State Compression of Markov Processes via Empirical Low-Rank Estimation*

Anru Zhang and Mengdi Wang

(February 9, 2018)

Abstract

Model reduction is a central problem in analyzing complex systems and high-dimensional data. We study the state compression of finite-state Markov process from its empirical trajectories. We adopt a low-rank model which is motivated by the state aggregation of controlled systems. A spectral method is proposed for estimating the frequency and transition matrices, estimating the compressed state spaces, and recovering the state aggregation structure if there is any. We provide upper bounds for the estimation and recovery errors and matching minimax lower bounds.

1 Introduction

Dimension reduction is a central problem in system engineering and data science. In scientific studies or engineering applications, one often needs to interact with unknown complex systems about which many noisy observations of system characteristics and system trajectories are available. The exact structures and dynamics of the system are typically masked by massive observations of noisy variables, many of which might not be relevant to the physical state of the system. It is often unclear how to describe the "state" of a system, when one can only access noisy observations. One may view each unique observation as a single state, however, this would generate a huge- or even infinite-dimensional process which is difficult to model or analyze. Although there exists a vast body of literatures on time series analysis [18], they typically require knowledge of specific models and might perform poorly when the models are misspecified.

*Anru Zhang is Assistant Professor, Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706, E-mail: anruzhang@stat.wisc.edu; Mengdi Wang is Assistant Professor, Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, E-mail: mengdiw@princeton.edu.

Let us focus on the model reduction of discrete-state Markov processes where the transition function is not known. Suppose that we are given the trajectoric data generated by a black-box Markov process. The goal is to "sketch" the unknown dynamics from the limited data as well as to "compress" the state space into compact representations. We are inspired by the state aggregation approach for reducing controlled systems, which corresponds to a special decomposition of the system's transition function. In this paper, we consider a more general framework for state compression of Markov processes based on low-rank models. There are a few basic questions of interest: How to compress a large state space into compact representations? How to find an approximate state-aggregation structure based on empirical data? How to quantify the estimation error? In this paper, we plan to take a substantial step towards answering these questions.

1.1 Motivating Examples

State compression finds wide applications in analyzing random dynamic processes. Examples of application include network analysis, community detection, reinforcement learning and ranking problems.

- *Network Analysis and Community Detection.* Suppose that the network is hidden and one needs to learn from the dynamic "state-transition" data, which are snapshots of a random walk associated with the implicit network. For example, records of taxi trips can be used to reveal the traffic network of a metropolitan [26, 3]. Each trip can be viewed as a fragmented sample path realized from a city-wide Markov chain that characterizes the traffic dynamics. Experiments suggest that one can use state compression to recover the zoning of Manhattan city [52]. Existing results for network partition do not address the Markov nature of state-transition data.

- *Reinforcement Learning.* Reinforcement learning applications such as autonomous driving and game AI are modeled as Markov decision processes [46]. Given trajectories of game snapshots by an expert player, it is of vital interest to identify the compressed representation of the "state" of game. Multiple efforts show that as long as a reduced model is given, on can solve the reinforcement learning in sample-efficient ways; see [44] for reinforcement learning with known soft aggregation models.

- *Ranking Problems.* Learning to rank is a basic problem in machine learning, where the goal is to reconstruct a rank-ordered list from training data that consist of partial-ordered lists [32, 13]. Such problems typically arise from e-commerce applications where one needs to analyze click streams. The click stream can be viewed as a random walk on the space of all possible clicks, governed by the advertising strategy the user's preference. State compression of click stream might provide insights into the consumers' preferences.

1.2 Related Work

This work is related to previous literatures on spectral clustering, latent-variable models, and state aggregation of controlled system. Network data arise in many applications and research areas. Examples include protein-protein interaction networks [22], phone communication networks [34], collaboration networks [6], correlation networks between stock prices and the gravitational interaction network of dark matter particles in cosmology [38, 27, 33]. Due to the highly complex nature of these networks, many efforts have been devoted to investigate reduced-order representations from high-dimensional data (e.g. [14, 39, 37, 12]). Previously, [45, 51, 20] considered the minimax-optimal estimation under various losses and specific discrete distributions with i.i.d. observations. In contrast, the estimation of Markov frequency matrix is a discrete distribution estimation based on non-i.i.d. random walks and low-rank structures. The spectral clustering is a class of powerful methods that exploits the spectrum structure, reduce the dimension, and perform clustering for the high-dimensional data [29, 36]. This method is widely used in various problems in statistics and machine learning, such as community detection [41, 35, 24], high-dimensional feature clustering [21, 8], imaging segmentation [43, 55]. For most of spectral clustering literature, the data are independent but not Markovian generated. Most of the existing statistical results on low-rank matrix estimation focus on independent and identically distributed data.

There is a large body of literatures on the model reduction of large-scale systems. One popular way to model complicated systems in reduced dimensions is the so-called *state aggregation* [4, 5, 31, 40]. Admitting an inherent state aggregation structure means that the large collection of states can be mapped into a small number of state clusters without affecting the system dynamics. In practical engineering systems, state aggregation is usually heuristically imposed by practitioners based on domain-specific knowledges.

1.3 Scope of This Paper

In this paper, we first study the discrete-state Markov processes where the state space is finite and known *a priori*. Supposing that the transition probability is unknown, we aim to estimate a compressed model of the transition matrix from empirical trajectories. We propose a spectral state compression method for state space reduction of the Markov processes, which is based on truncated singular value decomposition. Then we extend the analysis for continuous-state Markov processes. Our main results are as follows.

1. We establish upper bounds on the estimation error for the frequency and transition

matrices, and further establish matching minimax lower bounds for a large class of Markov processes. Our method and result also extends to the estimation of general stochastic matrices that are not necessarily square.

2. We show that our method recovers the leading low-dimensional subspace for Markov processes, which we refer to as the compressed state space, with high accuracy. We provide upper bounds and minimax lower bounds for the subspace recovery error.

3. In the special case where the Markov process admits a latent state-aggregation structure. We show that the state compression method applies to recovering the state clusters with high probability.

4. We also consider the continuous-state Markov process low-rank kernel estimation. Particularly, a functional truncated singular value decomposition is introduced based on local smoothed empirical surface. The estimation upper bounds are further established.

We provide a general framework for dimension reduction of Markov processes, which applies to special cases like latent-variable models and lumpability of Markov models. The state compression method can be further used as basis functions for approximating distributions or for recovering state aggregation structures.

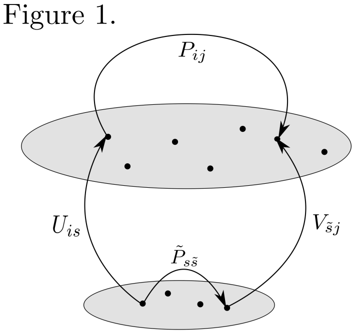
Outline Section 2 proposes the spectral method for low-rank approximation of Markov chains and provide recovery guarantees. Section 3 investigates the factorization approach to identify state aggregation structures for lumpable Markov chains. Section 4 extends the low-rank estimation results to continuous-state Markov processes and gives upper bound for the estimation error.

Notation and Preliminaries Let small case letters, such as x, y, z, to represent scalars and vectors. For $x, y \in \mathbb{R}$, we note $x \wedge y$ and $x \vee y$ as the maximum and the minimum of x and y, respectively. Especially, $(x)_+ = x \vee 0 = \max\{x, 0\}$ represents the non-negative part of x. For vector $v \in \mathbb{R}^p$, define its ℓ_q norm as $\|u\|_q = (\sum_{i=1}^p |v_i|^q)^{1/q}$, particularly $\|u\|_1 = \sum_{i=1}^p |u_i|$, $\|u\|_2 = (\sum_{i=1}^p u_i^2)^{1/2}$, and $\|u\|_\infty = \max_{1\le i\le p} |u_i|$ will be extensively used throughout the paper. We use boldface upper case letters, e.g. $\mathbf{F}$, $\mathbf{P}$, to represent matrices. For $\mathbf{X} \in \mathbb{R}^{p_1\times p_2}$ with singular value decomposition $\mathbf{X} = \sum_{k=1}^{p_1\wedge p_2} \sigma_k u_k v_k^\top$, denote $\sigma_k(\mathbf{X})$ as the k-th largest singular value of $\mathbf{X}$. Several matrix norms will be considered in this paper, including Frobenius norm $\|\mathbf{X}\|_F = \left(\sum_{i,j} \mathbf{X}_{ij}^2\right)^{1/2}$ and spectral norm $\|\mathbf{X}\| = \sup_{\|u\|_2\le 1} \|\mathbf{X}u\|_2$. The class of p-by-r orthogonal matrices is denoted as $\mathbb{O}_{p,r} = \{\mathbf{U} \in \mathbb{R}^{p\times r} : \mathbf{U}^\top\mathbf{U} = \mathbf{I}_r\}$. Finally, we use $C, C_0, C_1, \dots$ and $c, c_0, c_1, \dots$ to present the large and small constants respectively, whose actual values may vary from line to line.

Next, we briefly review the basic concepts of Markov process. Given a discrete Markov process on p states $\{1,\ldots,p\}$, denote its transition matrix as $\mathbf{P}\in\mathbb{R}^{p\times p}$. Given n states $\{X_0,\ldots,X_n\}$ from $\mathbf{P}$, one must have $\mathbb{P}(X_k=j|X_{k-1}=i,X_{k-2},\ldots,X_0)=\mathbf{P}_{ij}$ for all $k\geq 1$, $1\leq i,j\leq p$. When the Markov chain is ergodic on a finite-dimensional state space, a stationary distribution $\mu\in\mathbb{R}^p$ exists and characterizes the frequency of each state in a long time observation, $\mu_i=\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^n 1_{\{X_k=i\}}$. Furthermore, μ is a stationary distribution if and only if $\mu^\top\mathbf{P}=\mu^\top,\mu_i\geq 0$ and $\sum_{i=1}^p\mu_i=1$. For convenience we also denote $\mu_{\min}=\min_{1\leq i\leq p}\mu_i,\mu_{\max}=\max_{1\leq i\leq p}\mu_i$. The p-by-p frequency matrix $\mathbf{F}$ characterizes how frequent State i jumps to State j for each (i,j) pair in the long run: $\mathbf{F}_{ij}=\frac{1}{n}\sum_{i=1}^n 1_{\{X_k=i,X_{k+1}=j\}}$. Then $\mathbf{P}$ and $\mathbf{F}$ are related via $\mathbf{F}=\mathrm{diag}(\mu)\mathbf{P}$ (or $\mathbf{F}_{i\cdot}=\mu_i\mathbf{P}_{i\cdot}$ for any $1\leq i\leq p$). Some basic properties of $\mathbf{P}$ and $\mathbf{F}$ are collected in Lemma 2 in the proof section, which will be used in the technical analysis in this article.

