}^{(k)}))/\|\mathbf{P}\| \geq \delta_P$. Namely, $\mathbf{P}^{(k)} \in \mathcal{P}^*_{p,r,\delta_P}$, $k = 1, \ldots, m$. Since $\delta_P \leq 1/(4\sqrt{2})$, $3/4 \leq p\mathbf{P}^{(k)}_{ij} \leq 5/4$. Thus,

$$\forall 1 \leq i \leq p, \quad \|e_i^\top\mathbf{P}^{(k)} - \pi\|_1 \leq 1/4,$$

which implies $\tau_* := \tau(1/4) = 1$. In summary, $\mathbf{P}^{(k)} \in \mathcal{P}^*_{p,r,\delta_P}$.

Note that $(\beta^{(k)})^\top\beta^{(l)}$ is a sum of $(p/4)$ i.i.d. Rademacher random variables, by Bernstein's inequality

$$\mathbb{P}\left(\frac{1}{p/4}\left|(\beta^{(k)})^\top\beta^{(l)}\right| \geq 1/2\right) \leq 2\exp\left(-\frac{p/4 \cdot (1/2)^2}{2(1 + 1/3 \cdot 1/2)}\right),$$

then

$$\mathbb{P}\left(\exists k \neq l, \quad \text{s.t.} \quad \frac{1}{p/4}\left|(\beta^{(k)})^\top\beta^{(l)}\right| \geq \frac{1}{2}\right)$$

$$\leq 2 \cdot \frac{m(m-1)}{2}\exp(-p/28) < m^2\exp(-p/28). \qquad (46)$$

If we set $m = \lceil\exp(-p/56)\rceil$, the probability in the right hand side of (46) is strictly less than 1, which means there must exists fixed $\{\beta^{(k)}\}_{k=1}^m$ such that

$$|(\beta^{(k)})^\top\beta^{(l)}| < p/8, \quad \forall 1 \leq k < l \leq m. \qquad (47)$$

For the rest of the proof we assume (47) always hold. Now, for any $k \neq l$,

$$\left\|\sin\Theta\left(\mathbf{U}_P^{(k)}, \mathbf{U}_P^{(l)}\right)\right\| = \|\sin\Theta(u^{(k)}, u^{(l)})\|$$

$$= \sqrt{1 - ((u^{(k)})^\top v^{(l)})^2} = \sqrt{1 - \left(\frac{p/2 + 2\zeta^2(\beta^{(k)})^\top\beta^{(l)}}{p/2 + \zeta^2 p/2}\right)^2}$$

$$\geq \sqrt{1 - \left(\frac{p/2 + \zeta^2 p/4}{p/2 + \zeta^2 p/2}\right)^2} = \sqrt{1 - \left(\frac{1 + \zeta^2/2}{1 + \zeta^2}\right)^2}$$

$$= \sqrt{\frac{\zeta^2/2}{1 + \zeta^2} \cdot \left(1 + \frac{1 + \zeta^2/2}{1 + \zeta^2}\right)} \geq \sqrt{\frac{\zeta^2/2}{2}} = \frac{\zeta}{2}.$$

Now for each $1 \leq k \leq m$, suppose $X^{(k)} = \{x_0^{(k)}, \ldots, x_n^{(k)}\}$ is a Markov chain generated from transition matrix $\mathbf{P}^{(k)}$ and initial distribution $x_0^{(k)} \sim \frac{1}{p}1_p$. Then based on the calculation in

Theorem 2, the KL-divergence between $X^{(k)}$ and $X^{(l)}$ satisfies

$$
\begin{aligned}
D_{KL}\left(X^{(k)}\middle\|X^{(l)}\right) &= \frac{n}{p}\sum_{i=1}^{p} D_{KL}\left(\mathbf{P}_{[i,:]}^{(k)}\middle\|\mathbf{P}_{[i,:]}^{(l)}\right) \\
&\overset{\text{Lemma 4}}{\leq} \frac{20n}{9}\sum_{i=1}^{p}\|\mathbf{P}_{[i,:]}^{(k)} - \mathbf{P}_{[i,:]}^{(l)}\|_2^2 \\
&\overset{(44)}{\leq} \frac{20n}{9}\cdot\frac{2\delta_P^2}{p^2}\cdot\left(2\zeta^2 p^2\right) \leq \frac{80n\delta_P^2\zeta^2}{9}.
\end{aligned}
$$

Finally we set $\zeta = \sqrt{\frac{2\log(m)-\log 2}{80n\delta_P^2/9}}$. By generalized Fano's lemma,

$$
\begin{aligned}
\inf_{\tilde{\mathbf{U}}_P}\sup_{\mathbf{P}\in\{\mathbf{P}^{(1)},\ldots,\mathbf{P}^{(m)}\}} &\mathbb{E}\left\|\sin\Theta(\tilde{\mathbf{U}}_P, \mathbf{U}_P)\right\| \\
\geq \frac{\zeta}{2}\left(1 - \frac{80n\delta_P^2\zeta^2/9 + \log 2}{\log m}\right) &\geq \frac{\zeta}{4} \geq c\frac{\sqrt{p/n}}{\delta_P}
\end{aligned}
$$

for large p. We can finally finish the proof for the theorem by noting that $\{\mathbf{P}^{(1)}, \ldots, \mathbf{P}^{(m)}\} \subseteq \mathcal{P}_{p,r,\delta_P}^* \subseteq \mathcal{P}_{p,r,\delta_P}$. Note that the frequency matrix corresponding to $\mathbf{P}^{(k)}$ is $\mathbf{F}^{(k)} = \text{diag}(\pi)\mathbf{P}^{(k)} = \mathbf{P}^{(k)}/p$ for $k = 1, \ldots, m$, the proof for the lower bound of $\mathbf{U}_F$ exactly follows from the previous arguments. $\square$

D. Proof of Theorem 6

Let $\mathbf{P} = \mathbf{U}_P\mathbf{\Sigma}_P\mathbf{V}_P^\top$ be the singular value decomposition of $\mathbf{P}$, where $\mathbf{U}_P, \mathbf{V}_P \in \mathbb{O}_{p,r}$ and $\mathbf{\Sigma}_P$ has non-negative diagonal entries in descending order. Let $\mathbf{Z} \in \mathbb{R}^{p\times r}$ be the group membership indicator

$$
\mathbf{Z}_{ij} = \begin{cases} 1, & i \in \Omega_j; \\ 0, & i \notin \Omega_j, \end{cases}
$$

By Proposition 3, each column of $\mathbf{P}$ is piece-wise constant with respect to partitions $\Omega_1, \ldots, \Omega_r$ and $\mathbf{P}$ can be written as $\mathbf{P} = \mathbf{ZG}$. Since $\mathbf{U}_P$ and $\mathbf{P}$ share the same column space, we can write

$$
\mathbf{U}_P = \mathbf{ZX},
$$

where $\mathbf{X} \in \mathbb{R}^{r\times r}$ satisfies $\mathbf{X}_{kj} = (\mathbf{U}_P)_{ij}, \forall i \in [p], j \in [r], i \in \Omega_k$. Denote $n_k = |\Omega_k|, k = 1, \ldots, r$. Since the columns of $\mathbf{U}_P$ are orthonormal, we have $\mathbf{X}^\top\mathbf{Z}^\top\mathbf{ZX} = \mathbf{U}_P^\top\mathbf{U}_P = \mathbf{I}_r$ and

$$
(\mathbf{Z}^\top\mathbf{Z})_{kl} = \sum_{i=1}^{r}\mathbf{Z}_{ik}\mathbf{Z}_{il} = \sum_{i=1}^{r}\mathbf{1}_{\{i\in\Omega_k\text{ and }i\in\Omega_l\}} = |\Omega_k|\cdot\mathbf{1}_{\{k=l\}}.
$$

Thus, $\mathbf{Z}^\top\mathbf{Z} = \text{diag}(n_1, \ldots, n_r)$ and $\mathbf{X}^\top\text{diag}(n_1, \ldots, n_r)\mathbf{X} = \mathbf{X}^\top\mathbf{Z}^\top\mathbf{ZX} = \mathbf{I}_r$. This implies $\text{diag}(n_1^{1/2}, \ldots, n_r^{1/2})\mathbf{X}$ is an orthogonal matrix and

$$
\begin{aligned}
\mathbf{X}_{[k,:]}\mathbf{X}_{[l,:]}^\top &= (\mathbf{XX}^\top)_{kl} \\
&= \frac{1}{\sqrt{n_k n_l}}\left(\text{diag}(n_1^{1/2}, \ldots, n_r^{1/2})\mathbf{XX}^\top\text{diag}(n_1^{1/2}, \ldots, n_r^{1/2})\right)_{kl} \\
&= (n_k n_l)^{-1/2}\cdot\mathbf{1}_{\{k=l\}}, \quad \forall 1 \leq k, l \leq r.
\end{aligned}
$$

Therefore, for any two states i, j, if $i \in \Omega_k, j \in \Omega_l$, we have

$$
\begin{aligned}
\|(\mathbf{U}_P)_{[i,:]} - (\mathbf{U}_P)_{[j,:]}\|_2^2 &= \|(\mathbf{ZX})_{[i,:]} - (\mathbf{ZX})_{[j,:]}\|_2^2 \\
&= \|\mathbf{X}_{[k,:]} - \mathbf{X}_{[l,:]}\|_2^2 = \|\mathbf{X}_{[k,:]}\|_2^2 + \|\mathbf{X}_{[l,:]}\|_2^2 + 2\mathbf{X}_{[k,:]}^\top\mathbf{X}_{[l,:]} \\
&= \frac{1}{|\Omega_k|} + \frac{1}{|\Omega_l|} - 2\left(|\Omega_k|\cdot|\Omega_l|\right)^{-1/2}\cdot\mathbf{1}_{\{k=l\}} \\
&= \begin{cases} 0, & i \text{ and } j \text{ belong to the same group, i.e., } k = l; \\ \frac{1}{|\Omega_k|} + \frac{1}{|\Omega_l|}, & \\ & \text{otherwise.} \end{cases}
\end{aligned}
$$

Next, the k-means misclassification rate can be bounded by the $\sin\Theta$ distance between $\hat{\mathbf{U}}_P$ and $\mathbf{U}_P$ [42, Lemma 5.3]:

$$
\begin{aligned}
M(\hat{\Omega}_1, \ldots, \hat{\Omega}_r) &\leq \left(C\min_{\mathbf{O}\in\mathbb{O}_r}\|\hat{\mathbf{U}}_P - \mathbf{U}_P\mathbf{O}\|_F^2\right)\wedge r \\
&\leq \left(C\left\|\sin\Theta\left(\hat{\mathbf{U}}_P, \mathbf{U}_P\right)\right\|_F^2\right)\wedge r, \quad (48)
\end{aligned}
$$

where C is a uniform constant and $\mathbb{O}_r$ is the class of all r-by-r orthogonal matrices. Since $\mathbf{P}$ is state-aggregatable with respect to r groups, by Proposition 3, $\text{rank}(\mathbf{P}) \leq r$ and $\sigma_{r+1}(\mathbf{P}) = 0$. Based on the proof of Theorem 4, we have

$$
\begin{aligned}
\mathbb{P}(\mathcal{A}) \geq 1 - n^{-c}, \quad \mathcal{A} = \Bigg\{ &\left\|\sin\Theta(\hat{\mathbf{U}}_P, \mathbf{U}_P)\right\|_F^2 \\
&\leq \frac{C\|\mathbf{P}\|^2 pr\cdot\tau_*\log^2(n)\cdot\pi_{\max}/(\pi_{\min}^2 p)}{n\sigma_r^2(\mathbf{P})}\wedge r\Bigg\} \quad (49)
\end{aligned}
$$

for some $c > 1$. Combining (48), (49), and the trivial bound $M(\hat{\Omega}_1, \ldots, \hat{\Omega}_r) \leq r$, we have

$$
\begin{aligned}
&\mathbb{E}M(\hat{\Omega}_1, \ldots, \hat{\Omega}_r) \\
&= \mathbb{E}M(\hat{\Omega}_1, \ldots, \hat{\Omega}_r)\mathbf{1}_{\mathcal{A}} + \mathbb{E}M(\hat{\Omega}_1, \ldots, \hat{\Omega}_r)\mathbf{1}_{\mathcal{A}^c} \\
&\leq \left(\frac{C\|\mathbf{P}\|^2 pr\cdot\tau_*\log^2(n)\cdot\pi_{\max}/(\pi_{\min}^2 p)}{n\sigma_r^2(\mathbf{P})} + rn^{-c}\right)\wedge r.
\end{aligned}
$$