2 Low-Rank Estimation of Discrete-State Markov Chains

In this paper, we focus on the state compression of finite-state Markov process based on empirical state trajectories. Consider the Discrete-time Markov chain on p states $\{1,\ldots,p\}$.Let us consider the model that the system state transition matrix $\mathbf{P}\in\mathbb{R}^{p\times p}$ is an approximately low-rank matrix; see Figure 1.



$$P\approx U\tilde{P}V^T$$

Figure 1: State compression of high-dimensional Markov chains. Informally speaking, we want to find a low-rank decomposition that approximate the transition probability matrix. When the transition probability matrix $\mathbf{P}$ is low-rank, one can map high-dimensional states into low-dimensional states while preserving most of the system dynamics.

2.1 A Spectral Method for Markov Transition Matrix Estimation

We consider a Markov chain on p states $\{1,\ldots,p\}$ with transition matrix $\mathbf{P}\in\mathbb{R}^{p\times p}$ and frequency matrix $\mathbf{F}\in\mathbb{R}^{p\times p}$. Given $(n+1)$ observable states $\{X_0,\ldots,X_n\}$, it is natural to estimate $\mathbf{P}$ and $\mathbf{F}$ via the following *empirical frequency matrix* and *empirical transition matrix*,

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

$$\tilde{\mathbf{P}} = \left(\tilde{\mathbf{P}}_{ij}\right)_{1\leq i,j\leq p}, \quad \tilde{\mathbf{P}}_{ij} = \begin{cases} \frac{\sum_{k=1}^n 1_{\{X_{k-1}=i, X_k=j\}}}{\sum_{k=1}^n 1_{\{X_{k-1}=i\}}}, & \text{if } \sum_{k=1}^n 1_{\{X_{k-1}=i\}} \geq 1; \\ \frac{1}{p}, & \text{if } \sum_{k=1}^n 1_{\{X_{k-1}=i\}} = 0. \end{cases} \tag{2}$$

The empirical transition matrix is in fact the maximum likelihood estimator [1].

When $\mathbf{F}$ or $\mathbf{P}$ further satisfy low-rank assumption, we consider a spectral method for estimation. Suppose the singular value decomposition of $\tilde{\mathbf{F}}$ is $\tilde{\mathbf{F}} = \tilde{\mathbf{U}}_F \tilde{\mathbf{\Sigma}}_F \tilde{\mathbf{V}}_F^\top$, where $\tilde{\mathbf{U}}_F$ and $\tilde{\mathbf{V}}_F$ are p-by-p orthogonal matrices, $\tilde{\mathbf{\Sigma}}$ is p-by-p diagonal, we propose

$$\hat{\mathbf{F}}^{(r)} = \left(\tilde{\mathbf{U}}_{F,[:,1:r]} \tilde{\mathbf{\Sigma}}_{F,[1:r,1:r]} (\tilde{\mathbf{V}}_{F,[:,1:r]})^\top\right)_+, \tag{3}$$

and $\mathbf{P}$ is estimated after the row-wise normalization as

$$\hat{\mathbf{P}}^{(r)} \in \mathbb{R}^{p\times p}, \quad \hat{\mathbf{P}}^{(r)}_{[i,:]} = \begin{cases} \hat{\mathbf{F}}^{(r)}_{[i,:]} / \sum_{j=1}^p \hat{\mathbf{F}}^{(r)}_{ij}, & \text{if } \sum_{j=1}^p \hat{\mathbf{F}}^{(r)}_{ij} > 0, \\ \frac{1}{p} 1_p, & \text{if } \sum_{j=1}^p \hat{\mathbf{F}}^{(r)}_{ij} = 0. \end{cases} \tag{4}$$

Remark 1. The proposed $\hat{\mathbf{F}}^{(r)}$ and $\hat{\mathbf{P}}^{(r)}$ is related to the hard singular thresholding estimator, which has been applied to various settings, including matrix denoising [9, 42, 15, 8]; matrix completion [7, 10]. However, due to the additional transition matrix structure (see Lemma 2 in the proof section) and Markov dependency, the analysis is different from and far more complicated than most of the previous independent sampling settings.

2.2 Optimal Low-Rank Recovery of Transition Probability Matrices

Next, we investigate the theoretical performance of the proposed estimators $\hat{\mathbf{F}}^{(r)}$ and $\hat{\mathbf{P}}^{(r)}$. The result relies on a key quantity of the Markov mixing time. For any ergodic Markov transition matrix $\mathbf{P}$ with p states and stationary distribution μ, and any value $\varepsilon > 0$, the *Markov mixing time* is defined as

$$\tau(\varepsilon) = \max_{1\leq i\leq p} \min \left\{ k : \forall k' > k, \quad \frac{1}{2} \left\| (\mathbf{P}^{k'})_{[i,:]} - \mu^\top \right\|_1 \leq \varepsilon \right\}. \tag{5}$$

Clearly, $\tau(\varepsilon)$ is an non-negative, integer-valued, and non-increasing function. The Markov mixing time measures how many jumps one needs to ensure that the state is sufficiently random given any specific starting state. The interested readers are referred to [25] for a more comprehensive discussion of Markov mixing time. Based on $\tau := \tau(1/4)$ in particular, the following theoretical upper bound holds for $\hat{\mathbf{F}}^{(r)}$ and $\hat{\mathbf{P}}^{(r)}$.

Theorem 1 (Upper Bound). *Suppose* $\mathbf{P} \in \mathbb{R}^{p\times p}$ *and* $\mu \in \mathbb{R}^p$ *are the transition matrix and invariant distribution of some ergodic Markov process on* p *states. We observe* n *consecutive states from any starting points. Denote* $\tau = \tau(1/4)$*, where* $\tau(t)$ *is the Markov mixing time defined as* (5)*. If* $\text{rank}(\mathbf{P}) \leq r$ *and the truncation rank* $\hat{r}$ *satisfies* $r \leq \hat{r}$*, there exist universal* *constants C_0, C such that when $n \geq C_0 rp \cdot \frac{\mu_{\max}}{p\mu_{\min}^2} \cdot \tau \log^2(n)$, we have*

$$\mathbb{E} \sum_{i=1}^{p} \|\hat{\mathbf{F}}_{i\cdot}^{(\hat{r})} - \mathbf{F}_{i\cdot}\|_1 \leq \sqrt{\frac{C\hat{r}p}{n} \cdot \mu_{\max} p \cdot \tau \log^2(n)}, \tag{6}$$

$$\mathbb{E} \frac{1}{p} \sum_{i=1}^{p} \|\hat{\mathbf{P}}_{i\cdot}^{(\hat{r})} - \mathbf{P}_{i\cdot}\|_1 \leq \sqrt{\frac{C\hat{r}p}{n} \cdot \frac{\mu_{\max}}{\mu_{\min}^2 p} \cdot \tau \log^2(n)}. \tag{7}$$

Remark 2. The proof of Theorem 1 relies on a novel matrix Markov chain concentration inequality with mixing time (Lemma 9), which characterizes the spectral norm distance between $\tilde{\mathbf{F}}$ and $\mathbf{F}$. Then based on the low-rank assumption of $\mathbf{F}$ and $\mathbf{P}$, we perform a careful spectral analysis (Lemma 4) to obtain the error bounds for $\hat{\mathbf{F}}^{(\hat{r})}$ and $\hat{\mathbf{P}}^{(\hat{r})}$.

Remark 3 (Empirical estimators $\tilde{\mathbf{P}}$ and $\tilde{\mathbf{F}}$)**.** The empirical estimators $\tilde{\mathbf{P}}$ and $\tilde{\mathbf{F}}$ usually yield a larger convergence rate since it does not utilize the low-rank structure. Consider the case that $\mathbf{P} = \frac{1}{p} 1_p 1_p^\top$, the empirical frequency and probability estimation yields a larger error rate than $\hat{\mathbf{F}}^{(r)}$ and $\hat{\mathbf{P}}^{(r)}$: $\mathbb{E}\|\tilde{\mathbf{F}} - \mathbf{F}\|_1 \asymp \mathbb{E}\frac{1}{p}\|\tilde{\mathbf{P}} - \mathbf{P}\|_1 \asymp \sqrt{p^2/n}$. In order to obtain consistency estimation, the sufficient sample complexity for $\hat{\mathbf{P}}^{(r)}$ and $\hat{\mathbf{F}}^{(r)}$ is $O(pr\text{polylog}(p))$, which is smaller than $O(p^2)$, i.e., the sufficient sample complexity for $\tilde{\mathbf{F}}$ and $\tilde{\mathbf{P}}$.

The actual value of Markov mixing time $\tau(t)$ is generally difficult to evaluate in practice. On the other hand, $\tau(\varepsilon)$ can be characterized by the following Cheeger's constant and eigen-gap condition.

- *Cheeger's constant [11]:* the following Cheeger's constant is introduced to measure the degree of connectivity for $\mathbf{P}$,

$$\Phi = \min_{\Omega \subseteq \{1,\dots,p\}} \frac{\sum_{i\in\Omega, j\in\Omega^c} \mu_i \mathbf{P}_{ij}}{\min\{\sum_{i\in\Omega} \mu_i, \sum_{i\in\Omega^c} \mu_i\}}. \tag{8}$$

  Cheeger's constant essentially characterizes the connectivity of all different parts of the Markov chain. The larger Φ implies the less chance that the state will "stuck" in some subsets of the states for a long run.

- *Eigen-gap condition:* When $\mathbf{P}$ satisfies the *detailed balance condition*, i.e., $\mu_i \mathbf{P}_{ij} = \mu_j \mathbf{P}_{ji}$ 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. The largest eigenvalue of a reversible Markov transition matrix is always 1; 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 eigen-gap of reversible Markov processes can be estimated from the observable states via a plug-in estimator proposed by [19].

The following results hold as an extension of Theorem 1 based on Cheeger's constant or eigen-gap assumption.

Corollary 1. *Under the assumption of Theorem 1,*

1. *if the Markov chain is with the Cheeger's constant* Φ*, then*

$$\mathbb{E}\|\hat{\mathbf{F}}^{(\hat{r})} - \mathbf{F}\|_1 \leq \sqrt{\frac{C\hat{r}p}{n} \cdot (\mu_{\max} p) \cdot \frac{\log^2(n)}{\Phi^2}}, \tag{9}$$

$$\mathbb{E}\frac{1}{p}\sum_{i=1}^{p} \|\hat{\mathbf{P}}^{(\hat{r})}_{[i,:]} - \mathbf{P}_{[i,:]}\|_1 \leq \sqrt{\frac{C\hat{r}p}{n} \cdot \left(\frac{\mu_{\max}}{\mu_{\min}^2 p}\right) \cdot \frac{\log^2(n)}{\Phi^2}}; \tag{10}$$

2. *if* $\mathbf{P}$ *is reversible and with second largest eigenvalue* $\lambda_2 < 1$*, then*

$$\mathbb{E}\|\hat{\mathbf{F}}^{(\hat{r})} - \mathbf{F}\|_1 \leq \sqrt{\frac{C\hat{r}p}{n} \cdot (\mu_{\max} p) \cdot \frac{\log^2(n)}{1-\lambda_2}}, \tag{11}$$

$$\mathbb{E}\frac{1}{p}\sum_{i=1}^{p} \|\hat{\mathbf{P}}^{(\hat{r})}_{[i,:]} - \mathbf{P}_{[i,:]}\|_1 \leq \sqrt{\frac{C\hat{r}p}{n} \cdot \left(\frac{\mu_{\max}}{\mu_{\min}^2 p}\right) \cdot \frac{\log^2(n)}{1-\lambda_2}}. \tag{12}$$

Correspondingly, the following lower bound results hold for low-rank Markov transition and frequency matrices estimation.

Theorem 2 (Lower Bound)**.** *Consider the following class of low-rank transition and frequency matrices*

$$\begin{aligned} \mathcal{F}_{p,r} =& \Big\{\mathbf{F} \in \mathbb{R}^{p\times p} : \mathbf{F} \in \mathcal{F}, \mathrm{rank}(\mathbf{F}) \leq r\Big\}, \\ \mathcal{P}_{p,r} =& \Big\{\mathbf{P} \in \mathbb{R}^{p\times p} : \mathbf{P} \in \mathcal{P}, \mathrm{rank}(\mathbf{P}) \leq r\Big\}. \end{aligned} \tag{13}$$

where $\mathcal{F}$ *and* $\mathcal{P}$ *are the classes of transition and frequency matrices whose formal characterizations are given in* (53) *and* (52)*. Suppose* $(n+1)$ *consecutive transition states* $\{x_0, \dots, x_n\}$ *from the corresponding Markov chain with starting point randomly generated from the invariant distribution. Then following minimax lower bound for estimation of* $\mathbf{P}$ *and* $\mathbf{F}$ *hold,*

$$\inf_{\hat{\mathbf{F}}} \sup_{\mathbf{F}\in\mathcal{F}_{p,r}} \mathbb{E}\sum_{i=1}^{p} \left\|\hat{\mathbf{F}}_{[i,:]} - \mathbf{F}_{[i,:]}\right\|_1 \geq c\left(\sqrt{\frac{rp}{n}} \wedge 1\right), \tag{14}$$

$$\inf_{\hat{\mathbf{P}}} \sup_{\mathbf{P}\in\mathcal{P}_{p,r}} \mathbb{E}\frac{1}{p}\sum_{i=1}^{p} \left\|\hat{\mathbf{P}}_{[i,:]} - \mathbf{P}_{[i,:]}\right\|_1 \geq c\left(\sqrt{\frac{rp}{n}} \wedge 1\right). \tag{15}$$

where $c > 0$ *is some uniform constant.*

Combining Theorems 1 and 2 together, we have shown that the spectral estimator $\hat{\mathbf{F}}^{(r)}$ and $\hat{\mathbf{P}}^{(r)}$ achieves near minimax-optimal up to logarithmic terms, when $\hat{r}/r$, $\theta_{\max}/\theta_{\min}$, and τ are of constant order.

Remark 4. The proof of Theorem 1 relies on the careful construction of a series of low-rank Markov transition and frequency matrices. Then one aims to show that these low-rank objects are non-separable based on a length-$(n+1)$ Markov train, and the generalized Fano's lemma is applied to develop the desired lower bound results.

2.3 Optimal Compression of State Space

The matrix factorization is an important tool for dimension reduction for high-dimensional data. We further consider the factorization of approximately low-rank Markov process in this section. Suppose the singular value decomposition of $\mathbf{P}$ and $\mathbf{F}$ are

$$\mathbf{P} = [\mathbf{U}_P \ \mathbf{U}_{P\perp}] \begin{bmatrix} \mathbf{\Sigma}_{P1} & 0 \\ 0 & \mathbf{\Sigma}_{P2} \end{bmatrix} \cdot \begin{bmatrix} \mathbf{V}_P \\ \mathbf{V}_{P\perp} \end{bmatrix}, \quad \mathbf{F} = [\mathbf{U}_F \ \mathbf{U}_{F\perp}] \begin{bmatrix} \mathbf{\Sigma}_{F1} & 0 \\ 0 & \mathbf{\Sigma}_{F2} \end{bmatrix} \cdot \begin{bmatrix} \mathbf{V}_F \\ \mathbf{V}_{F\perp} \end{bmatrix}, \tag{16}$$

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 $\mathbf{F} = \mathrm{diag}(\mu) \cdot \mathbf{P}$ and the left multiplication does not change the right singular subspace. We consider the following estimators for the leading singular vectors of $\mathbf{F}$ and $\mathbf{P}$,

$$\begin{aligned} \hat{\mathbf{U}}_F =& \mathrm{SVD}_r\left(\tilde{\mathbf{F}}\right) = \text{leading } r \text{ left singular vectors of } \tilde{\mathbf{F}}; \\ \hat{\mathbf{V}}_F =& \mathrm{SVD}_r\left(\tilde{\mathbf{F}}^\top\right) = \text{leading } r \text{ right singular vectors of } \tilde{\mathbf{F}}; \\ \hat{\mathbf{U}}_P =& \mathrm{SVD}_r\left(\tilde{\mathbf{P}}\right) = \text{leading } r \text{ left singular vectors of } \tilde{\mathbf{P}}; \\ \hat{\mathbf{V}}_P =& \mathrm{SVD}_r\left(\tilde{\mathbf{P}}^\top\right) = \text{leading } r \text{ right singular vectors of } \tilde{\mathbf{P}}. \end{aligned} \tag{17}$$

The proposed estimators satisfy the following theoretical properties.

Theorem 3 (Upper Bounds for Low-rank Stochastic Matrix Factorization)**.** *Suppose we observe n states from an ergodic Markov chain with transition matrix $\mathbf{P}$ from any starting point, where $n \geq \frac{C_0 r \tau \log^2(n) \mu_{\max}}{\mu_{\min}^2}$. The mixing time $\tau := \tau(1/4)$ is defined as (5). Then the proposed estimator (17) satisfies*

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

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

Particularly if we focus on the following class of approximately low-rank stochastic matrices,

$$\begin{aligned}\mathcal{P}_{p,r,\lambda_P} &= \left\{\mathbf{P} \in \mathbb{R}^{p\times p} : \mathbf{P} \in \mathcal{P}, \mathrm{rank}(\mathbf{P}) \leq r, \sigma_r(\mathbf{P}) - \sigma_{r+1}(\mathbf{P}) \geq \lambda_P\right\};\\ \mathcal{F}_{p,r,\lambda_F} &= \left\{\mathbf{P} \in \mathbb{R}^{p\times p} : \mathbf{F} \in \mathcal{F}, \mathrm{rank}(\mathbf{F}) \leq r, \sigma_r(\mathbf{F}) - \sigma_{r+1}(\mathbf{F}) \geq \lambda_F\right\},\end{aligned} \tag{18}$$

we can develop the following lower bound results.

Theorem 4 (Lower Bound for Low-rank Stochastic Matrix Factorization)**.** *Suppose* $2 \leq r \leq p/2$*, the following lower bound holds for estimations of* U_P, V_P, U_F, V_F*.*

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

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

Remark 5. Theorems 3 and 4 together yields a near-optimal rate of convergence for factorizing $\mathbf{P}$ and $\mathbf{F}$, when $\mu_{\max}/\mu_{\min}$ and τ are in the constant order.

3 Optimal State Aggregation for Lumpable Markov Processes

Motivated by previous discussions on low-rank stochastic matrix estimation and factorization for Markov processes, we consider the lumpable complex network in this section.

3.1 Lumpable Markov Chains and State Aggregation

Suppose the targeting stochastic network is lumpable, in the sense that the node can be partitioned into smaller number of groups, which still form a Markov chain. Our goal is to partition these nodes into sub-groups according to lumpability (see, e.g. [29, 16]).