Since $\|\mathbf{P}\|/\sigma_r(\mathbf{P}) \geq 1$ and $c > 1$, one has $r/n^c \leq \frac{C\|\mathbf{P}\|^2 pr\cdot\tau_*\log^2(n)\cdot\pi_{\max}/(\pi_{\min}^2 p)}{n\sigma_r^2(\mathbf{P})}$. Then,

$$
\mathbb{E}M(\hat{\Omega}_1, \ldots, \hat{\Omega}_r) \leq \frac{C\|\mathbf{P}\|^2 pr\cdot\tau_*\log^2(n)\cdot\pi_{\max}/(\pi_{\min}^2 p)}{n\sigma_r^2(\mathbf{P})}\wedge r,
$$

which has finished the proof for Theorem 6. $\square$

E. Proof of Theorem 7

Denote $\mathbf{E} = \tilde{\mathbf{F}} - \mathbf{F}$. Recall from Prop. 4 and the discussions in its proof, $\mathbf{F}$ can be decomposed as $\mathbf{F} = \mathbf{F}_1 + \mathbf{F}_2$, where $\mathbf{F}_1$ is a rank-r matrix and the right singular vectors $\mathbf{V}_{F_1}$ has piece-wise constant structure, i.e., $(\mathbf{V}_{F_1})_{[i,:]} = (\mathbf{V}_{F_1})_{[i',:]}$ whenever i, i' belong to the same group. Based on the problem set-up, we can assume that the SVDs of $\mathbf{F}_1$ and $\tilde{\mathbf{F}}$ are

$$
\mathbf{F}_1 = \mathbf{U}_{F_1}\mathbf{\Sigma}_{F_1}\mathbf{V}_{F_1}^\top, \quad \tilde{\mathbf{F}} = \hat{\mathbf{U}}_F\hat{\mathbf{\Sigma}}_F\hat{\mathbf{V}}_F^\top + \hat{\mathbf{U}}_{F,\perp}\hat{\mathbf{\Sigma}}_{F,\perp}\hat{\mathbf{V}}_{F,\perp}^\top.
$$

Here, $\mathbf{U}_{F_1}, \mathbf{V}_{F_1}, \hat{\mathbf{U}}_F, \hat{\mathbf{V}}_F \in \mathbb{O}_{p,r}$, $\hat{\mathbf{U}}_{F,\perp}, \hat{\mathbf{V}}_{F,\perp} \in \mathbb{O}_{p,p-r}$ are the orthogonal complement of $\hat{\mathbf{U}}_F, \hat{\mathbf{V}}_F$. $\mathbf{\Sigma}_{F_1}, \hat{\mathbf{\Sigma}}_F$, and $\hat{\mathbf{\Sigma}}_{F,\perp}$ are diagonal matrices with non-negative and non-increasing diagonal entries; $\hat{\mathbf{U}}_F\hat{\mathbf{\Sigma}}_F\hat{\mathbf{V}}_F^\top$ correspond to the leading r principal components of $\tilde{\mathbf{F}}$, while $\hat{\mathbf{U}}_{F,\perp}\hat{\mathbf{\Sigma}}_{F,\perp}\hat{\mathbf{V}}_{F,\perp}^\top$ correspond to

the remainders. Since $\tilde{\mathbf{F}} - \mathbf{F}_1 = (\tilde{\mathbf{F}} - \mathbf{F}) + \mathbf{F} - \mathbf{F}_1 = \mathbf{E} + \mathbf{F}_2$, Wedin's perturbation lemma [71] implies

$$
\begin{aligned}
&\left\| \sin \Theta(\hat{\mathbf{V}}_F, \mathbf{V}_{F_1}) \right\|_F \\
&\leq \frac{\max\{\|(\mathbf{E} + \mathbf{F}_2)\hat{\mathbf{V}}_F\|_F, \|\hat{\mathbf{U}}_F(\mathbf{E} + \mathbf{F}_2)\|_F\}}{\sigma_{\min}(\hat{\boldsymbol{\Sigma}}_F) - 0} \wedge \sqrt{r}.
\end{aligned}
$$

Note that for any matrix $\mathbf{M}$,

$$
\begin{aligned}
\|\mathbf{M}\|_F &= \left(\sum_{i=1}^{\mathrm{rank}(\mathbf{M})} \sigma_i^2(\mathbf{M}) \right)^{1/2} \leq \sqrt{\mathrm{rank}(\mathbf{M})}\sigma_1(\mathbf{M}) \\
&= \sqrt{\mathrm{rank}(\mathbf{M})}\|\mathbf{M}\|.
\end{aligned}
$$

Provided that $\hat{\mathbf{U}}_F$ and $\hat{\mathbf{V}}_F$ are p-by-r matrices with orthogonal columns, we have

$$
\max \left\{ \mathrm{rank}(\mathbf{F}_2\hat{\mathbf{V}}_F), \mathrm{rank}(\mathbf{E}\hat{\mathbf{V}}_F), \mathrm{rank}(\hat{\mathbf{U}}_F^\top \mathbf{F}_2), \mathrm{rank}(\hat{\mathbf{U}}_F^\top \mathbf{E}) \right\} \leq r
$$

and

$$
\begin{aligned}
&\max \left\{ \|(\mathbf{E} + \mathbf{F}_2)\hat{\mathbf{V}}_F\|_F, \|\hat{\mathbf{U}}_F^\top(\mathbf{E} + \mathbf{F}_2)\|_F \right\} \\
&\leq \max \left\{ \|\mathbf{E}\hat{\mathbf{V}}_F\|_F, \|\hat{\mathbf{U}}_F^\top \mathbf{E}\|_F \right\} \\
&\quad + \max \left\{ \|\mathbf{F}_2\hat{\mathbf{V}}_F\|_F, \|\hat{\mathbf{U}}_F^\top \mathbf{F}_2\|_F \right\} \\
&\leq \max \left\{ \sqrt{\mathrm{rank}(\mathbf{E}\hat{\mathbf{V}}_F)}\|\mathbf{E}\hat{\mathbf{V}}_F\|, \sqrt{\mathrm{rank}(\hat{\mathbf{U}}_F^\top \mathbf{E})}\|\hat{\mathbf{U}}_F^\top \mathbf{E}\| \right\} \\
&\quad + (\sqrt{r}\|\mathbf{F}_2\|) \wedge \|\mathbf{F}_2\|_F \\
&\leq \sqrt{r}\|\mathbf{E}\| + (\sqrt{r}\|\mathbf{F}_2\|) \wedge \|\mathbf{F}_2\|_F.
\end{aligned}
$$

Since $\mathbf{F}_1\mathbf{F}_2^\top = 0$, Lemma 2 in [44] implies $\sigma_r(\mathbf{F}) = \sigma_r(\mathbf{F}_1 + \mathbf{F}_2) \geq \sigma_r(\mathbf{F}_1)$; by Weyl's perturbation bound [72], $|\sigma_r(\tilde{\mathbf{F}}) - \sigma_r(\mathbf{F})| \leq \|\tilde{\mathbf{F}} - \mathbf{F}\| = \|\mathbf{E}\|$. These two inequalities together imply

$$
\sigma_{\min}(\hat{\boldsymbol{\Sigma}}_F) = \sigma_r(\tilde{\mathbf{F}}) \geq \sigma_r(\mathbf{F}) - \|\mathbf{E}\| \geq \sigma_r(\mathbf{F}_1) - \|\mathbf{E}\|.
$$

Therefore,

$$
\left\| \sin \Theta(\hat{\mathbf{V}}_F, \mathbf{V}_{F_1}) \right\|_F \leq \frac{\sqrt{r}\|\mathbf{E}\| + (\sqrt{r}\|\mathbf{F}_2\|) \wedge \|\mathbf{F}_2\|_F}{\sigma_r(\mathbf{F}_1) - \|\mathbf{E}\|} \wedge \sqrt{r}.
$$

Note that for any real values $z \geq 0, y \geq x \geq 0$,

$$
x/y = 1 - (y - x)/y \leq 1 - (y - x)/(y + z) = (x + z)/(y + z).
$$

Thus, if $\|\mathbf{E}\| + \|\mathbf{F}_2\| \wedge (\|\mathbf{F}_2\|_F/\sqrt{r}) \leq \sigma_r(\mathbf{F}_1) - \|\mathbf{E}\|$,

$$
\begin{aligned}
&\frac{1}{\sqrt{r}} \left\| \sin \Theta(\hat{\mathbf{V}}_F, \mathbf{V}_{F_1}) \right\|_F \\
&\leq \frac{\|\mathbf{E}\| + \|\mathbf{F}_2\| \wedge (\|\mathbf{F}_2\|_F/\sqrt{r})}{\sigma_r(\mathbf{F}_1) - \|\mathbf{E}\|} \wedge 1 \\
&\leq \frac{\|\mathbf{E}\| + \|\mathbf{F}_2\| \wedge (\|\mathbf{F}_2\|_F/\sqrt{r}) + \|\mathbf{E}\|}{\sigma_r(\mathbf{F}_1) - \|\mathbf{E}\| + \|\mathbf{E}\|} \wedge 1 \\
&= \frac{2\|\mathbf{E}\| + \|\mathbf{F}_2\| \wedge (\|\mathbf{F}_2\|_F/\sqrt{r})}{\sigma_r(\mathbf{F}_1)} \wedge 1;
\end{aligned}
$$

if $\|\mathbf{E}\| + \|\mathbf{F}_2\| \wedge (\|\mathbf{F}_2\|_F/\sqrt{r}) > \sigma_r(\mathbf{F}_1) - \|\mathbf{E}\|$,

$$
\begin{aligned}
&\frac{1}{\sqrt{r}} \left\| \sin \Theta(\hat{\mathbf{V}}_F, \mathbf{V}_{F_1}) \right\|_F \\
&\leq \frac{\|\mathbf{E}\| + \|\mathbf{F}_2\| \wedge (\|\mathbf{F}_2\|_F/\sqrt{r})}{\sigma_r(\mathbf{F}_1) - \|\mathbf{E}\|} \wedge 1 \\
&= 1 = \frac{2\|\mathbf{E}\| + \|\mathbf{F}_2\| \wedge (\|\mathbf{F}_2\|_F/\sqrt{r})}{\sigma_r(\mathbf{F}_1)} \wedge 1.
\end{aligned}
$$

Therefore, we always have

$$
\left\| \sin \Theta(\hat{\mathbf{V}}_F, \mathbf{V}_{F_1}) \right\|_F \leq \frac{2\sqrt{r}\|\mathbf{E}\| + (\sqrt{r}\|\mathbf{F}_2\|) \wedge \|\mathbf{F}_2\|_F}{\sigma_r(\mathbf{F}_1)} \wedge \sqrt{r}.
$$

By Lemma 7, there exists constants $C > 0$ such that

$$
\mathbb{P} \left(\|\mathbf{E}\| = \left\| \tilde{\mathbf{F}} - \mathbf{F} \right\| \leq C\sqrt{\frac{\pi_{\max}\tau_* \log^2(n)}{n}} \right) \geq 1 - n^{-c}.
$$

for some $c > 1$. This implies the following upper bound for the $\sin \Theta$ loss of $\hat{\mathbf{V}}_F$,

$$
\begin{aligned}
&\mathbb{P}(\mathcal{A}) \geq 1 - n^{-c}, \\
&\text{where } \mathcal{A} = \left\{ \left\| \sin \Theta(\hat{\mathbf{V}}_F, \mathbf{V}_{F_1}) \right\|_F \right. \\
&\left. \leq \frac{C\sqrt{\pi_{\max}r\tau_* \log^2(n)/n} + (\sqrt{r}\|\mathbf{F}_2\|) \wedge \|\mathbf{F}_2\|_F}{\sigma_r(\mathbf{F}_1)} \wedge \sqrt{r} \right\}.
\end{aligned}
$$

(50)

Next, we prove the upper bound for the misclassification rate of r-means based on (50). By Proposition 4, each column of $\mathbf{V}_{F_1}$ is piece-wise constant with respect to partitions $\Omega_1, \ldots, \Omega_r$ and we can write $\mathbf{V}_{F_1} = \mathbf{ZX}$, where $\mathbf{Z} \in \mathbb{R}^{p \times r}$ is the membership indicator,

$$
\mathbf{Z}_{ij} = \begin{cases} 1, & i\text{-th state} \in \Omega_j; \\ 0, & i\text{-th state} \notin \Omega_j, \end{cases}
$$

and $\mathbf{X} \in \mathbb{R}^{r \times r}, \mathbf{X}_{kj} = (\mathbf{V}_{F_1})_{ij}, \forall i \in [p], j \in [r], i \in \Omega_k$. Since the columns of $\mathbf{V}_{F_1}$ are orthonormal, $\mathbf{X}^\top \mathbf{Z}^\top \mathbf{ZX} = \mathbf{V}_{F_1}^\top \mathbf{V}_{F_1} = \mathbf{I}_r$. Denote $n_k = |\Omega_k|, k = 1, \ldots, r$. Note that

$$
(\mathbf{Z}^\top \mathbf{Z})_{kl} = \sum_{i=1}^r \mathbf{Z}_{ik}\mathbf{Z}_{il} = \sum_{i=1}^r 1_{\{i \in \Omega_k \text{ and } i \in \Omega_l\}} = |\Omega_k| \cdot 1_{\{k=l\}}.
$$

Thus, $\mathbf{Z}^\top \mathbf{Z} = \mathrm{diag}(n_1, \ldots, n_r)$ and $\mathbf{X}^\top \mathrm{diag}(n_1, \ldots, n_r)\mathbf{X} = \mathbf{X}^\top \mathbf{Z}^\top \mathbf{ZX} = \mathbf{I}_r$. This implies $\mathrm{diag}(n_1^{1/2}, \ldots, n_r^{1/2})\mathbf{X}$ is an orthogonal matrix and

$$
\begin{aligned}
\mathbf{X}_{[k,:]}\mathbf{X}_{[l,:]}^\top &= (\mathbf{XX}^\top)_{kl} \\
&= n_k^{-1/2}\left(\mathrm{diag}(n_1^{1/2}, \ldots, n_r^{1/2})\mathbf{XX}^\top \mathrm{diag}(n_1^{1/2}, \ldots, n_r^{1/2}) \right)_{kl} \\
&\quad \times n_l^{-1/2} \\
&= (n_k \cdot n_l)^{-1/2} \cdot 1_{\{k=l\}}, \quad \forall 1 \leq k, l \leq r.
\end{aligned}
$$

(51)

Therefore, for any two states i, j, if $i \in \Omega_k$, $j \in \Omega_l$, then

$$
\begin{aligned}
& \|(\mathbf{V}_{F_1})_{[i,:]} - (\mathbf{V}_{F_1})_{[j,:]}\|_2^2 \\
= & \|(\mathbf{Z}\mathbf{X})_{[i,:]} - (\mathbf{Z}\mathbf{X})_{[j,:]}\|_2^2 = \|\mathbf{X}_{[k,:]} - \mathbf{X}_{[l,:]}\|_2^2 \\
= & \|\mathbf{X}_{[k,:]}\|_2^2 + \|\mathbf{X}_{[l,:]}\|_2^2 + 2\mathbf{X}_{[k,:]}^\top \mathbf{X}_{[l,:]} \\
= & \frac{1}{|\Omega_k|} + \frac{1}{|\Omega_l|} - 2\left(|\Omega_k| \cdot |\Omega_l|\right)^{-1/2} \cdot 1_{\{k=l\}} \\
= & \begin{cases} 0, & i \text{ and } j \text{ belong to the same group, i.e., } k = l; \\ \frac{1}{|\Omega_k|} + \frac{1}{|\Omega_l|}, & \\ & \text{otherwise.} \end{cases}
\end{aligned}
$$
(52)

Next, the error bound of k-means approximation [42, Lemma 5.3] yields

$$
\begin{aligned}
M(\hat{\Omega}_1, \ldots, \hat{\Omega}_r) & \leq C \min_{\mathbf{O} \in \mathbb{O}_r} \|\hat{\mathbf{V}}_F - \mathbf{V}_{F_1} \mathbf{O}\|_F^2 \wedge r \\
& \leq C \left\| \sin\Theta\left(\hat{\mathbf{V}}_F, \mathbf{V}_{F_1}\right) \right\|_F^2 \wedge r,
\end{aligned}
$$
(53)

where C is a uniform constant and $\mathbb{O}_r$ is the class of all r-by-r orthogonal matrices. Combining (50) and (53) and the trivial bound $M(\hat{\Omega}_1, \ldots, \hat{\Omega}_r) \leq r$, we have

$$
\begin{aligned}
& \mathbb{E} M(\hat{\Omega}_1, \ldots, \hat{\Omega}_r) \\
= & \mathbb{E} M(\hat{\Omega}_1, \ldots, \hat{\Omega}_r) 1_{\mathcal{A}} + \mathbb{E} M(\hat{\Omega}_1, \ldots, \hat{\Omega}_r) 1_{\mathcal{A}^c} \\
\leq & \left(\frac{C\left(\pi_{\max} r \tau_* \log^2(n)/n + (r\|\mathbf{F}_2\|^2) \wedge \|\mathbf{F}_2\|_F^2\right)}{\sigma_r^2(\mathbf{F}_1)} \right. \\
& \left. + r/n^c \right) \wedge r.
\end{aligned}
$$

By the proof of Theorem 4, one has $\sigma_r(\mathbf{F}) \leq 1$. Thus, if $c \geq 1$, one has $r/n^c \leq \frac{C\pi_{\max} r \tau_* \log^2(n)/n}{\sigma_r^2(\mathbf{F})}$ and

$$
\begin{aligned}
& \mathbb{E} M(\hat{\Omega}_1, \ldots, \hat{\Omega}_r) \\
\leq & \frac{C\left(\pi_{\max} r \tau_* \log^2(n)/n + (r\|\mathbf{F}_2\|^2) \wedge \|\mathbf{F}_2\|_F^2\right)}{\sigma_r^2(\mathbf{F}_1)} \wedge r,
\end{aligned}
$$

which has finished the proof for Theorem 7.

XI. Technical Lemmas

We collect the technical lemmas for the main results in this section. The first Lemma 1 demonstrates a sufficient and necessary condition for being transition and frequency matrices of some ergodic Markov chain.

Lemma 1 (Properties of transition and frequency matrices for ergodic Markov process). $\mathbf{P}, \mathbf{F} \in \mathbb{R}^{p \times p}$ *are the transition matrix and frequency matrix of some ergodic finite-state-space Markov process if and only if*

$$
\mathbf{P} \in \mathcal{P}_p = \left\{ \mathbf{P} : \begin{array}{l} 0 \leq \mathbf{P}_{ij} \leq 1; \forall 1 \leq i \leq p, \sum_{j=1}^p \mathbf{P}_{ij} = 1; \\ \forall I \subseteq \{1, \ldots, p\}, \mathbf{P}_{[I, I^c]} \neq 0 \end{array} \right\},
$$
(54)

and

$$
\mathbf{F} \in \mathcal{F}_p = \left\{ \mathbf{F} \in \mathbb{R}^{p \times p} : \begin{array}{l} \mathbf{F} 1_p = \mathbf{F}^\top 1_p, \quad 1_p^\top \mathbf{F} 1_p = 1, \\ \forall I \subseteq \{1, \ldots, p\}, \mathbf{F}_{[I, I^c]} \neq 0 \end{array} \right\}.
$$
(55)

Proof of Lemma 1. The proof for the transition matrix (54) is by definition. Then we consider the condition for $\mathbf{F}$. When $\mathbf{F} \in \mathbb{R}^{p \times p}$ is the frequency matrix of some ergodic Markov chain, we have $\mathbf{F} = \operatorname{diag}(\pi)\mathbf{P}$, where π and $\mathbf{P}$ are the corresponding invariant distribution and stochastic matrix. Then

$$
\begin{aligned}
\mathbf{F} 1_p & = \operatorname{diag}(\pi)\mathbf{P} 1_p = \operatorname{diag}(\pi) 1_p = \pi, \\
\mathbf{F}^\top 1_p & = \mathbf{P}^\top \operatorname{diag}(\pi) 1_p = \mathbf{P}^\top \pi = \pi = \mathbf{F} 1_p, \\
& 1_p^\top \mathbf{F} 1_p = 1_p^\top \pi = 1.
\end{aligned}
$$

Here we used the fact that $\pi^\top \mathbf{P} = \pi^\top$ and $\mathbf{P} 1_p = 1_p$. Next, since the finite-state-space Markov process is ergodic, $\pi_i > 0$ for any i. Thus for any $I \subseteq \{1, \ldots, p\}$, $\mathbf{F}_{[I, I^c]} = \operatorname{diag}(\pi_I) \cdot \mathbf{P}_{[I, I^c]} \neq 0$. This implies $\mathbf{F} \in \mathcal{F}_p$.

On the other hand when $\mathbf{F} \in \mathcal{F}_p$, we define $\pi = \mathbf{F} 1_p$, $\mathbf{P} = \operatorname{diag}(\pi^{-1})\mathbf{F}$. Since $\mathbf{F}_{[\{i\}, \{i\}^c]} \neq 0$, we have $\pi_i \neq 0$ for any $1 \leq i \leq p$. Then $\mathbf{P}$ is well-defined. In addition, π and $\mathbf{P}$ satisfies the following properties

$$
\begin{aligned}
& 1_p^\top \pi = 1_p^\top \mathbf{F} 1_p = 1, \quad \mathbf{P}_{ij} \geq 0, \\
& \mathbf{P} 1_p = \operatorname{diag}(\pi^{-1})\mathbf{F} 1_p = \operatorname{diag}(\pi^{-1})\pi = 1_p, \\
& \pi^\top \mathbf{P} = \pi^\top \operatorname{diag}(\pi)^{-1}\mathbf{F} = 1_p^\top \mathbf{F} = (\mathbf{F}^\top 1_p)^\top = (\mathbf{F} 1_p)^\top = \pi, \\
& \forall I \subseteq \{1, \ldots, p\}, \mathbf{P}_{[I, I^c]} = \operatorname{diag}(\pi_I^{-1}) \cdot \mathbf{F}_{[I, I^c]} \neq 0.
\end{aligned}
$$
(56)

By comparing above properties with the definition of ergodic transition matrix (54), we can see $\mathbf{F}$ is indeed a frequency matrix of some ergodic Markov process. Thus, we have finished the proof of this lemma. $\square$

The next Lemma 2 characterizes the ℓ_1 distance between two vectors after ℓ_1 normalization, which will be used in the upper bound argument in the main context of the paper.

Lemma 2. *Suppose $u, v \neq 0$ are two vectors of the same dimension, then*

$$
\left\| \frac{u}{\|u\|_1} - \frac{v}{\|v\|_1} \right\|_1 \leq \frac{2\|u - v\|_1}{\max\{\|u\|_1, \|v\|_1\}}.
$$
(57)

Proof of Lemma 2.

$$
\begin{aligned}
& \left\| \frac{u}{\|u\|_1} - \frac{v}{\|v\|_1} \right\|_1 \leq \left\| \frac{u - v}{\|u\|_1} \right\|_1 + \left\| \frac{v}{\|u\|_1} - \frac{v}{\|v\|_1} \right\|_1 \\
= & \frac{\|u - v\|_1}{\|u\|_1} + \frac{|\|u\|_1 - \|v\|_1|}{\|u\|_1} \leq \frac{2\|u - v\|_1}{\|u\|_1}.
\end{aligned}
$$

Similarly, $\left\| \frac{u}{\|u\|_1} - \frac{v}{\|v\|_1} \right\|_1 \leq \frac{2\|u-v\|_1}{\|v\|_1}$, which implies (57). $\square$

The following Lemma 3 demonstrate the error for truncated singular value decomposition.