Definition 1 (Lumpability of Stochastic Network [23])**.** *A stochastic process with Markov transition matrix* $\mathbf{P}$ *is lumpable with respect to partition* $G_1, \ldots, G_r$*, if for any* $1 \leq k < l \leq r$*,* $G_k \cap G_l = \emptyset$*,* $G_1 \cup \cdots \cup G_r = \{1, \ldots, p\}$*, and for any two states in the same group, i.e.,* $i, i' \in G_k$*,*

$$\sum_{j\in G_k} \mathbf{P}_{i,j} = \sum_{j\in G_k} \mathbf{P}_{i',j}. \tag{19}$$

When the Markov chain is lumpable with r groups, say $G_1, \ldots, G_r \subseteq \{1, \ldots, p\}$, $\mathbf{P}$ exhibits a block-wise structure after permutation (see $\mathbf{P}_\sigma$ in Figure 2). The original p states can be compressed into r groups, where the law of walkers on $\{G_1, \ldots, G_r\}$ remains a Markov chain. For convenience, let $\mathbf{Z} \in \mathbb{R}^{p\times r}$ be the group membership indicator, such that $Z_{ik} = 1_{\{i\in G_k\}}$. Following the statements in [29] and [16], the stochastic matrices have the following decomposition.

Proposition 1 (Lumpability Transition Matrix Decomposition)**.** *Suppose* $\mathbf{P}$ *and* $\mathbf{F}$ *are the transition and frequency matrices of a lumpable Markov chain. Then there exists the* compressed Markov transition matrix $\bar{\mathbf{P}} \in \mathbb{R}^{r\times r}$ *for state space* $\{G_1, \ldots, G_r\}$*, such that*

$$\bar{\mathbf{P}}_{kl} = \sum_{j\in G_l} \mathbf{P}_{ij}, \quad \text{where} \quad i \in G_k. \tag{20}$$

$\mathbf{P}$ *and* $\mathbf{F}$ *can be further decomposed as follows,*

$$\mathbf{P} = \mathbf{P}_1 + \mathbf{P}_2, \quad (\mathbf{P}_1)_{ij} = \frac{1}{|G_l|}\sum_{j\in G_l} \mathbf{P}_{ij}, \quad \text{if} \quad j \in G_l, \tag{21}$$

$$\mathbf{F} = \mathbf{F}_1 + \mathbf{F}_2, \quad (\mathbf{F}_1)_{ij} = \frac{1}{|G_l|}\sum_{j\in G_l} (\mathbf{F}_1)_{ij}, \quad \text{if} \quad j \in G_l. \tag{22}$$

$\mathbf{P}_2, \mathbf{F}_2$ *can be seen as the remainders of low-rank approximation. Equivalently,* $\mathbf{P}_1$ *can be written as* $\mathbf{P}_1 = \mathbf{Z}\cdot\bar{\mathbf{P}}\cdot \mathrm{diag}(|G_1|^{-1}, \ldots, |G_r|^{-1})\cdot\mathbf{Z}^\top$*. Moreover,* $\mathbf{P}_1\mathbf{P}_2^\top = 0$*,* $\mathbf{F}_1\mathbf{F}_2^\top = 0$*.*

If $\mathbf{P}_1 = \mathbf{U}_{P_1}\mathbf{\Sigma}_{P_1}\mathbf{V}_{P_1}^\top, \mathbf{F}_1 = \mathbf{U}_{F_1}\mathbf{\Sigma}_{F_1}\mathbf{V}_{F_1}^\top$ *are the singular value decompositions,* $\mathbf{U}_{P_1}$*,* $\mathbf{V}_{P_1}, \mathbf{V}_{F_1}$ *are piecewise linear based on group partitions. To be specific,* $(\mathbf{U}_{P_1})_{[i,:]} = (\mathbf{U}_{P_1})_{[i',:]}$*,* $(\mathbf{V}_{P_1})_{[i,:]} = (\mathbf{V}_{P_1})_{[i',:]}$*, and* $(\mathbf{V}_{F_1})_{[i,:]} = (\mathbf{U}_{F_1})_{[i',:]}$ *for any* $i, i' \in G_k$*. On the other hand, unless the Markov chain is reversible,* $\mathbf{U}_{F_1}$ *does not necessarily have piecewise linear structure.*

Particularly, the lumpable transition matrix $\mathbf{P}$ and its decomposition are illustrated in Figure 2.

3.2 The State Aggregation Procedure and High-Probably Recovery of Meta-States

Given $\mathrm{rank}(\mathbf{P}_1) = r$ and $\mathbf{P}_2$ is of small amplitude, $\mathbf{P}$ can be seen as an approximately rank-r matrix. Thus, the proposed low-rank factorization method sheds light to the partitions structure estimation for lumpable stochastic processes. When we implement k-means algorithm on the rows of $\hat{\mathbf{U}}_{[:,1:r]}$ to partition the p states $\{1, \ldots, p\}$ into r groups, the following results on misclassification rate hold.

Theorem 5. *Suppose all assumptions in Theorem 1 holds, the targeting Markov process is lumpable with respect to partition* $G_1, \ldots, G_r$*. Let* $\hat{\mathbf{U}}_F$*,* $\hat{\mathbf{V}}_F$*,* U*, and* $\mathbf{V}$ *be the leading* r *left and right singular vectors of* $\hat{\mathbf{F}}$ *and* $\mathbf{F}$*, respectively. Then*

$$\begin{aligned}&\mathbb{E}\left(\|\sin\Theta(\hat{\mathbf{U}}_F, \mathbf{U}_{F_1})\| \vee \|\sin\Theta(\hat{\mathbf{V}}_F, \mathbf{V}_{F_1})\|\right)\\ &\leq \left(C\frac{\sqrt{1/(pn)\cdot(p\mu_{\max})\tau\log^2(n)}}{\sigma_r(\mathbf{F}_1)} + \frac{2\|\mathbf{F}_2\|}{\sigma_r(\mathbf{F}_1)}\right)\wedge 1.\end{aligned} \tag{23}$$

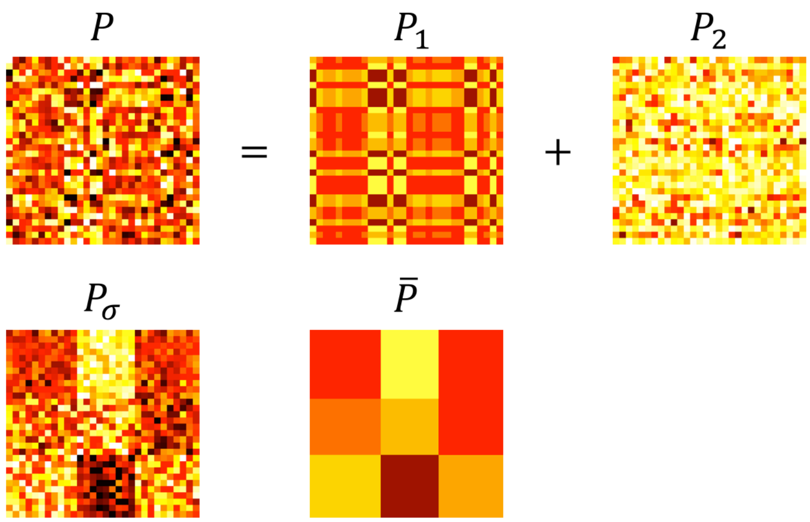


Figure 2: Heat map illustration of P, P_1, P_2, P_σ, and $\bar{P}$

When the Markov chain is further reversible, we have

$$\mathbb{E}\left(\|\sin\Theta(\hat{\mathbf{U}}_F, \mathbf{U}_{F_1})\| \vee \|\sin\Theta(\hat{\mathbf{V}}_F, \mathbf{V}_{F_1})\|\right) \leq \left(C\frac{\sqrt{1/(pn)\cdot p\mu_{\max}\tau\log^2(n)}}{\sigma_r(\mathbf{F}_1)}\right) \wedge 1. \quad (24)$$

Corollary 2 (Misclassification Rate)**.** *Suppose one performs k-means on all p rows of $\hat{\mathbf{V}}_F$ and partition into r groups, i.e., calculate*

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

If (23) *holds, there exists a permutation among r group memberships, π, such that*

$$\sum_{j=1}^{r}\frac{|\{i: i\in G_j,\ \ but\ i\notin \hat{G}_{\pi(j)}\}|}{|G_j|} \leq \left(\frac{Cr/(pn)}{\sigma_r^2(\mathbf{F}_1)}\cdot p\mu_{\max}\cdot\tau\log^2(n) + \frac{4\|\mathbf{F}_2\|^2}{\sigma_r^2(\mathbf{F}_1)}\right)\wedge r.$$

If the Markov chain is further reversible, one has

$$\sum_{j=1}^{r}\frac{|\{i: i\in G_j,\ \ but\ i\notin \hat{G}_{\pi(j)}\}|}{|G_j|} \leq \left(\frac{Cr/(pn)}{\sigma_r^2(\mathbf{F}_1)}\cdot p\mu_{\max}\cdot\tau\log^2(n)\right)\wedge r.$$

Here $\left|\{i: i\in G_j,\ \ but\ i\notin\hat{G}_{\pi(j)}\}\right|$ *presents the number of nodes in G_j but was classified into the other group by mistake.*

Remark 6. We prefer performing k-means on $\hat{\mathbf{V}}_F$ for lumpability network partition among the other singular subspace estimations, since (a) the estimation of $\hat{\mathbf{U}}_P$ and $\hat{\mathbf{V}}_P$ yields larger bounds; (b) $\mathbf{U}_F$ typically does not have piece-wise linear structure (see Proposition 1), then the outcome from k-means on $\hat{\mathbf{U}}_F$ does not provide valid partition estimations.

4 Continuous-State Markov Process Low-rank Kernel Estimation

Now we further consider the transition kernel estimation for continuous-state low-rank Markov chain. Given an ergodic Markov process on the continuous-state space $\mathcal{S}$, let $\mathbf{P}, \mathbf{F} : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ be the transition and frequency kernels, $\mu : \mathcal{S} \to \mathbb{R}$ be the invariant distribution. If one observes $(n+1)$ states $X = \{x_0, \ldots, x_n\}$ from such Markov process, the goal is to estimate $\mathbf{P}$ and $\mathbf{F}$ based on these state transitions. Similarly as the discrete-state Markov processes, $\mathbf{P}$, $\mathbf{F}$ and μ are related as $\mathbf{F}(a,b) = \mu(a)\mathbf{P}(a,b), \forall a, b \in \mathcal{S}$.