Lemma 3. *For any matrix $\mathbf{M}$ with singular value decomposition $\mathbf{M} = \sum_{k \geq 1} \sigma_k u_k v_k^\top$ and $r \geq 1$, we define $\mathbf{M}_{\max(r)} = \sum_{k=1}^r \sigma_k u_k v_k^\top$ and $\mathbf{M}_{-\max(r)} = \sum_{k \geq r+1} \sigma_k u_k v_k^\top = \mathbf{M} - \mathbf{M}_{\max(r)}$ as the leading and non-leading parts of $\mathbf{M}$. Suppose $\tilde{\mathbf{A}}$ and $\mathbf{A}$ are any two matrices of the same dimension. Then,*

$$
\begin{aligned}
& \left\| \tilde{\mathbf{A}}_{\max(r)} - \mathbf{A} \right\|_F \\
\leq & 2\sqrt{2r} \left\| \tilde{\mathbf{A}} - \mathbf{A} \right\| + 2\sqrt{2r} \|\mathbf{A}_{-\max(r)}\| + \|\mathbf{A}_{-\max(r)}\|_F.
\end{aligned}
$$
(58)

*Particularly, if rank(**A**) $\leq r$, we also have*

$$\left\| \tilde{\mathbf{A}}_{\max(r)} - \mathbf{A} \right\|_F \leq 2\|\tilde{\mathbf{A}} - \mathbf{A}\|_F. \tag{59}$$

Proof of Lemma 3. Note that $\tilde{\mathbf{A}}_{\max(r)}$ and $\mathbf{A}_{\max(r)}$ are both of rank-r, thus $\tilde{\mathbf{A}}_{\max(r)} - \mathbf{A}_{\max(r)}$ is of rank at most $2r$, and $\|\tilde{\mathbf{A}}_{\max(r)} - \mathbf{A}_{\max(r)}\|_F \leq \sqrt{2r}\|\tilde{\mathbf{A}}_{\max(r)} - \mathbf{A}_{\max(r)}\|$. By Weyl's inequality [72], $\sigma_{r+1}(\tilde{\mathbf{A}}) \leq \sigma_{r+1}(\mathbf{A}) + \|\mathbf{A} - \tilde{\mathbf{A}}\|$ for any r. Therefore,

$$\|\tilde{\mathbf{A}}_{\max(r)} - \mathbf{A}\|_F$$
$$\leq \|\tilde{\mathbf{A}}_{\max(r)} - \mathbf{A}_{\max(r)}\|_F + \|\mathbf{A}_{-\max(r)}\|_F$$
$$\leq \sqrt{2r}\|\tilde{\mathbf{A}}_{\max(r)} - \mathbf{A}_{\max(r)}\| + \|\mathbf{A}_{-\max(r)}\|_F$$
$$\leq \sqrt{2r}\left(\|\tilde{\mathbf{A}} - \mathbf{A}\| + \|\tilde{\mathbf{A}}_{-\max(r)}\| + \|\mathbf{A}_{-\max(r)}\| \right)$$
$$\quad + \|\mathbf{A}_{-\max(r)}\|_F$$
$$= \sqrt{2r}\left(\|\tilde{\mathbf{A}} - \mathbf{A}\| + \sigma_{r+1}(\tilde{\mathbf{A}}) + \sigma_{r+1}(\mathbf{A}) \right) + \|\mathbf{A}_{-\max(r)}\|_F$$
$$\overset{(a)}{\leq} \sqrt{2r}\left(\|\tilde{\mathbf{A}} - \mathbf{A}\| + 2\sigma_{r+1}(\mathbf{A}) + \|\tilde{\mathbf{A}} - \mathbf{A}\| \right) + \|\mathbf{A}_{-\max(r)}\|_F$$
$$= 2\sqrt{2r}\|\tilde{\mathbf{A}} - \mathbf{A}\| + 2\sqrt{2r}\|\mathbf{A}_{-\max(r)}\| + \|\mathbf{A}_{-\max(r)}\|_F,$$

which yields (58). Here, (a) is due to the Weyl's inequality. Additionally, if rank(A) $\leq r$, we have

$$\left\| \tilde{\mathbf{A}}_{\max(r)} - \mathbf{A} \right\|_F \leq \|\tilde{\mathbf{A}}_{\max(r)} - \tilde{\mathbf{A}}\|_F + \|\tilde{\mathbf{A}} - \mathbf{A}\|_F$$
$$= \min_{\text{rank}(\mathbf{M}) \leq r} \|\tilde{\mathbf{A}} - \mathbf{M}\|_F + \|\tilde{\mathbf{A}} - \mathbf{A}\|_F$$
$$\leq \|\tilde{\mathbf{A}} - \mathbf{A}\|_F + \|\tilde{\mathbf{A}} - \mathbf{A}\|_F = 2\|\tilde{\mathbf{A}} - \mathbf{A}\|_F.$$

which yields (59). $\square$

Our next lemma characterizes the relation between KL divergence and ℓ_2 distance between two discrete distribution vectors.

Lemma 4. *For any two distributions $u, v \in \mathbb{R}^p$, such that $\sum_{i=1}^p u_i = 1$, $\sum_{i=1}^p v_i = 1$. If there exists $0 < a \leq 1/p \leq b$ such that $a \leq u_i, v_i \leq b$ for $1 \leq i \leq p$, then the KL-divergence and ℓ_2 norm distance are equivalent, in the sense that,*

$$\frac{a}{2b^2}\|u - v\|_2^2 \leq D_{KL}(u\|v) \leq \frac{b}{2a^2}\|u - v\|_2^2, \tag{60}$$

Here $D_{KL}(u\|v) = \sum_{i=1}^p u_i \log(u_i/v_i)$ is the KL-divergence between u and v.

Proof of Lemma 4. By Taylor's expansion, there exists ξ_i between u_i and v_i, such that

$$\log(v_i/u_i) = \log(v_i) - \log(u_i) = \frac{v_i - u_i}{u_i} - \frac{(v_i - u_i)^2}{2\xi_i^2},$$

Thus,

$$D_{KL}(u\|v) = \sum_{i=1}^p -u_i \log(v_i/u_i)$$
$$= \sum_{i=1}^p \left\{ -(v_i - u_i) + \frac{u_i(v_i - u_i)^2}{2\xi_i^2} \right\}$$
$$\leq \sum_{i=1}^p \frac{b(u_i - v_i)^2}{2a^2} = \frac{b}{2a^2}\|u - v\|_2^2;$$

$$D_{KL}(u\|v) = \sum_{i=1}^p -u_i \log(v_i/u_i)$$
$$= \sum_{i=1}^p \left\{ -(v_i - u_i) + \frac{u_i(v_i - u_i)^2}{2\xi_i^2} \right\}$$
$$\geq \sum_{i=1}^p \frac{a(u_i - v_i)^2}{2b^2} = \frac{a}{2b^2}\|u - v\|_2^2,$$

which has finished the proof for this lemma. $\square$

The following Lemma 5 establishes a Markov mixing time comparison inequality between $\tau(\varepsilon)$ and $\tau(\delta)$ for any values ε and δ. This result is slightly more general than Theorem 4.9 in [52].

Lemma 5 (Markov Mixing Rate). *Let $\tau(\varepsilon)$ be the mixing time defined in (1) where $\varepsilon \leq \delta < 1/2$, then*

$$\tau(\varepsilon) \leq \tau(\delta) \cdot \left(\left\lceil \frac{\log(\varepsilon/\delta)}{\log(2\delta)} \right\rceil + 1 \right). \tag{61}$$

Proof of Lemma 5. We denote $\{e^{(i)}\}_{i=1}^p$ as the canonical basis for $\mathbb{R}^p$, namely $e^{(i)}$ is equal to 1 in its i-th entry and equal to 0 elsewhere. For any vector $\theta \in \mathbb{R}^p$, we also use $\theta_+, \theta_- \in \mathbb{R}^p$ to denote the positive and negative parts of θ, respectively, i.e.

$$(\theta_+)_j = \min\{\theta_j, 0\}, \quad (\theta_-)_j = -\max\{\theta_j, 0\}, \quad 1 \leq j \leq p. \tag{62}$$

Clearly $\theta_+ \geq 0, \theta_- \geq 0$, and $\theta = \theta_+ - \theta_-$. Suppose $k = \tau(\delta)$, then for any distribution $\theta \in \mathbb{R}^p$ with $\sum_i \theta_i = 1, \theta_i \geq 0$, and any integer $k' \geq k$, we must have

$$\frac{1}{2}\left\| (\mathbf{P}^\top)^{k'}\theta - \pi \right\|_1 = \frac{1}{2}\left\| \sum_{i=1}^p (\mathbf{P}^\top)^{k'}\theta_i e^{(i)} - \pi \right\|_1$$
$$\leq \sum_{i=1}^p |\theta_i| \cdot \frac{1}{2}\left\| (\mathbf{P}^\top)^{k'} e^{(i)} - \pi \right\|_1 \leq \sum_{i=1}^p |\theta_i| \cdot \delta = \delta. \tag{63}$$

When θ and π are both distributions, $\sum_{j=1}^p (\mathbf{P}^\top)^k \theta_j = \sum_{j=1}^p \pi_j = 1$, then $\sum_{j=1}^p ((\mathbf{P}^\top)^k \theta - \pi)_j = 0$, and

$$\left\| \left((\mathbf{P}^\top)^k \theta - \pi \right)_+ \right\|_1 = \left\| \left((\mathbf{P}^\top)^{k'} \theta - \pi \right)_- \right\|_1$$
$$= \frac{1}{2}\left\| (\mathbf{P}^\top)^k \theta - \pi \right\|_1. \tag{64}$$

Next, we consider any integer $k' \geq 2k$, then $k' - k$. One can calculate that

$$
\begin{aligned}
\frac{1}{2}\left\| (\mathbf{P}^\top)^{k'}\theta - \pi \right\|_1 &= \frac{1}{2}\left\| (\mathbf{P}^\top)^{k'-k}((\mathbf{P}^\top)^k\theta - \pi) \right\|_1 \\
&= \frac{1}{2}\left\| (\mathbf{P}^\top)^{k'-k}\left[((\mathbf{P}^\top)^k\theta - \pi)_+ - ((\mathbf{P}^\top)^k e_i - \pi)_- \right] \right\|_1 \\
&\leq \frac{1}{2}\left\| (\mathbf{P}^\top)^{k'-k}\frac{((\mathbf{P}^\top)^k\theta - \pi)_+}{\|((\mathbf{P}^\top)^k\theta - \pi)_+\|_1} - \pi \right\|_1 \\
&\quad \cdot \|((\mathbf{P}^\top)^k\theta - \pi)_+\|_1 \\
&\quad + \frac{1}{2}\left\| (\mathbf{P}^\top)^{k'-k}\frac{((\mathbf{P}^\top)^k\theta - \pi)_-}{\|((\mathbf{P}^\top)^k\theta - \pi)_-\|_1} - \pi \right\|_1 \\
&\quad \cdot \|((\mathbf{P}^\top)^k\theta - \pi)_-\|_1 \\
&\stackrel{(63)(64)}{\leq} \delta\left(\left\| ((\mathbf{P}^\top)^k\theta - \pi)_+ \right\|_1 + \left\| ((\mathbf{P}^\top)^k\theta - \pi)_- \right\|_1 \right) \\
&\leq \delta\|(\mathbf{P}^\top)^k\theta - \pi\| \leq \frac{1}{2}(2\delta)^2.
\end{aligned}
$$

By induction, one can show for any integers l, we must have

$$
\forall k' \geq lk, \quad \frac{1}{2}\|(\mathbf{P}^\top)^{k'}\theta - \pi\|_1 \leq \frac{1}{2}(2\delta)^l.
$$

Note that $\delta < 1/2$, $\varepsilon \leq \delta$, we set $l = \lceil \frac{\log(\varepsilon/\delta)}{\log(2\delta)} \rceil + 1$. Then for any $k' \geq kl$,

$$
\begin{aligned}
\frac{1}{2}\left\| (\mathbf{P}^\top)^{k'}\theta - \pi \right\|_1 &\leq \frac{1}{2}(2\delta)^l \leq \frac{1}{2}(2\delta)^{\frac{\log(\varepsilon/\delta)}{\log(2\delta)}+1} \\
&= \frac{1}{2}2\delta \cdot (\varepsilon/\delta) = \varepsilon, \tag{65}
\end{aligned}
$$

which implies $\tau(\varepsilon) \leq kl = \tau(\delta) \cdot (\lceil \log(\varepsilon/\delta)/\log(2\delta) \rceil + 1)$, and complete the proof for (61). Thus we have finished the proof for Lemma 5. $\square$

The next Lemma 6 relates the Markov mixing time to the eigengap condition.

Lemma 6 (Markov Mixing Time and Eigengap Condition ([52]. , Theorem 12.3)] *Suppose* $\mathbf{P} \in \mathbb{R}^{p \times p}$ *is the transition matrix of an ergodic and reversible Markov chain with invariant distribution* π. *Suppose* λ_2 *is its second largest eigenvalue, then* $\lambda_2 \in \mathbb{R}$, $|\lambda_2| \leq 1$, *and*

$$
\tau(\varepsilon) \leq \frac{1}{1 - \lambda_2} \log\left(\frac{1}{\varepsilon \pi_{\min}}\right). \tag{66}
$$

Lemma 7 (Markov Chain Concentration Inequality). *Suppose* $\mathbf{P} \in \mathbb{R}^{p \times p}$ *is an ergodic Markov chain transition matrix on* p *states* $\{1, \ldots, p\}$. $\mathbf{P}$ *is with invariant distribution* π *and the Markov mixing time* $\tau(\varepsilon)$ *defined as (1). Recall the frequency matrix is* $\mathbf{F} = \operatorname{diag}(\pi)\mathbf{P}$. *Given a Markov trajectory with* $(n+1)$ *observable states* $X = \{x_0, x_1, \ldots, x_n\}$ *from any initial state, we denote the empirical invariant distribution* $\tilde{\pi}$ *and empirical frequency matrix as*

$$
\tilde{\pi} = \frac{1}{n}\sum_{k=1}^{n} e_{x_k}, \quad \text{where } e_{x_k} \text{ is the indicator such that,}
$$

$$
(e_{x_k})_i = \begin{cases} 1, & x_k = i; \\ 0, & x_k \neq i; \end{cases} \tag{67}
$$

$$
\tilde{\mathbf{F}} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{E}_k, \quad \text{where } \mathbf{E}_k \in \mathbb{R}^{p \times p},
$$

$$
(\mathbf{E}_k)_{ij} = \begin{cases} 1, & (x_{k-1}, x_k) = (i, j); \\ 0, & \text{otherwise}. \end{cases} \tag{68}
$$

Let $t > 0$, $\alpha = \tau((t/2) \wedge \pi_{\max}) + 1$. *Recall* $\|\cdot\|$ *is defined as the matrix 2-norm,* $\|\cdot\|_\infty$ *is defined as the vector* ℓ_∞ *norm. Then*

$$
\forall t > 0, \mathbb{P}\left(\left\| \tilde{\mathbf{F}} - \mathbf{F} \right\| \geq t \right) \leq 2\alpha p \exp\left(-\frac{nt^2/8}{2\pi_{\max}\alpha + t\alpha/6} \right), \tag{69}
$$

$$
\forall t > 0, \mathbb{P}\left(\|\tilde{\pi} - \pi\|_\infty \geq t \right) \leq 2\alpha p \exp\left(-\frac{nt^2/8}{2\pi_{\max}\alpha + t\alpha/6} \right). \tag{70}
$$

For any constant $c_0 > 0$, *there exists constant* $C > 0$ *such that if* $n \geq C\left(\tau(\sqrt{\pi_{\max}/n})\log(n)/\pi_{\max} \vee p \right)$, *we have*

$$
\mathbb{P}\left(\left\| \tilde{\mathbf{F}} - \mathbf{F} \right\| \geq C\sqrt{\frac{\pi_{\max}\tau(\sqrt{\pi_{\max}/n})\log(n)}{n}} \right) \leq n^{-c_0}, \tag{71}
$$

$$
\mathbb{P}\left(\|\tilde{\pi} - \pi\|_\infty \geq C\sqrt{\frac{\pi_{\max}\tau(\sqrt{\pi_{\max}/n})\log(n)}{n}} \right) \leq n^{-c_0}, \tag{72}
$$

Additionally, let $\tau_* = \tau(1/4)$. *For any constant* $c_0 > 0$, *there exists constant* $C > 0$ *such that if* $n \geq C\tau_* p\log^2(n)$, *then*

$$
\mathbb{P}\left(\left\| \tilde{\mathbf{F}} - \mathbf{F} \right\| \geq C\sqrt{\frac{\pi_{\max}\tau_*\log^2(n)}{n}} \right) \leq n^{-c_0}, \tag{73}
$$

$$
\mathbb{P}\left(\|\tilde{\pi} - \pi\|_\infty \geq C\sqrt{\frac{\pi_{\max}\tau_*\log^2(n)}{n}} \right) \leq n^{-c_0}. \tag{74}
$$

When $\mathbf{P}$ *is reversible with second largest eigenvalue* $\lambda_2 < 1$ *and* $c_0 > 0$ *is any constant, there exists constant* $C > 0$ *such that if* $n \geq Cp\log(n)\log(n/\pi_{\min})$, *then*

$$
\mathbb{P}\left(\left\| \tilde{\mathbf{F}} - \mathbf{F} \right\| \geq C\sqrt{\frac{\pi_{\max}\log(n/\pi_{\min})\log(n)}{n(1 - \lambda_2)}} \right) \leq n^{-c_0}, \tag{75}
$$

$$
\mathbb{P}\left(\|\tilde{\pi} - \pi\|_\infty \geq C\sqrt{\frac{\pi_{\max}\log(n/\pi_{\min})\log(n)}{n(1 - \lambda_2)}} \right) \leq n^{-c_0}. \tag{76}
$$

Proof of Lemma 7. Let $n_0 = \lfloor n/\alpha \rfloor$. Without loss of generality, assume n is a multiple of α. We introduce the "thin" sequences as

$$
\begin{aligned}
\tilde{e}_k^{(l)} &= e_{x_{k\alpha+l}} - \mathbb{E}\left(e_{x_{k\alpha+l}} \big| x_{(k-1)\alpha+l} \right), \\
&\quad l = 1, \ldots, \alpha; k = 1, \ldots, n_0; \tag{77}
\end{aligned}
$$

$$
\begin{aligned}
\tilde{\mathbf{E}}_k^{(l)} &= \mathbf{E}_{k\alpha+l} - \mathbb{E}\left(\mathbf{E}_{k\alpha+l} | \mathbf{E}_{(k-1)\alpha+l} \right), \\
&\quad l = 1, \ldots, \alpha; k = 1, \ldots, n_0. \tag{78}
\end{aligned}
$$

By Jensen's inequality, for any $l = 1, \ldots, \alpha$, $k = 1, \ldots, n_0$,

$$\left\| \mathbb{E} \left(e_{x_{k\alpha+l}} \middle| e_{x_{(k-1)\alpha+l}} \right) \right\|_2 \leq \mathbb{E} \| e_{x_{k\alpha+l}} \|_2 \leq 1,$$
$$\left\| \mathbb{E} \left(\mathbf{E}_{k\alpha+l} \middle| \mathbf{E}_{(k-1)\alpha+l} \right) \right\| \leq \mathbb{E} \| \mathbf{E}_{k\alpha+l} \| \leq 1,$$

which implies

$$\left\| \tilde{e}_k^{(l)} \right\|_2 \leq 2, \quad \left\| \tilde{\mathbf{E}}_k^{(l)} \right\| \leq 2. \tag{79}$$

Now we develop the concentration inequalities of the partial sum sequences $\sum_{k=1}^{n_0} \tilde{\mathbf{E}}_k^{(l)}$ for any fixed l. Note that for any given $\tilde{\mathbf{E}}_{k-1}^{(l)}$ and $e_{x_{k-1}^{(l)}}$, i.e. given the values of $(x_{k\alpha+l-1}, x_{k\alpha+l})$ pair, the conditional distribution of $e_{x_{k\alpha+l-1}}$ satisfies

$$x_{k\alpha+l-1} | x_{(k-1)\alpha+l} \sim e_{x_{(k-1)\alpha+l}}^\top \mathbf{P}^{\alpha-1}, \quad k = 1, \ldots, n_0.$$

For convenience, we denote $\tilde{\pi}_{(k,l)} = \left(e_{x_{(k-1)\alpha+l}}^\top \mathbf{P}^{\alpha-1} \right)^\top \in \mathbb{R}^p$. By the choice of α and the mixing time property,

$$\| \tilde{\pi}_{(k,l)} - \pi \|_1 = \left\| e_{x_{(k-1)\alpha+l}}^\top \mathbf{P}^{\alpha-1} - \pi \right\|_1 \leq \min\{ t/2, \pi_{\max} \}. \tag{80}$$

(80) will be crucial to our later analysis. Note that

$$\tilde{\mathbf{E}}_k^{(l)} = \mathbf{E}_{k\alpha+l} - \mathbb{E} \left(\mathbf{E}_{k\alpha+l} \middle| x_{(k-1)\alpha+l} \right),$$
$$\text{where} \quad \mathbf{E}_{k\alpha+l} = e_{x_{k\alpha+l-1}} \cdot e_{x_{k\alpha+l}}^\top, \tag{81}$$
$$\mathbb{P} \left(\mathbf{E}_{k\alpha+l} = e_i e_j^\top \middle| x_{(k-1)\alpha+l} \right)$$
$$= \mathbb{P} \left((x_{k\alpha+l-1}, x_{k\alpha+l}) = (i, j) \middle| x_{(k-1)\alpha+l} \right)$$
$$= \left(e_{x_{(k-1)\alpha+l}}^\top \mathbf{P}^{\alpha-1} \right)_i \cdot \mathbf{P}_{ij} = (\tilde{\pi}_{(k,l)})_i \mathbf{P}_{ij}, \tag{82}$$

we can further calculate that

$$\mathbb{E} \left(\mathbf{E}_{k\alpha+l} \mathbf{E}_{k\alpha+l}^\top \middle| x_{(k-1)\alpha+l} \right)$$
$$= \sum_{i=1}^p \sum_{j=1}^p e_i e_i^\top (\tilde{\pi}_{(k,l)})_i \mathbf{P}_{ij} = \sum_{i=1}^p e_i e_i^\top (\tilde{\pi}_{(k,l)})_i$$
$$= \text{diag} \left(\tilde{\pi}_{(k,l)} \right) = \text{diag}(\pi) + \text{diag} \left(\tilde{\pi}_{(k,l)} - \pi \right)$$
$$\preceq \pi_{\max} \mathbf{I}_p + \left\| \tilde{\pi}_{(k,l)} - \pi \right\|_1 \cdot \mathbf{I}_p \preceq 2\pi_{\max} \mathbf{I}_p; \tag{83}$$

$$\mathbb{E} \left(\mathbf{E}_{k\alpha+l}^\top \mathbf{E}_{k\alpha+l} \middle| x_{(k-1)\alpha+l} \right) = \sum_{i=1}^p \sum_{j=1}^p e_j e_j^\top \{ (\tilde{\pi}_{(k,l)})_i \mathbf{P}_{ij} \}$$
$$= \sum_{i=1}^p \sum_{j=1}^p e_j e_j^\top \{ \pi_i \mathbf{P}_{ij} \} + \sum_{i=1}^p \sum_{j=1}^p e_j e_j^\top \{ ((\tilde{\pi}_{(k,l)})_i - \pi)_i \mathbf{P}_{ij} \}$$
$$\preceq \sum_{j=1}^p e_j e_j^\top \pi_j + \sum_{j=1}^p e_j e_j^\top \left\| \tilde{\pi}_{(k,l)} - \pi \right\|_1 \cdot \max_{ij} \mathbf{P}_{ij}$$
$$\text{(since } \pi^\top \mathbf{P} = \pi \text{)}$$
$$\preceq \pi_{\max} \mathbf{I}_p + \left\| \tilde{\pi}_{(k,l)} - \pi \right\|_1 \cdot \mathbf{I}_p \preceq 2\pi_{\max} \mathbf{I}_p. \tag{84}$$

Therefore,

$$0 \preceq \mathbb{E} \left(\tilde{\mathbf{E}}_k^{(l)} (\tilde{\mathbf{E}}_k^{(l)})^\top \middle| \tilde{\mathbf{E}}_{k-1}^{(l)} \right)$$
$$= \mathbb{E} \Bigg\{ \left(\mathbf{E}_{k\alpha+l} - \mathbb{E}(\mathbf{E}_{k\alpha+l} | x_{(k-1)\alpha+l}) \right)$$
$$\cdot \left(\mathbf{E}_{k\alpha+l} - \mathbb{E}(\mathbf{E}_{k\alpha+l} | x_{(k-1)\alpha+l}) \right)^\top \middle| x_{(k-1)\alpha+l} \Bigg\}$$
$$= \mathbb{E} \left\{ \mathbf{E}_{k\alpha+1} \mathbf{E}_{k\alpha+1}^\top \middle| x_{(k-1)\alpha+l} \right\}$$
$$- \mathbb{E} \left\{ \mathbf{E}_{k\alpha+l} \middle| x_{(k-1)\alpha+l} \right\} \mathbb{E} \left\{ \mathbf{E}_{k\alpha+l}^\top \middle| x_{(k-1)\alpha+l} \right\}$$
$$\preceq \mathbb{E} \left\{ \mathbf{E}_{k\alpha+1} \mathbf{E}_{k\alpha+1}^\top \middle| x_{(k-1)\alpha+l} \right\} \preceq 2\pi_{\max} \mathbf{I}_p. \tag{85}$$