4.1 Low-Rank Kernel Estimation of Continuous Markov Processes

We consider the method for estimating low-rank continuous-state Markov processes.

1. Let $\mathbf{K} : \mathbb{R}^2 \to \mathbb{R}$ be a bivariate smooth kernel and $h > 0$ be the bandwidth. Calculate the following local smoothed empirical surface $\tilde{\mathbf{F}} : \mathbb{R}^2 \to \mathbb{R}$,

$$\tilde{\mathbf{F}}(a,b) = \frac{1}{n}\sum_{k=1}^{n} \mathbf{K}\left(\frac{a - x_{k-1}}{h}, \frac{b - x_k}{h}\right). \tag{25}$$

2. Perform functional SVD on $\tilde{\mathbf{F}}$ and extract the first r singular components,

$$\hat{\lambda}_i \geq 0, \quad \hat{u}_i \in \mathcal{L}^2(\mathbb{R}), \quad \hat{v}_i \in \mathcal{L}^2(\mathbb{R}), \quad i = 1, \ldots, r. \tag{26}$$

   Numerically one can obtain these objects via discretization and matrix singular value decomposition.

3. Then we construct the estimators for frequency and transition kernels and the singular subspaces,

$$\hat{\mathbf{F}} = \left(\sum_{i=1}^{r} \hat{\lambda}_i \hat{u}_i \otimes \hat{v}_i\right)_+, \quad \text{i.e.} \quad \hat{\mathbf{F}}(a,b) = \left(\sum_{i=1}^{r} \hat{\lambda}_i \hat{u}_i(a)\hat{v}_i(b)\right)_+,$$

$$\hat{\mathbf{P}} : \mathbb{R}^2 \to \mathbb{R}, \quad \hat{\mathbf{P}}(a,b) = \frac{\hat{\mathbf{F}}(a,b)}{\int \hat{\mathbf{F}}(a,b')db'},$$

4.2 Upper Bounds of Estimation Errors

Suppose kernel $\mathbf{F}$ has functional singular value decomposition $\mathbf{F} = \sum_{k=1}^{\infty} \lambda_k u_k \otimes v_k$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$, $\{u_i(\cdot)\}_{i=1}^{\infty}$ and $\{v_i(\cdot)\}_{i=1}^{\infty}$ are two orthonormal systems in ℓ_2 finite space $\mathcal{L}^2(\mathbb{R})$. The Markov chain mixing time can be defined as follows,

$$\tau(\varepsilon) = \sup_{a \in \mathbb{R}} \min\left\{k : \forall k' \geq k, \frac{1}{2}\|\mathbf{P}^{(k)}(a,\cdot) - \mu(\cdot)\|_1 = \frac{1}{2}\int |\mathbf{P}^{(k)}(a,b) - \mu(b)|db \leq \varepsilon\right\}. \tag{27}$$

Here, $\mathbf{P}^{(k)}$ is the k-th power of $\mathbf{P}$,

$$\mathbf{P}^{(k)}: \mathbb{R}^2 \to \mathbb{R}, \quad \mathbf{P}^{(k)}(a,b) = \int_{\mathbb{R}^{k-1}} \mathbf{P}(a,x_1)\mathbf{P}(x_1,x_2)\cdots\mathbf{P}(x_{k-1},b)dx_1\cdots dx_{k-1}. \tag{28}$$

Next, we establish the upper bound for estimation errors of $\hat{\mathbf{F}}$ and $\hat{\mathbf{P}}$.

Theorem 6. *Suppose a continuous-state Markov chain is with transition kernel* $\mathbf{P}$*, frequency kernel* $\mathbf{F}$*, compact support* $\mathcal{S}$ *with constant measure, and invariant distribution* μ*. If* $\mathbf{K}$ *is an order-*$\lfloor\beta\rfloor$ *kernel and* $\mathbf{F}$ *belongs to the 2-dimensional Nikol'ski class [48, Chapter 1]:*

$$\mathbf{F} \in \mathcal{H}(\beta, L) = \Bigg\{ \mathbf{F}: \mathbb{R}^2 \to \mathbb{R} \text{ is } \lfloor\beta\rfloor\text{-th differentiable};$$
$$\left(\sum_{\alpha \le \lfloor\beta\rfloor} \int_{\mathbb{R}^2} \left(D^\alpha \mathbf{F}(x_1+t_1, x_2+t_2) - D^\alpha \mathbf{F}(x_1,x_2)\right)^2 dx_1 dx_2 \right)^{1/2} \le L(t_1^2+t_2^2)^{(\beta-\lfloor\beta\rfloor)/2} \Bigg\}.$$

for constant $L > 0$*. Here,* $D^\alpha \mathbf{F}$ *is the differential operator with* $\alpha = (\alpha_1, \alpha_2) \in \mathbb{R}^2$ *as the order of derivatives. When the bandwidth* $h = n^{-\frac{1}{4\beta+2}}$*, the estimators* $\hat{\mathbf{F}}$ *and* $\hat{\mathbf{P}}$ *yield the following estimation upper bounds,*

$$\mathbb{E}\int |\hat{\mathbf{F}}(a,b) - \mathbf{F}(a,b)| da db \le C\left(\frac{\log(n)\tau}{n}\right)^{\frac{2\beta}{4\beta+2}}; \tag{29}$$

$$\mathbb{E}\int |\hat{\mathbf{P}}(a,b) - \mathbf{P}(a,b)| da db \le C\mu_{\min}^{-1}\left(\frac{\log(n)\tau}{n}\right)^{\frac{2\beta}{4\beta+2}}. \tag{30}$$

To better characterize the continuous Markov chain mixing time, we introduce the following continuous-state Markov chain Cheeger's constant and eigen-gap.

Definition 2 (Cheeger's constant for continuous-state Markov chain)**.** *The Cheeger's constant for continuous-state Markov chain is defined as (see, e.g. [2]),*

$$\Phi = \inf_{\Omega \subseteq \mathbb{R}} \frac{\int_{\Omega\times\Omega^c} \mu(a)\mathbf{P}(a,b)dadb}{\min\{\int_\Omega \mu(a)da, \int_{\Omega^c}\mu(a)da\}}.$$

Definition 3 (Eigen-gap for reversible continuous-state Markov chain)**.** *Suppose* $\mathbf{P}$ *and* μ *are the transition kernel of a continuous-state Markov chain. Assume the Markov chain is reversible, in the sense that* $\mu(a)\mathbf{P}(a,b) = \mu(b)\mathbf{P}(b,a)$ *or equivalently* $\mathbf{F}$ *is self-adjoint. Recall that* $\lambda_k(\mathbf{P})$ *is defined as the* k*-th largest eigenvalue of* $\mathbf{P}$*, then the eigen-gap for reversible continuous-state Markov chain is defined as* $1 - \lambda_2(\mathbf{P})$*.*

With Cheeger's constant or eigen-gap condition, the following upper bounds hold for estimation errors.

Corollary 3 (Continuous Markov Chain Transition Operator Estimation under Cheeger's Constant and Spectral Gap)**.** *Under the assumption of Theorem 6,*

1. *if the Markov process is with Cheeger's constant* Φ*, then* (29) *and* (30) *with* τ *replaced by* $1/\Phi^2$ *hold;*

2. *if the reversible Markov process is with second largest eigenvalue* $\lambda_2 < 1$*, then* (29) *and* (30) *with* τ *replaced by* $1/(1-\lambda_2)$ *hold.*

5 Proofs of Main Results

We collect the proofs for the technical results in this section.

5.1 Proof of Theorem 1

First, for any matrix $\mathbf{A}$ with singular value decomposition $\mathbf{A} = \sum_i \sigma_i u_i v_i^\top$, we define $\mathbf{A}_{\max(r)} = \sum_{k=1}^r \sigma_k u_k v_k^\top$ and $\mathbf{A}_{-\max(r)} = \sum_{k=r+1}^{p_1\wedge p_2} \sigma_k u_k v_k^\top = \mathbf{A} - \mathbf{A}_{\max(r)}$ as the leading and non-leading parts of $\mathbf{A}$.

Now we consider the proof for Theorem 1. By Lemma 9, there exists constants $C > 0$ and $c > 1$ such that

$$\left\|\tilde{\mathbf{F}} - \mathbf{F}\right\| \leq C\sqrt{\frac{\mu_{\max}\cdot\tau(\sqrt{\mu_{\max}/n})\cdot\log(n)}{n}}, \tag{31}$$

$$\|\tilde{\mu} - \mu\|_\infty \leq C\sqrt{\frac{\mu_{\max}\cdot\tau(\sqrt{\mu_{\max}/n})\cdot\log(n)}{n}}, \tag{32}$$

with probability at least $1 - O(n^{-c})$. By Lemma 6, one further has $\tau(\sqrt{\mu_{\max}/n}) \leq C\tau(\log(n) + \log(1/\mu_{\max})) \leq C\tau\log(n)$, provided that $n \geq C_0 rp\frac{\mu_{\max}}{\mu_{\min}^2 p}\cdot\tau\log^2(n) \geq C_0 rp \geq C_0/\mu_{\max}$. Thus, (31) and (32) implies

$$\mathbb{P}\left\{\max\left\{\left\|\tilde{\mathbf{F}} - \mathbf{F}\right\|, \|\tilde{\mu} - \mu\|_\infty\right\} \leq C\sqrt{\frac{\mu_{\max}\tau\log^2(n)}{n}}\right\} \geq 1 - n^{-c}. \tag{33}$$