Similarly,

$$0 \preceq \mathbb{E} \left((\tilde{\mathbf{E}}_k^{(l)})^\top \tilde{\mathbf{E}}_k^{(l)} \middle| \tilde{\mathbf{E}}_{k-1}^{(l)} \right)$$
$$= \mathbb{E} \Bigg\{ \left(\mathbf{E}_{k\alpha+l} - \mathbb{E}(\mathbf{E}_{k\alpha+l} | x_{(k-1)\alpha+l}) \right)^\top$$
$$\left(\mathbf{E}_{k\alpha+l} - \mathbb{E}(\mathbf{E}_{k\alpha+l} | x_{(k-1)\alpha+l}) \right) \middle| x_{(k-1)\alpha+l} \Bigg\}$$
$$= \mathbb{E} \left\{ \mathbf{E}_{k\alpha+1}^\top \mathbf{E}_{k\alpha+1} \middle| x_{(k-1)\alpha+l} \right\}$$
$$- \mathbb{E} \left\{ \mathbf{E}_{k\alpha+l}^\top \middle| x_{(k-1)\alpha+l} \right\} \mathbb{E} \left\{ \mathbf{E}_{k\alpha+l} \middle| x_{(k-1)\alpha+l} \right\}$$
$$\preceq \mathbb{E} \left\{ \mathbf{E}_{k\alpha+1}^\top \mathbf{E}_{k\alpha+1} \middle| x_{(k-1)\alpha+l} \right\} \preceq 2\pi_{\max} \mathbf{I}_p, \tag{86}$$

which means for $1 \leq k \leq n_0$, $1 \leq l \leq \alpha$,

$$\max \left\{ \left\| \mathbb{E} \left((\tilde{\mathbf{E}}_k^{(l)})^\top \tilde{\mathbf{E}}_k^{(l)} \middle| \tilde{\mathbf{E}}_{k-1}^{(l)} \right) \right\|, \right.$$
$$\left. \left\| \mathbb{E} \left(\tilde{\mathbf{E}}_k^{(l)} (\tilde{\mathbf{E}}_k^{(l)})^\top \middle| \tilde{\mathbf{E}}_{k-1}^{(l)} \right) \right\| \right\} \leq 2\pi_{\max} \mathbf{I}_p. \tag{87}$$

Next, the predictable quadratic variation process of the martingale $\left\{ \tilde{\mathbf{E}}_k^{(l)} \right\}_{k=1}^{n_0}$ satisfies

$$\left\| \sum_{k=1}^{n_0} \mathbb{E} \left(\tilde{\mathbf{E}}_k^{(l)} (\tilde{\mathbf{E}}_k^{(l)})^\top \middle| \tilde{\mathbf{E}}_{k-1}^{(l)} \right) \right\|$$
$$\leq \sum_{k=1}^{n_0} \left\| \mathbb{E} \left(\tilde{\mathbf{E}}_k^{(l)} (\tilde{\mathbf{E}}_k^{(l)})^\top \middle| \tilde{\mathbf{E}}_{k-1}^{(l)} \right) \right\| \leq 2n_0 \pi_{\max},$$
$$\left\| \sum_{k=1}^{n_0} \mathbb{E} \left((\tilde{\mathbf{E}}_k^{(l)})^\top \tilde{\mathbf{E}}_k^{(l)} \middle| \tilde{\mathbf{E}}_{k-1}^{(l)} \right) \right\|$$
$$\leq \sum_{k=1}^{n_0} \left\| \mathbb{E} \left((\tilde{\mathbf{E}}_k^{(l)})^\top \tilde{\mathbf{E}}_k^{(l)} \middle| \tilde{\mathbf{E}}_{k-1}^{(l)} \right) \right\| \leq 2n_0 \pi_{\max}.$$

Now by matrix Freedman's inequality (Corollary 1.3 in [73]), we know

$$\mathbb{P} \left(\left\| \frac{1}{n_0} \sum_{k=1}^{n_0} \tilde{\mathbf{E}}_k^{(l)} \right\| \geq t/2 \right) \leq 2p \exp \left(- \frac{(tn_0)^2/8}{2n_0 \pi_{\max} + tn_0/6} \right). \tag{88}$$

Here, $\|\cdot\|$ represents the matrix 2-norm. Next, we shall note that

$$\mathbb{E}\left(\mathbf{E}_{k\alpha+l}\Big|x_{(k-1)\alpha+l}\right) - \mathrm{diag}(\pi)\mathbf{P}$$

$$=\sum_{i=1}^{p}\sum_{j=1}^{p}e_i\left(e_{x_{(k-1)\alpha+l}}^{\top}\mathbf{P}^{\alpha-1}\right)_i \mathbf{P}_{ij}e_j^{\top} - \mathrm{diag}(\pi)\mathbf{P}$$

$$=\mathrm{diag}\left(e_{x_{(k-1)\alpha+l}}^{\top}\mathbf{P}^{\alpha-1}\right)\mathbf{P} - \mathrm{diag}(\pi)\mathbf{P},$$

thus

$$\left\|\mathbb{E}\left(\mathbf{E}_{k\alpha+l}\Big|x_{(k-1)\alpha+l}\right) - \mathrm{diag}(\pi)\mathbf{P}\right\| \le \left\|(\tilde{\pi}_{(k,l)} - \pi)\mathbf{P}\right\|$$

$$= \max_{\substack{u,v\in\mathbb{R}^p \\ \|u\|_2=\|v\|_2=1}} u^{\top}\mathrm{diag}(\tilde{\pi}_{(k,l)} - \pi)\mathbf{P}v$$

$$\le \max_{\substack{u,v\in\mathbb{R}^p \\ \|u\|_2=\|v\|_2=1}} \sum_{i=1}^{p}\left|u_i((\tilde{\pi}_{(k,l)})_i - \pi_i)\mathbf{P}_{ij}v_j\right|$$

$$\le \sum_{i=1}^{p}\sum_{j=1}^{p}\left|((\tilde{\pi}_{(k,l)})_i - \pi_i)\mathbf{P}_{ij}\right| \le \|\tilde{\pi}_{(k,l)} - \pi\|_1 \overset{(80)}{\le} t/2.$$
(89)

The last but one equality is due to $\sum_{j=1}^{p}|\mathbf{P}_{ij}| = \sum_{j=1}^{p}\mathbf{P}_{ij} = 1$ for all i. Combining (78), (88), and (89), we have for any $l = 1,\ldots,\alpha$,

$$\mathbb{P}\left(\left\|\frac{1}{n_0}\sum_{k=1}^{n_0}\mathbf{E}_{k\alpha+l}-\mathbf{F}\right\|\ge t\right)\le 2p\exp\left(-\frac{(tn_0)^2/8}{2n_0\pi_{\max}+tn_0/6}\right).$$
(90)

Finally, we only need to combine these "thin" summation sequences by using a union bound,

$$\mathbb{P}\left(\|\tilde{\mathbf{F}} - \mathbf{F}\| \ge t\right) = \mathbb{P}\left(\left\|\frac{1}{\alpha}\sum_{l=1}^{\alpha}\frac{1}{n_0}\sum_{k=1}^{n_0}\mathbf{E}_{k\alpha+l} - \mathbf{F}\right\| \ge t\right)$$

$$\le \mathbb{P}\left(\max_{1\le l\le\alpha}\left\|\sum_{k=1}^{n_0}\frac{1}{n_0}\mathbf{E}_{k\alpha+l} - \mathbf{F}\right\| \ge t\right)$$

$$\le \alpha\max_{1\le l\le p}\mathbb{P}\left(\left\|\sum_{k=1}^{n_0}\frac{1}{n_0}\mathbf{E}_{k\alpha+l} - \mathbf{F}\right\| \ge t\right)$$

$$\le 2\alpha p\exp\left(-\frac{(tn_0)^2/8}{2n_0\pi_{\max} + tn_0/6}\right),$$
(91)

which proves (69). Particularly by setting $t = C\sqrt{\frac{\pi_{\max}\tau(\sqrt{\pi_{\max}/n})\log(n)}{n}}$ for large constant C, one further obtains (7). When $\tau_* = \tau(1/4)$, Lemma 5 implies

$$\tau(\sqrt{\pi_{\max}/n}) \le C\tau_*\log(\sqrt{n/\pi_{\max}}) \le C\tau_*\log(\sqrt{np})$$
$$\le C\tau_*\log(n),$$

thus (73) immediately follows from (7).

When $\mathbf{P}$ is reversible and with second largest eigenvalue $\lambda_2 < 1$, Lemma 6 implies

$$\tau\left(\sqrt{\pi_{\max}/n}\right) \le \frac{1}{1-\lambda_2}\log\left(\frac{\sqrt{n/\pi_{\max}}}{2\pi_{\min}}\right)$$

$$\le \frac{C}{1-\lambda_2}\left(\log(n) + \log(1/\pi_{\min})\right) = \frac{C}{1-\lambda_2}\log(n/\pi_{\min}).$$

Then (7) follows from (7).

The proof for the upper bounds $\|\tilde{\pi} - \pi\|_\infty$ is similar. Recall the definition of $\tilde{e}_k^{(l)}$ in (77). Note that for any index $j \in \{1,\ldots,p\}$,

$$\left(\tilde{e}_k^{(l)}\right)_j = (e_{x_{k\alpha+l}})_j - \mathbb{E}\left((e_{x_{k\alpha+l}})_j\Big|x_{(k-1)\alpha+l}\right)$$

$$=1_{\{x_{k\alpha+l}=j\}} - \mathbb{E}\left(1_{\{x_{k\alpha+l}=j\}}\Big|x_{(k-1)\alpha+l}\right).$$

Clearly $0 \le \mathbb{E}\left(1_{\{x_{k\alpha+l}=j\}}\Big|x_{(k-1)\alpha+l}\right) \le 1$, which implies $\left|(\tilde{e}_k^{(l)})_j\right| \le 1$. Additionally,

$$\mathbb{E}\left(\tilde{e}_k^{(l)}\right)_j^2 = \mathrm{Var}\left(1_{\{x_{k\alpha+l}=j\}}\Big|x_{(k-1)\alpha+l}\right) \le \mathbb{E}\left(1_{\{x_{k\alpha+l}=j\}}^2\right)$$

$$= \left(e_{x_{(k-1)\alpha+l}}^{\top}\mathbf{P}^{\alpha}\right)_j$$

$$\le \pi_j + \left(e_{x_{(k-1)\alpha+l}}^{\top}\mathbf{P}^{\alpha} - \pi^{\top}\right)_j \le 2\pi_{\max}.$$

By Freedman's inequality (e.g. Theorem 1.6 in [74] and Theorem 1.1 in [73]), for any $1 \le j \le p$,

$$\mathbb{P}\left(\left|\sum_{k=1}^{n_0}(\tilde{e}_k^{(l)})_j\right| \ge t/2\right) \le 2\exp\left(\frac{-t^2/8}{2n_0\pi_{\max} + t/6}\right)$$

On the other hand,

$$\left\|\mathbb{E}\left(e_{x_{k\alpha+l}}\Big|e_{x_{(k-1)\alpha+l}}\right) - \pi\right\|_\infty = \left\|e_{x_{(k-1)\alpha+l}}^{\top}\mathbf{P}^{\alpha} - \pi^{\top}\right\|_\infty$$

$$\le \left\|e_{x_{(k-1)\alpha+l}}^{\top}\mathbf{P}^{\alpha} - \pi^{\top}\right\|_1 \le \frac{t}{2}\wedge\pi_{\max}.$$

Combining the two inequality above and the definition (77), we have for any $1 \le j \le p, 1 \le l \le \alpha$,

$$\mathbb{P}\left(\left|\sum_{k=1}^{n_0}(e_{k\alpha+l})_j - n_0\pi_j\right| \ge t\right) \le 2\exp\left(\frac{-t^2/8}{2n_0\pi_{\max}+t/6}\right).$$
(92)

By a union bound, one can show

$$\mathbb{P}\left(\|\tilde{\pi} - \pi\|_\infty \ge t\right) = \mathbb{P}\left(\left\|\frac{1}{\alpha}\sum_{l=1}^{\alpha}\frac{1}{n_0}\sum_{k=1}^{n_0}e_{x_{k\alpha+l}} - \pi\right\|_\infty \ge t\right)$$

$$\le \mathbb{P}\left(\max_{1\le l\le\alpha}\max_{1\le j\le p}\left|\sum_{k=1}^{n_0}\frac{1}{n_0}(e_{x_{k\alpha+l}})_j - \pi_j\right| \ge t\right)$$

$$\le 2\alpha p\exp\left(\frac{-(tn_0)^2/8}{2n_0\pi_{\max} + tn_0/6}\right),$$
(93)

which has developed the upper bound for $\|\tilde{\pi} - \pi\|_\infty$ (70). Finally, the proofs of (7), (74), and (7) are essentially follows from the previous argument for $\|\tilde{\mathbf{F}} - \mathbf{F}\|$. $\square$

Lemma 8 (Rowwise Markov Concentration Inequality). *Suppose $\mathbf{V} \in \mathbb{O}_{p,r}$ is a fixed orthogonal matrices satisfying $\max_i \|\mathbf{V}^{\top}e_i\|_2 \le \delta\sqrt{r/p}$. Assume $n \ge Cp\tau_*\log^2(n)$ and $\tau_* := \tau(1/4)$. Under the same setting as Lemma 7, for any $c_0 > 0$ there exists $C > 0$,*

$$\max_{1\le i\le p}\|(\tilde{\mathbf{F}}\mathbf{V})_{i\cdot} - (\mathbf{F}\mathbf{V})_{i\cdot}\|_2 \le C\sqrt{\frac{\pi_{\max}\delta^2\,r\tau_*\log^2(n)}{np}}$$

with probability at least $1 - Cn^{-c_0}$.