Assume (33) holds, define $\hat{\mathbf{F}}_0^{(\hat{r})} = \tilde{\mathbf{U}}_{F,[:,1:r]}\tilde{\mathbf{\Sigma}}_{F,[1:r,1:r]}\tilde{\mathbf{V}}_{F,[:,1:r]}^\top$, then $\hat{\mathbf{F}}^{(\hat{r})} = (\hat{\mathbf{F}}_0^{(\hat{r})})_+$. By Lemma 4 and $\hat{\mathbf{F}}_0^{(\hat{r})}$ is the leading $\hat{r}$ principal components of $\tilde{\mathbf{F}}$, we have

$$\begin{aligned}\left\|\hat{\mathbf{F}}_0^{(\hat{r})} - \mathbf{F}\right\|_F &\leq 2\sqrt{2\hat{r}}\left\|\tilde{\mathbf{F}} - \mathbf{F}\right\| + 2\sqrt{2\hat{r}}\|\mathbf{F}_{-\max(\hat{r})}\| + \|\mathbf{F}_{-\max(\hat{r})}\|_F \\ \leq & C\sqrt{\frac{\hat{r}\mu_{\max}\tau\log^2(n)}{n}} + C\sqrt{\hat{r}}\left\|\mathbf{F}_{-\max(\hat{r})}\right\| + C\|\mathbf{F}_{-\max(\hat{r})}\|_F.\end{aligned} \tag{34}$$

By Hölder's inequality,

$$\begin{aligned}
&\left\|\hat{\mathbf{F}}^{(\hat{r})}-\mathbf{F}\right\|_1=\sum_{i=1}^p\sum_{j=1}^p|\hat{\mathbf{F}}^{(\hat{r})}_{0,ij}-\mathbf{F}_{ij}|\leq\sum_{j=1}^p|\hat{\mathbf{F}}^{(\hat{r})}_{ij}-\mathbf{F}_{ij}|\\
\leq& p\left(\sum_{i=1}^p\sum_{j=1}^p|\hat{\mathbf{F}}^{(\hat{r})}_{ij}-\mathbf{F}_{ij}|^2\right)^{1/2}=p\left\|\hat{\mathbf{F}}^{(\hat{r})}-\mathbf{F}\right\|_F\\
\leq& Cp\sqrt{\frac{\hat{r}\mu_{\max}\tau\log^2(n)}{n}}+C\sqrt{\hat{r}}p\left\|\mathbf{F}_{-\max(\hat{r})}\right\|+Cp\|\mathbf{F}_{-\max(\hat{r})}\|_F\\
=&C\sqrt{\frac{\hat{r}p}{n}\cdot p\mu_{\max}\cdot\tau\log^2(n)}+C\sqrt{\hat{r}}p\left\|\mathbf{F}_{-\max(\hat{r})}\right\|+Cp\|\mathbf{F}_{-\max(\hat{r})}\|_F.
\end{aligned}\tag{35}$$

provided that (31), (32) hold. On the other hand, we always have the following loose upper bound,

$$\left\|\hat{\mathbf{F}}^{(\hat{r})}-\mathbf{F}\right\|_1\leq\left\|\hat{\mathbf{F}}^{(\hat{r})}\right\|+\|\mathbf{F}\|_1\leq 2.\tag{36}$$

Since (31) and (32) hold with probability $1-O(n^{-c})$, we finally have

$$\begin{aligned}
\mathbb{E}\left\|\hat{\mathbf{F}}^{(\hat{r})}-\mathbf{F}\right\|_1=&\mathbb{E}\left\|\hat{\mathbf{F}}^{(\hat{r})}-\mathbf{F}\right\|_1 1_{\{(31)(32)\text{ hold}\}}+\mathbb{E}\left\|\hat{\mathbf{F}}^{(\hat{r})}-\mathbf{F}\right\|_1 1_{\{(31)(32)\text{ hold}\}^c}\\
\overset{(35)(36)}{\leq}&C\sqrt{\frac{\hat{r}p}{n}\cdot p\mu_{\max}\cdot\tau\log^2(n)}+C\sqrt{\hat{r}}p\left\|\mathbf{F}_{-\max(\hat{r})}\right\|+Cp\|\mathbf{F}_{-\max(\hat{r})}\|_F\\
&+2\cdot\mathbb{P}\left(\text{either (35) or (36) not holds}\right)\\
\leq&C\sqrt{\frac{\hat{r}p}{n}\cdot p\mu_{\max}\cdot\tau\log^2(n)}+C\sqrt{\hat{r}}p\left\|\mathbf{F}_{-\max(\hat{r})}\right\|+Cp\|\mathbf{F}_{-\max(\hat{r})}\|_F+2n^{-c},
\end{aligned}$$

which implies the error bound for $\hat{\mathbf{F}}^{(\hat{r})}$ provided that $\mathrm{rank}(\mathbf{F})\leq\hat{r}$.

Next we consider the error bound for $\hat{\mathbf{P}}^{(\hat{r})}$. Since $\hat{\mathbf{P}}^{(\hat{r})}_{i\cdot}=\frac{\hat{\mathbf{F}}^{(\hat{r})}_{i\cdot}}{\|\hat{\mathbf{F}}^{(\hat{r})}_{i\cdot}\|_1},\mathbf{P}^{(\hat{r})}_{i\cdot}=\frac{\mathbf{F}^{(\hat{r})}_{i\cdot}}{\|\mathbf{F}^{(\hat{r})}_{i\cdot}\|_1}$, and $\|\mathbf{F}_{i\cdot}\|_1=\mu_i\geq\mu_{\min}$, therefore

$$\begin{aligned}
\frac{1}{p}\left\|\hat{\mathbf{P}}^{(\hat{r})}-\mathbf{P}\right\|_1=&\frac{1}{p}\sum_{i=1}^p\|\hat{\mathbf{P}}^{(\hat{r})}_{i\cdot}-\mathbf{P}_{i\cdot}\|_1\overset{\text{Lemma 3}}{\leq}\sum_{i=1}^p\frac{2\|\hat{\mathbf{F}}^{(\hat{r})}_{i\cdot}-\mathbf{F}_{i\cdot}\|_1}{p\mu_{\min}}\\
\leq&C\sqrt{\frac{\hat{r}}{np}\cdot\frac{\mu_{\max}}{p\mu_{\min}^2}\cdot\tau\log^2(n)}+\frac{C\sqrt{r}\left\|\mathbf{F}_{-\max(\hat{r})}\right\|+C\|\mathbf{F}_{-\max(\hat{r})}\|_F}{\mu_{\min}}.
\end{aligned}$$

One can similarly show

$$\frac{1}{p}\left\|\hat{\mathbf{P}}^{(\hat{r})}-\mathbf{P}\right\|_1\leq C\sqrt{\frac{\hat{r}}{np}\cdot\frac{\mu_{\max}}{p\mu_{\min}^2}\cdot\tau\log^2(n)}+\frac{C\sqrt{r}\left\|\mathbf{F}_{-\max(\hat{r})}\right\|+C\|\mathbf{F}_{-\max(\hat{r})}\|_F}{\mu_{\min}}.$$

Therefore, we have finished the proof for this theorem. □

5.2 Proof of Corollary 1

We only focus on the case with Cheeger's constant as the proof for eigen-gap-based scenario essentially follows. By Lemma 7, one has

$$\tau(\varepsilon) \leq \frac{2}{\Phi^2}\log(1/(2\varepsilon\mu_{\min})), \quad \forall\varepsilon > 0. \tag{37}$$

Combing (37) with Lemma 9 and $n \geq C\frac{\mu_{\max}r\tau\log^2(n)}{\mu_{\min}^2}$ for large constant $C > 0$, we have

$$\begin{aligned}\max\left\{\left\|\tilde{\mathbf{F}} - \mathbf{F}\right\|, \|\tilde{\mu} - \mu\|_\infty\right\} &\leq C\sqrt{\frac{\mu_{\max}\tau(\sqrt{\mu_{\max}/n})\log(n)}{n}} \\ \leq C\sqrt{\frac{\mu_{\max}\log(\sqrt{n}/(2\mu_{\min}\sqrt{\mu_{\max}}))\log(n)}{n\Phi^2}} &\leq C\sqrt{\frac{\mu_{\max}\log^2(n)}{n\Phi^2}},\end{aligned} \tag{38}$$

for large constant $C > 0$ and with probability at least $1 - O(n^{-c})$. Applying (38), the rest of the proof exactly follows from the one of Theorem 1. $\square$

5.3 Proof of Theorem 2

First, it is helpful to study the Kullback-Leibler divergence between two Markov processes with same the same state space $\{1, \ldots, p\}$ but different transition matrices $\mathbf{P}$ and $\mathbf{Q}$. Suppose μ is the stationary distribution of $\mathbf{P}$, $X^{(1)} = \{x_0^{(1)}, \ldots, x_n^{(1)}\}$ and $X^{(2)} = \{x_0^{(2)}, \ldots, x_n^{(2)}\}$ are two Markov chains generated from $\mathbf{P}$ and $\mathbf{Q}$, and $x_0^{(1)} \sim \mu$, i.e. the starting point of $X^{(1)}$ is from its stationary distribution. Then, clearly $x_0^{(1)}, \ldots, x_n^{(1)}$ identically satisfy the distribution of μ (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,

$$\begin{aligned}
&D_{KL}\left(X^{(1)}||X^{(2)}\right) := \sum_{X\in[p]^{n+1}} p_{X^{(1)}}(X)\log\left(\frac{p_{X^{(1)}}(X)}{p_{X^{(2)}}(X)}\right) \\
=&\sum_{i_0,\ldots,i_n\in[p]^{n+1}} \mathbb{P}\left(X^{(1)} = (i_1, \ldots, i_n)\right)\cdot\log\left(\frac{\mathbb{P}\left(X^{(1)} = (i_0, \ldots, i_n)\right)}{\mathbb{P}\left(X^{(2)} = (i_0, \ldots, i_n)\right)}\right) \\
=&\sum_{i_0,\ldots,i_n\in[p]^{n+1}} \mu_{i_0}\mathbf{P}_{i_0,i_1}\cdots\mathbf{P}_{i_{n-1},i_n}\log\left(\frac{\mu_{i_0}\mathbf{P}_{i_0,i_1}\cdots\mathbf{P}_{i_{n-1},i_n}}{\mu_{i_0}\mathbf{Q}_{i_0,i_1}\cdots\mathbf{Q}_{i_{n-1},i_n}}\right) \\
=&\sum_{i_0,\ldots,i_{n-1}\in[p]^n}\sum_{i_n\in[p]} \mu_{i_0}\mathbf{P}_{i_0,i_1}\cdots\mathbf{P}_{i_{n-1},i_n}\left\{\log\left(\frac{\mu_{i_0}\mathbf{P}_{i_0,i_1}\cdots\mathbf{P}_{i_{n-2},i_{n-1}}}{\mu_{i_0}\mathbf{Q}_{i_0,i_1}\cdots\mathbf{Q}_{i_{n-2},i_{n-1}}}\right) + \log\left(\frac{\mathbf{P}_{i_{n-1},i_n}}{\mathbf{Q}_{i_{n-1},i_n}}\right)\right\} \\
=&D_{KL}\left(\{x_0^{(1)}, \ldots, x_{n-1}^{(1)}\}||\{x_0^{(2)}, \ldots, x_{n-1}^{(2)}\}\right) + \sum_{i_{n-1}\in[p]}\mu_{i_{n-1}}\sum_{i_n\in[p]}\mathbf{P}_{i_{n-1},i_n}\log\left(\frac{\mathbf{P}_{i_{n-1},i_n}}{\mathbf{Q}_{i_{n-1},i_n}}\right) \\
=&D_{KL}\left(\{x_0^{(1)}, \ldots, x_{n-1}^{(1)}\}||\{x_0^{(2)}, \ldots, x_{n-1}^{(2)}\}\right) + \sum_{i\in[p]}\mu_i D_{KL}\left(\mathbf{P}_{i\cdot}||\mathbf{Q}_{i\cdot}\right).
\end{aligned}$$

Then it is easy to use induction to show that

$$
\begin{aligned}
&D_{KL}\left(X^{(1)}||X^{(2)}\right) = D_{KL}\left(\{x_0^{(1)},\dots,x_{n-1}^{(1)}\}||\{x_0^{(2)},\dots,x_{n-1}^{(2)}\}\right) + \sum_{i\in[p]}\mu_i D_{KL}\left(\mathbf{P}_{i\cdot}||\mathbf{Q}_{i\cdot}\right) \\
&= \cdots = D_{KL}\left(x_0^{(1)}\middle\|x_0^{(2)}\right) + n\sum_{i\in[p]}\mu_i D_{KL}\left(\mathbf{P}_{i\cdot}||\mathbf{Q}_{i\cdot}\right).
\end{aligned} \tag{39}
$$

Next, we consider the proof the frequency matrix estimation lower bound. Let $p_0 = \lfloor p/2 \rfloor$, $l_0 = \lfloor p_0/\{2(r-1)\}\rfloor$, and construct

$$
\mathbf{P}^{(k)} = \frac{1}{p}1_p 1_p^\top + \frac{\tau}{2p}\begin{bmatrix} \overbrace{\mathbf{R}^{(k)} \ \cdots \ \mathbf{R}^{(k)}}^{l_0} & \overbrace{-\mathbf{R}^{(k)} \ \cdots \ -\mathbf{R}^{(k)}}^{l_0} & \mathbf{0}_{p_0\times(p-2l_0(r-1))} \\ -\mathbf{R}^{(k)} \ \cdots \ -\mathbf{R}^{(k)} & \mathbf{R}^{(k)} \ \cdots \ \mathbf{R}^{(k)} & \mathbf{0}_{p_0\times(p-2l_0(r-1))} \\ \mathbf{0}_{(p-2p_0)\times(l_0(r-1))} & \mathbf{0}_{(p-2p_0)\times(l_0(r-1))} & \mathbf{0}_{(p-2p_0)\times(p-2l_0(r-1))} \end{bmatrix} \tag{40}
$$

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 $0 < \tau \leq 1$ is some constant to be determined later. Then clearly, $\mathbf{P}^{(k)}$ is a transition matrix, and $\frac{1}{p}1_p$ is the stationary distribution, then the corresponding frequency matrix $\mathbf{F}^{(k)} = \frac{1}{p}\mathbf{P}^{(k)}$. Now for any $k \neq l$,

$$
\|\mathbf{F}^{(k)} - \mathbf{F}^{(l)}\|_1 = \frac{1}{p}\|\mathbf{P}^{(k)} - \mathbf{P}^{(l)}\|_1 = \frac{2l_0\tau}{p^2}\|\mathbf{R}^{(k)} - \mathbf{R}^{(l)}\|_1 = \frac{2l_0\tau}{p^2}\sum_{i=1}^{p_0}\sum_{j=1}^{r-1}\left|\mathbf{R}_{ij}^{(k)} - \mathbf{R}_{ij}^{(l)}\right|.
$$

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

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

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

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

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

$$
\begin{aligned}
&\mathbb{P}\left(\forall 1 \leq k < l \leq m, \quad \frac{l_0\tau p_0(r-1)}{p^2} \leq \left\|\mathbf{F}^{(k)} - \mathbf{F}^{(l)}\right\|_F^2 \leq \frac{3l_0\tau p_0(r-1)}{p^2}\right) \\
\geq &1 - m(m-1)\exp\left(\frac{-p_0(r-1)}{28}\right) > 1 - m^2\exp\left(\frac{-p_0(r-1)}{28}\right) > 0.
\end{aligned}
$$

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

$$
\forall 1 \leq k < l \leq m, \quad \frac{l_0\tau p_0(r-1)}{p^2} \leq \left\|\mathbf{F}^{(k)} - \mathbf{F}^{(l)}\right\|_1 \leq \frac{3l_0\tau p_0(r-1)}{p^2}. \tag{41}
$$

We thus assume $\{\mathbf{R}^{(1)}, \ldots, \mathbf{R}^{(m)}\} \subseteq \{-1, 1\}^{p_0 \times (r-1)}$ satisfies (41).

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 jump according to $\mathbf{P}^{(k)}$ and $\mathbf{F}^{(k)}$. Based on the calculation in (39),

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

Base on Lemma 5 and $1/(2p) \leq \mathbf{P}_{ij}^{(k)} \leq 3/(2p)$, we further have $D_{KL}\left(\mathbf{P}_{i\cdot}^{(k)} \middle\| \mathbf{P}_{i\cdot}^{(l)}\right) \leq 3p\|\mathbf{P}_{i\cdot}^{(k)} - \mathbf{P}_{i\cdot}^{(l)}\|_2^2$. Thus, for any $1 \leq k < l \leq m$,

$$\begin{aligned} D_{KL}\left(X^{(k)} \middle\| X^{(l)}\right) &\leq 3n \sum_{i=1}^{p} \|\mathbf{P}_{i\cdot}^{(k)} - \mathbf{P}_{i\cdot}^{(l)}\|_2^2 = 3n \sum_{i,j=1}^{p} \left(\mathbf{P}_{ij}^{(k)} - \mathbf{P}_{ij}^{(l)}\right)^2 \\ &= \frac{6n\tau}{p} \sum_{i,j=1}^{p} \left|\mathbf{P}_{ij}^{(k)} - \mathbf{P}_{ij}^{(l)}\right| \leq 6n\tau \cdot \|\mathbf{F}^{(k)} - \mathbf{F}^{(l)}\|_1 \leq \frac{18n\tau^2 l_0 p_0 (r-1)}{p^2}. \end{aligned}$$

Now, by generalized Fano's lemma (see, e.g., [54, 53]),

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

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

$$\begin{aligned} \inf_{\hat{\mathbf{F}}} \sup_{\mathbf{F} \in \mathcal{F}_{p,r}} \left\|\hat{\mathbf{F}} - \mathbf{F}\right\|_1 &\geq \inf_{\hat{\mathbf{F}}} \sup_{\mathbf{F} \in \{\mathbf{F}^{(1)}, \ldots, \mathbf{F}^{(m)}\}} \left\|\hat{\mathbf{F}} - \mathbf{F}\right\|_1 \\ &\geq \frac{p_0 l_0 (r-1)}{2p^2} \cdot \sqrt{\frac{p^2 \cdot \left(\frac{1}{2}\log(m) - \log(2)\right)}{18np_0l_0(r-1) \wedge 1}} \geq c\sqrt{\frac{pr}{n} \wedge 1}. \end{aligned}$$

Then we develop the lower bound for transition matrix estimation. To simplify the notations, it is without loss of generality to assume that p is a multiple of $4(r-1)$. For $1 \leq k \leq m$, let

$$\begin{aligned} \mathbf{P}^{(k)} =& \left[\tfrac{2-\gamma}{p} \cdot \mathbf{1}_{p \times (p/2)} \quad \tfrac{\gamma}{p} \cdot \mathbf{1}_{p \times (p/2)}\right] \\ &+ \frac{\tau(2-\gamma)}{2p} \begin{bmatrix} \mathbf{0}_{(p/2)\times(p/4)} & \mathbf{0}_{(p/2)\times(p/4)} & \mathbf{0}_{(p/2)\times(p/2)} \\ \mathbf{R}^{(k)} \ \cdots \ \mathbf{R}^{(k)} & -\mathbf{R}^{(k)} \ \cdots \ -\mathbf{R}^{(k)} & \mathbf{0}_{(p/4)\times(p/2)} \\ \underbrace{-\mathbf{R}^{(k)} \ \cdots \ -\mathbf{R}^{(k)}}_{l_0} & \underbrace{\mathbf{R}^{(k)} \ \cdots \ \mathbf{R}^{(k)}}_{l_0} & \mathbf{0}_{(p/4)\times(p/2)} \end{bmatrix}, \end{aligned} \tag{42}$$

where $l_0 = p/4(r-1)$, $\mathbf{R}^{(k)} \in \mathbb{R}^{(p/4)\times(r-1)}$, $\mathbf{R}^{(k)}$ is a $(p/4)$-by-$(r-1)$ matrix with i.i.d. Bernoulli random values, $\gamma = \theta_{\min} p$, and τ is some positive value to be determined later. It is easy to see that $\text{rank}(\mathbf{P}^{(k)}) \leq r$. Let

$$\mu' = \left[\frac{2-\gamma}{p} 1_{p/2}^\top \quad \frac{\gamma}{p} 1_{p/2}^\top\right]^\top. \tag{43}$$