Proof of Lemma 8. We first focus on the s-th row of $\|\tilde{\mathbf{F}}\mathbf{V}_{i\cdot} - \mathbf{F}\mathbf{V}_{i\cdot}\|_2$. Similarly as the proof of Lemma 7, let $\alpha = \tau\left(\min\{t/2, \pi_{\max}\}\right) + 1$, $\mathbf{E}_k = e_{x_k}e_{x_{k+1}}^\top$, $k = 1, \ldots, n$. t is to be determined later. We similarly assume n is a multiple of α and define $n_0 = n/\alpha$. We further define

$$\mathbf{T}_k \in \mathbb{R}^{1\times r}, \quad \mathbf{T}_k = e_s^\top \mathbf{E}_k \mathbf{V}, \quad k = 1, \ldots, n;$$

and the "thin" matrix sequences for $l = 1, \ldots, \alpha$, $k = 1, \ldots, n_0$,

$$\begin{aligned}
\tilde{\mathbf{T}}_k^{(l)} &= \mathbf{T}_{k\alpha+l} - \mathbb{E}\left(\mathbf{T}_{k\alpha+l}|\mathbf{T}_{(k-1)\alpha+l}\right) \\
&= e_s^\top \mathbf{E}_{k\alpha+l}\mathbf{V} - \mathbb{E}\left(e_s^\top \mathbf{E}_{k\alpha+l}\mathbf{V}\Big|x_{(k-1)\alpha+l}\right).
\end{aligned}$$

Then $\mathbf{T}_{k\alpha+l}$ and $\tilde{\mathbf{T}}_k^{(l)}$ satisfy the following 2-norm upper bound

$$\|\mathbf{T}_{k\alpha+l}\| = \max_{1\le i,j\le p}\|e_s^\top e_i e_j^\top \mathbf{V}\| = \max_j \|e_j^\top \mathbf{V}\|_2 \le \delta\sqrt{r/p}.$$

By Jensen's inequality, $\|\mathbb{E}(\mathbf{T}_{k\alpha+l}|\mathbf{T}_{(k-1)\alpha+l})\| \le \delta\sqrt{r/p}$, thus

$$\|\tilde{\mathbf{T}}_k^{(l)}\| \le 2\delta\sqrt{r/p} \quad \text{almost surely.}$$

Next, we define $\tilde{\pi}_{(k,l)} = \left(e_{x_{(k-1)\alpha+l}}^\top \mathbf{P}^{\alpha-1}\right)^\top \in \mathbb{R}^p$. By the choice of α and the mixing time property,

$$\|\tilde{\pi}_{(k,l)} - \pi\|_1 = \left\|e_{x_{(k-1)\alpha+l}}^\top \mathbf{P}^{\alpha-1} - \pi\right\|_1 \le \min\{t/2, \pi_{\max}\}.$$

Then,

$$\begin{aligned}
&\mathbb{E}\left(\mathbf{T}_{k\alpha+l}\mathbf{T}_{k\alpha+l}^\top\Big|x_{(k-1)\alpha+l}\right) \\
&= \sum_{i,j=1}^p e_s^\top e_i e_j^\top \mathbf{V}\mathbf{V}^\top e_j e_i^\top e_s (\tilde{\pi}_{(k,l)})_i \mathbf{P}_{ij} \\
&= \sum_{i,j=1}^p e_s^\top e_i e_i^\top e_s (\tilde{\pi}_{(k,l)})_i \mathbf{P}_{ij}\|\mathbf{V}^\top e_j\|_2^2 \\
&\le \sum_{i,j=1}^p e_s^\top e_i e_i^\top e_s (\tilde{\pi}_{(k,l)})_i \mathbf{P}_{ij}\cdot \delta r/p \\
&= \sum_{i=1}^p e_s^\top e_i e_i^\top e_s (\tilde{\pi}_{(k,l)})_i \delta^2\ r/p \\
&\le \delta^2(r/p)\max_i(\tilde{\pi}_{(k,l)})_i \cdot e_s^\top \sum_{i=1}^p e_i e_i^\top e_s \\
&\le \delta^2(r/p)\max_i(\tilde{\pi}_{(k,l)})_i \le \delta^2(r/p)\left(\pi_{\max} + \|\pi - \tilde{\pi}_{(k,l)}\|_1\right) \\
&\le 2\pi_{\max}\delta^2(r/p),
\end{aligned}$$

By Jensen's inequality,

$$\begin{aligned}
&\left\|\mathbb{E}\left(\mathbf{T}_{k\alpha+l}^\top \mathbf{T}_{k\alpha+l}\Big|x_{(k-1)\alpha+l}\right)\right\| \\
&\le \mathbb{E}\left(\left\|\mathbf{T}_{k\alpha+l}^\top \mathbf{T}_{k\alpha+l}\right\|\Big|x_{(k-1)\alpha+l}\right) \\
&\le \mathbb{E}\left(\mathbf{T}_{k\alpha+l}\mathbf{T}_{k\alpha+l}^\top\Big|x_{(k-1)\alpha+l}\right) \le 2\pi_{\max}\delta^2(r/p).
\end{aligned}$$

Similarly as (85) and (86) in the proof of Lemma 7, we can show

$$0 \le \mathbb{E}\left(\tilde{\mathbf{T}}_k^{(l)}(\tilde{\mathbf{T}}_k^{(l)})^\top\Big|\tilde{\mathbf{T}}_k^{(l)}\right) \le 2\pi_{\max}\delta^2 r/p,$$

$$0 \preceq \mathbb{E}\left((\tilde{\mathbf{T}}_k^{(l)})^\top\tilde{\mathbf{T}}_k^{(l)}\Big|\tilde{\mathbf{T}}_k^{(l)}\right) \preceq 2\pi_{\max}\delta^2(r/p)\mathbf{I}_r.$$

Then the predictable quadratic variation process satisfies

$$\max\left\{\left\|\sum_{k=1}^{n_0}\mathbb{E}\left(\tilde{\mathbf{T}}_k^{(l)}(\tilde{\mathbf{T}}_k^{(l)})^\top\Big|\tilde{\mathbf{T}}_k^{(l)}\right)\right\|,\right.$$
$$\left.\left\|\sum_{k=1}^{n_0}\mathbb{E}(\tilde{\mathbf{T}}_k^{(l)})^\top\left(\tilde{\mathbf{T}}_k^{(l)}\Big|\tilde{\mathbf{T}}_k^{(l)}\right)\right\|\right\} \le 2n_0\pi_{\max}\delta^2 r/p.$$

By the Freedman's inequality (Corollary 1.3 in [73]),

$$P\left(\left\|\sum_{k=1}^{n_0}\tilde{\mathbf{T}}_k^{(l)}\right\| \ge t/2\right)$$
$$\le (r+1)\exp\left(-\frac{(tn_0)^2/8}{2n_0\pi_{\max}\delta^2 r/p + tn_0\delta\sqrt{r/p}/3}\right).$$

Next, similarly as (89), (90), and (91) in the proof of Lemma 7, we can show

$$\left\|\mathbb{E}\left(\mathbf{T}_{k\alpha+l}\Big|x_{(k-1)\alpha+l}\right) - \text{diag}(\pi)\mathbf{P}\right\| \le t/2,$$

$$\mathbb{P}\left(\left\|\frac{1}{n_0}\sum_{k=1}^{n_0}\mathbf{T}_{k\alpha+l} - e_s^\top \mathbf{F}\mathbf{V}\right\| \ge t\right)$$
$$\le (r+1)\exp\left(-\frac{(tn_0)^2/8}{2n_0\pi_{\max}\delta^2 r/p + tn_0\delta\sqrt{r/p}/3}\right),$$

and

$$\mathbb{P}\left(\left\|e_s^\top\tilde{\mathbf{F}}\mathbf{V} - e_s^\top\mathbf{F}\mathbf{V}\right\| \ge t\right)$$
$$\le \alpha(r+1)\exp\left(-\frac{(tn_0)^2/8}{2n_0\pi_{\max}\delta^2 r/p + tn_0\delta\sqrt{r/p}/3}\right)$$
$$\le \alpha(r+1)\exp\left(-\frac{t^2 n/(8\alpha)}{2\pi_{\max}\delta^2 r/p + t\delta\sqrt{r/p}/3}\right).$$

By Lemma 5, $\alpha = \tau\left(\min(t/2, \pi_{\max})\right) + 1 \le C\tau_*\log(1/(t \wedge \pi_{\max}))$. Next, for any $c_0 > 0$, we set

$$t = C\sqrt{\frac{\pi_{\max}\delta^2 r\tau_*\log^2(n)}{np}} + C\frac{\delta\sqrt{r/p}\cdot\tau_*\log^2(n)}{n}$$

for large constant $C > 0$. By $n \ge Cp\tau_*\log^2(n) \ge Cr$ and $\pi_{\max} \ge 1/p$, we have

$$\mathbb{P}\left(\left\|e_s^\top\tilde{\mathbf{F}}\mathbf{V} - e_s^\top\mathbf{F}\mathbf{V}\right\| \ge C\sqrt{\frac{\pi_{\max}\delta^2 r\tau_*\log^2(n)}{np}}\right.$$
$$\left. + C\frac{\delta\sqrt{r/p}\cdot\tau_*\log^2(n)}{n}\right) \le n^{-c_0-1}$$

and

$$\max_i\|(\tilde{\mathbf{F}}\mathbf{V})_{i\cdot} - (\mathbf{F}\mathbf{V})_{i\cdot}\|_2 = \max_{1\le s\le p}\left\|e_s^\top\tilde{\mathbf{F}}\mathbf{V} - e_s^\top\mathbf{F}\mathbf{V}\right\|$$
$$\le C\sqrt{\frac{\pi_{\max}\delta^2 r\tau_*\log^2(n)}{np}} + C\frac{\delta\sqrt{r/p}\cdot\tau_*\log^2(n)}{n}$$
$$\le C\sqrt{\frac{\pi_{\max}\delta^2 r\tau_*\log^2(n)}{np}}$$

with probability at least $1 - n^{-c_0-1}p \ge 1 - n^{-c_0}$. $\quad\square$

References

[1] Y. Liu, C. Kang, Y. Xiao, Y. Tian, and S. Gao, "Understanding intra-urban trip patterns from taxi trajectory data," *J. Geogr. Syst.*, vol. 14, no. 4, pp. 463–483, 2012.

[2] A. R. Benson, D. F. Gleich, and L.-H. Lim, "The spacey random walk: A stochastic process for higher-order data," *SIAM Rev.*, vol. 59, no. 2, pp. 321–345, 2017.

[3] L. F. Yang, V. Braverman, T. Zhao, and M. Wang, "Dynamic partition of complex networks," 2017, *arXiv:1705.07881*. [Online]. Available: https://arxiv.org/abs/1705.07881

[4] S. Negahban, S. Oh, and D. Shah, "Rank centrality: Ranking from pairwise comparisons," *Oper. Res.*, vol. 65, no. 1, pp. 266–287, 2016.

[5] Y. Chen, J. Fan, C. Ma, and K. Wang, "Spectral method and regularized MLE are both optimal for top-K ranking," 2017, *arXiv:1707.09971*. [Online]. Available: https://arxiv.org/abs/1707.09971

[6] D. F. Rogers, R. D. Plante, R. T. Wong, and J. R. Evans, "Aggregation and disaggregation techniques and methodology in optimization," *Oper. Res.*, vol. 39, no. 4, pp. 553–582, 1991.

[7] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*. Belmont, MA, USA: Athena Scientific, 1996.

[8] J. N. Tsitsiklis and B. Van Roy, "Feature-based methods for large scale dynamic programming," *Mach. Learn.*, vol. 22, no. 1, pp. 59–94, 1996.

[9] S. P. Singh, T. Jaakkola, and M. I. Jordan, "Reinforcement learning with soft state aggregation," in *Proc. Adv. Neural Inf. Process. Syst.*, 1995, pp. 361–368.

[10] L. Molgedey and H. G. Schuster, "Separation of a mixture of independent signals using time delayed correlations," *Phys. Rev. Lett.*, vol. 72, no. 23, p. 3634, 1994.

[11] G. Pérez-Hernández, F. Paul, T. Giorgino, G. De Fabritiis, and F. Noé, "Identification of slow molecular order parameters for Markov model construction," *J. Chem. Phys.*, vol. 139, no. 1, 2013, Art. no. 015102.

[12] P. J. Schmid, "Dynamic mode decomposition of numerical and experimental data," *J. Fluid Mech.*, vol. 656, pp. 5–28, Jan. 2010.

[13] K. K. Chen, J. H. Tu, and C. W. Rowley, "Variants of dynamic mode decomposition: Boundary condition, Koopman, and Fourier analyses," *J. Nonlinear Sci.*, vol. 22, no. 6, pp. 887–915, Mar. 2012.

[14] S. Klus F. Nüske, P. Koltai, H. Wu, I. Kevrekidis, C. Schütte, and F. Noé, "Data-driven model reduction and transfer operator approximation," *J. Nonlinear Sci.*, vol. 28, no. 3, pp. 985–1010, 2018.

[15] A. W. Moore, "Variable resolution dynamic programming: Efficiently learning action maps in multivariate real-valued state-spaces," in *Machine Learning Proceedings 1991*. Amsterdam, The Netherlands: Elsevier, 1991, pp. 333–337.

[16] Z. Ren and B. H. Krogh, "State aggregation in Markov decision processes," in *Proc. 41st IEEE Conf. Decis. Control*, vol. 4, Dec. 2002, Aug. pp. 3819–3824.

[17] J. Johns and S. Mahadevan, "Constructing basis functions from directed graphs for value function approximation," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 385–392.

[18] S. Mahadevan, "Proto-value functions: Developmental reinforcement learning," in *Proc. 22nd Int. Conf. Mach. Learn.*, 2005, pp. 553–560.

[19] R. Parr, C. Painter-Wakefield, L. Li, and M. Littman, "Analyzing feature generation for value-function approximation," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 737–744.

[20] M. Petrik, "An analysis of Laplacian methods for value function approximation in MDPs," in *Proc. IJCAI*, 2007, pp. 2574–2579.

[21] S. Mahadevan, "Learning representation and control in Markov decision processes: New frontiers," *Found. Trends Mach. Learn.*, vol. 1, no. 4, pp. 403–565, 2008.

[22] H. Steinhaus, "The problem of estimation," *Ann. Math. Statist.*, vol. 28, no. 3, pp. 633–648, 1957.

[23] M. Wilczyński, "Minimax estimation for the multinomial and multivariate hypergeometric distributions," *Sankhya, Indian J. Statist., A*, vol. 47, no. 1, pp. 128–132, 1985.

[24] E. L. Lehmann and G. Casella, *Theory of Point Estimation*. Berlin, Germany: Springer, 2006.

[25] Y. Han, J. Jiao, and T. Weissman, "Minimax estimation of discrete distributions under ℓ_1 loss," *IEEE Trans. Inf. Theory*, vol. 61, no. 11, pp. 6343–6354, Nov. 2015.

[26] S. Kamath, A. Orlitsky, D. Pichapati, and A. T. Suresh, "On learning distributions from their samples," in *Proc. Conf. Learn. Theory*, 2015, pp. 1066–1100.

[27] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, p. 717, Dec. 2009.

[28] B. Recht, "A simpler approach to matrix completion," *J. Mach. Learn. Res.*, vol. 12, pp. 3413–3430, Jan. 2011.

[29] K. C. Toh and S. Yun, "An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems," *Pacific J. Optim.*, vol. 6, nos. 615–640, p. 15, 2010.

[30] Y. Chen and M. J. Wainwright, "Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees," 2015, *arXiv:1509.03025*. [Online]. Available: https://arxiv.org/abs/1509.03025

[31] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, 2010.

[32] S. Chatterjee, "Matrix estimation by universal singular value thresholding," *Ann. Statist.*, vol. 43, no. 1, pp. 177–214, 2015.

[33] J. D. Lee, B. Recht, N. Srebro, J. Tropp, and R. R. Salakhutdinov, "Practical large-scale optimization for max-norm regularization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1297–1305.

[34] T. Cai and W.-X. Zhou, "A max-norm constrained minimization approach to 1-bit matrix completion," *J. Mach. Learn. Res.*, vol. 14, pp. 3619–3647, Dec. 2013.

[35] D. Hsu, S. M. Kakade, and T. Zhang, "A spectral algorithm for learning hidden Markov models," *J. Comput. Syst. Sci.*, vol. 78, no. 5, pp. 1460–1480, 2012.

[36] Q. Huang, S. M. Kakade, W. Kong, and G. Valiant, "Recovering structured probability matrices," 2016, *arXiv:1602.06586*. [Online]. Available: https://arxiv.org/abs/1602.06586

[37] X. Li, M. Wang, and A. Zhang, "Estimation of Markov chain via rank-constrained likelihood," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1–16.

[38] M. Meila and J. Shi, "A random walks view of spectral segmentation," Tech. Rep., 2001.

[39] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. Adv. Neural Inf. Process. Syst.*, 2002, pp. 849–856.

[40] K. Rohe, S. Chatterjee, and B. Yu, "Spectral clustering and the high-dimensional stochastic blockmodel," *Ann. Statist.*, vol. 39, no. 4, pp. 1878–1915, 2011.

[41] M. E. J. Newman, "Spectral methods for community detection and graph partitioning," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 88, no. 4, 2013, Art. no. 042822.

[42] J. Lei *et al.*, "Consistency of spectral clustering in stochastic block models," *Ann. Statist.*, vol. 43, no. 1, pp. 215–237, 2015.

[43] J. Jin and W. Wang, "Influential features PCA for high dimensional clustering," *Ann. Statist.*, vol. 44, no. 6, pp. 2323–2359, 2016.

[44] T. T. Cai and A. Zhang, "Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics," *Ann. Statist.*, vol. 46, no. 1, pp. 60–89, 2018.

[45] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

[46] S. Zeng, R. Huang, Z. Kang, and N. Sang, "Image segmentation using spectral clustering of Gaussian mixture models," *Neurocomputing*, vol. 144, pp. 346–356, Nov. 2014.

[47] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from a few entries," *IEEE Trans. Inf. Theory*, vol. 56, no. 6, pp. 2980–2998, Jun. 2010.

[48] E. Weinan, T. Li, and E. Vanden-Eijnden, "Optimal partition and effective dynamics of complex networks," *Proc. Nat. Acad. Sci. USA*, vol. 105, no. 23, pp. 7907–7912, 2008.

[49] Y. Duan, Z. T. Ke, and M. Wang, "State aggregation learning from Markov transition data," in *Proc. Neural Inf. Process. Syst. (NeurIPS)*, 2019.

[50] Y. Sun, Y. Duan, H. Gong, and M. Wang, "Learning low-dimensional state embeddings and metastable clusters from time series data," in *Proc. Neural Inf. Process. Syst. (NeurIPS)*, 2019.

[51] J. R. Norris, *Markov Chains*, no. 2. Cambridge, U.K.: Cambridge Univ. Press, 1998.

[52] D. A. Levin, Y. Peres, and E. L. Wilmer, *Markov Chains and Mixing Times*. Providence, RI, USA: AMS, 2009.

[53] J. T. Runnenburg, "Markov processes in waiting-time and renewal theory," Ph.D. dissertation, Amsterdam, The Netherlands, 1966.

[54] A. Hoekstra, "Markov chains of finite rank," Ph.D. dissertation, Math. Centre, Amsterdam, The Netherlands, 1983.

[55] A. H. Hoekstra and F. W. Steutel, "Limit theorems for Markov chains of finite rank," *Linear Algebra Appl.*, vol. 60, pp. 65–77, Aug. 1984.

[56] D. P. Bertsekas, *Dynamic Programming and Optimal Control*. Belmont, MA, USA: Athena Scientific, 2007.

[57] A. George, W. B. Powell, and S. R. Kulkarni, "Value function approximation using multiple aggregation for multiattribute resource management," *J. Mach. Learn. Res.*, vol. 9, pp. 2079–2111, Oct. 2008.

[58] J. G. Kemeny and J. L. Snell, *Finite Markov Chains*, vol. 356. Princeton, NJ, USA: David Van Nostrand, 1960.

[59] P. Buchholz, "Exact and ordinary lumpability in finite Markov chains," *J. Appl. Probab.*, vol. 31, no. 1, pp. 59–75, 1994.

[60] T. W. Anderson and L. A. Goodman, "Statistical inference about Markov chains," *Ann. Math. Statist.*, vol. 28, no. 1, pp. 89–110, 1957.

[61] W. Wang and M. Á. Carreira-Perpiñán, "Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application," 2013, *arXiv:1309.1541*. [Online]. Available: https://arxiv.org/abs/1309.1541

[62] E. J. Candès, C. A. Sing-Long, and J. D. Trzasko, "Unbiased risk estimates for singular value thresholding and spectral estimators," *IEEE Trans. Signal Process.*, vol. 61, no. 19, pp. 4643–4657, Oct. 2012.

[63] D. Donoho and M. Gavish, "Minimax risk of matrix denoising by singular value thresholding," *Ann. Statist.*, vol. 42, no. 6, pp. 2413–2440, 2014.

[64] I. Jolliffe, *Principal Component Analysis*, 2nd ed. New York, NY, USA: Springer, 2002.

[65] D. J. Hsu, A. Kontorovich, and C. Szepesvári, "Mixing time estimation in reversible Markov chains from a single sample path," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1459–1467.

[66] B. Jiang, Q. Sun, and J. Fan, "Bernstein's inequality for general Markov chains," 2018, *arXiv:1805.10721*. [Online]. Available: https://arxiv.org/abs/1805.10721

[67] D. Paulin, "Concentration inequalities for Markov chains by Marton couplings and spectral methods," *Electron. J. Probab.*, vol. 20, no. 79, p. 32, 2015.

[68] I. Jolliffe, *Principal Component Analysis*. New York, NY, USA: Springer-Verlag, 1986.

[69] B. Yu, "Assouad, Fano, and Le Cam," in *Festschrift for Lucien Le Cam*. New York, NY, USA: Springer, 1997, pp. 423–435.

[70] Y. Yang and A. Barron, "Information-theoretic determination of minimax rates of convergence," *Ann. Statist.*, vol. 27, no. 5, pp. 1564–1599, 1999.

[71] P.-Å. Wedin, "Perturbation bounds in connection with singular value decomposition," *BIT Numer. Math.*, vol. 12, no. 1, pp. 99–111, 1972.

[72] H. Weyl, "Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung)," *Math. Ann.*, vol. 71, no. 4, pp. 441–479, 1912.

[73] J. A. Tropp, "Freedman's inequality for matrix martingales," *Electron. Commun. Probab.*, vol. 16, pp. 262–270, Jan. 2011.

[74] D. A. Freedman, "On tail probabilities for martingales," *Ann. Probab.*, vol. 3, no. 1, pp. 100–118, 1975.

Anru Zhang received the B.S. degree from Peking University, Beijing, China, in 2010, and the Ph.D. degree from the University of Pennsylvania, Philadelphia, PA, USA, in 2015. He is currently an Assistant Professor of Statistics with the University of Wisconsin-Madison, Madison, WI, USA. His current research interests include high-dimensional statistical inference, tensor data analysis, statistical learning theory, dimension reduction, and convex/non-convex optimization.

Mengdi Wang received the B.S. degree in automation from Tsinghua University, Beijing, China, in 2007, and the Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology in 2013. She is currently an Associate Professor with the Department of Electrical Engineering and the Center for Statistics and Machine Learning, Princeton University. Her research interests include information-theoretic reinforcement learning, stochastic optimization, statistical dimension reduction, and system identification.