Then we can check that $\mu^\top \mathbf{P}^{(k)} = \mu^\top$, $\min_{1\leq i\leq p} \mu'_i \geq \theta_{\min}$, then μ' is the invariant distribution of $\mathbf{P}^{(k)}$, and $\mathbf{P}^{(k)} \in \mathcal{P}_{p,r,\theta_{\min}}$. Similarly as the proof for previous part, there exists fixed matrices $\{\mathbf{R}^{(1)}, \ldots, \mathbf{R}^{(m)}\} \subseteq \{-1,1\}^{(p/4)\times(r-1)}$ such that $m = \lfloor \exp(cp(r-1)) \rfloor$ and

$$\forall 1 \leq k < l \leq m, \quad p(r-1)/4 \leq \left\|\mathbf{R}^{(k)} - \mathbf{R}^{(l)}\right\|_1 \leq 3p(r-1)/4.$$

Then for any $1 \leq k <\leq m$, we further have

$$\left\|\mathbf{P}^{(k)} - \mathbf{P}^{(l)}\right\|_1 = \frac{\tau(2-\gamma)}{2p} \cdot 4l_0 \cdot \|\mathbf{R}^{(k)} - \mathbf{R}^{(l)}\|_1 \geq \frac{\tau(2-\gamma)p}{8}.$$

$$D_{KL}\left(\{X_0^{(k)}, \ldots, X_n^{(k)}\}\middle\|\{X_0^{(k)}, \ldots, X_n^{(k)}\}\right) = n\sum_{i\in[p]} \mu_i \cdot D_{KL}\left(\mathbf{P}_{i\cdot}^{(k)}\middle\|\mathbf{P}_{i\cdot}^{(l)}\right)$$

$$= n\sum_{i=p/2+1}^{p}\sum_{j=1}^{p/2} \frac{\gamma}{p} \cdot \mathbf{P}_{ij}^{(k)} \log\left(\mathbf{P}_{ij}^{(k)}/\mathbf{P}_{ij}^{(l)}\right) = \frac{2n\gamma}{p}\sum_{i=1}^{p/4} \frac{(2-\gamma)}{2} D_{KL}\left(u_i^{(k)}\middle\|u_i^{(l)}\right),$$

where $u_i^{(k)} = \frac{2}{p}1_{p/2} + \frac{\tau}{p}\left[\mathbf{R}_{i\cdot}^{(k)} \ \cdots \ \mathbf{R}_{i\cdot}^{(k)} \ -\mathbf{R}_{i\cdot}^{(k)} \ \cdots \ -\mathbf{R}_{i\cdot}^{(k)}\right]$ is a $(p/2)$-dimensional distribution. Similarly by Lemma 5, we have

$$D_{KL}\left(u_i^{(k)}\middle\|u_i^{(l)}\right) \leq \frac{3p}{2} \cdot l_0\frac{\tau^2}{p^2}\left\|\mathbf{R}_{i\cdot}^{(k)} - \mathbf{R}_{i\cdot}^{(l)}\right\|_2^2 \leq \frac{3 \cdot l_0\tau^2}{p} \cdot \left\|\mathbf{R}_{i\cdot}^{(k)} - \mathbf{R}_{i\cdot}^{(l)}\right\|_1.$$

Thus,

$$D_{KL}\left(\{X_0^{(k)}, \ldots, X_n^{(k)}\}\middle\|\{X_0^{(k)}, \ldots, X_n^{(k)}\}\right)$$

$$= \frac{n\gamma(2-\gamma)}{p}\sum_{i=1}^{p/4}\frac{3l_0\tau^2}{p} \cdot \left\|\mathbf{R}_{i\cdot}^{(k)} - \mathbf{R}_{i\cdot}^{(l)}\right\|_1 \leq \frac{6n\gamma l_0\tau^2}{p^2}\|\mathbf{R}^{(k)} - \mathbf{R}^{(l)}\|_1 \leq \frac{6n\gamma l_0\tau^2}{p^2} \cdot \frac{3p(r-1)}{4} = 18n\tau^2\gamma.$$

Similarly as previous part, we set $\tau = \frac{\frac{1}{2}\log m - \log 2}{18n\gamma} \wedge 1$,

$$\inf_{\hat{\mathbf{P}}}\sup_{\mathbf{F}\in\mathcal{P}_{p,r}(\theta_{\min})}\left\|\hat{\mathbf{P}} - \mathbf{P}\right\|_1 \geq \inf_{\hat{\mathbf{P}}}\sup_{\mathbf{P}\in\{\mathbf{P}^{(1)},\ldots,\mathbf{P}^{(m)}\}}\left\|\hat{\mathbf{P}} - \mathbf{P}\right\|_1$$

$$\geq \frac{\tau(2-\gamma)p}{16} \cdot \left(1 - \frac{18n\tau^2\gamma - \log 2}{\log m}\right) \geq \frac{p}{32} \cdot \sqrt{\frac{cpr}{18np\gamma} \wedge 1}$$

$$\geq cp\left(\sqrt{\frac{pr}{n} \cdot \frac{1}{p\theta_{\min}}} \wedge 1\right),$$

which implies the lower bound for transition matrix estimation. To sum up, we have finished the proof for this theorem. □

5.4 Proof of Theorem 3

By Lemmas 6 and 9, one has

$$\mathbb{P}\left(\|\tilde{\mathbf{F}} - \mathbf{F}\| \geq C\sqrt{\mu_{\max}\tau\log^2(n)/n}\right) \leq n^{-c_0},$$

By Wedin's lemma [49], one has

$$\max\left\{\|\sin\Theta(\hat{\mathbf{U}}_F,\mathbf{U}_F)\|,\|\sin\Theta(\hat{\mathbf{V}}_F,\mathbf{V}_F)\|\right\}\leq\frac{C\sqrt{\mu_{\max}\tau\log^2(n)/n}}{\sigma_r(\mathbf{F})-\sigma_{r+1}(\mathbf{F})}$$

with probability at least $1-n^{-c_0}$. Since $\max\{\|\sin\Theta(\hat{\mathbf{U}}_F,\mathbf{U}_F)\|,\|\sin\Theta(\hat{\mathbf{U}}_F,\mathbf{U}_F)\|\}\leq 1$ by definition,

$$\begin{aligned}&\mathbb{E}\max\left\{\|\sin\Theta(\hat{\mathbf{U}}_F,\mathbf{U}_F)\|,\|\sin\Theta(\hat{\mathbf{V}}_F,\mathbf{V}_F)\|\right\}\\ \leq&\mathbb{E}\max\left\{\|\sin\Theta(\hat{\mathbf{U}}_F,\mathbf{U}_F)\|,\|\sin\Theta(\hat{\mathbf{V}}_F,\mathbf{V}_F)\|\right\}1_Q\\ &+\mathbb{E}\max\left\{\|\sin\Theta(\hat{\mathbf{U}}_F,\mathbf{U}_F)\|,\|\sin\Theta(\hat{\mathbf{V}}_F,\mathbf{V}_F)\|\right\}1_{Q^c}\\ \leq&\frac{C\sqrt{\mu_{\max}\tau\log^2(n)/n}}{\sigma_r(\mathbf{F})-\sigma_{r+1}(\mathbf{F})}+1\cdot\mathbb{P}(Q^c)\leq\frac{C\sqrt{\mu_{\max}\tau\log^2(n)/n}}{\sigma_r(\mathbf{F})-\sigma_{r+1}(\mathbf{F})}\wedge 1.\end{aligned}$$

Since $\sin\Theta(\hat{\mathbf{U}}_F,\mathbf{U}_F)$ and $\sin\Theta(\hat{\mathbf{V}}_F,\mathbf{V}_F)$ are of r-by-r, we additionally have

$$\mathbb{E}\max\left\{\|\sin\Theta(\hat{\mathbf{U}}_F,\mathbf{U}_F)\|_F,\|\sin\Theta(\hat{\mathbf{V}}_F,\mathbf{V}_F)\|_F\right\}\leq\frac{C\sqrt{(rp/n)\cdot p\mu_{\max}\cdot\tau\log^2(n)}}{\sigma_r(\mathbf{F})-\sigma_{r+1}(\mathbf{F})}\wedge\sqrt{r}.$$

Next we consider $\hat{\mathbf{U}}_P$, and $\hat{\mathbf{V}}_P$. Let $\tilde{\mu}$ be the empirical distribution of μ,

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

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

$$\|\tilde{\mu}-\mu\|_\infty\leq C\sqrt{\frac{\mu_{\max}\tau\log^2(n)}{n}}\leq\frac{1}{2}\mu_{\min}.$$

Then

$$\min_i\tilde{\mu}_i\geq\min_i\mu_i-\|\tilde{\mu}-\mu\|_\infty\geq\frac{1}{2}\mu_{\min},\tag{44}$$

and

$$|\mu_i/\tilde{\mu}_i-1|=\frac{|\mu_i-\tilde{\mu}_i|}{\tilde{\mu}_i}\leq 2\mu_{\min}^{-1}\cdot C\sqrt{\frac{\mu_{\max}\tau\log^2(n)}{n}}.\tag{45}$$

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

$$\begin{aligned}
\left\|\tilde{\mathbf{P}} - \mathbf{P}\right\| &= \left\|\tilde{\mu}^{-1}\tilde{\mathbf{F}} - \mathrm{diag}(\mu)^{-1}\mathbf{F}\right\| \\
&\leq \left\|\tilde{\mu}^{-1}(\tilde{\mathbf{F}} - \mathbf{F})\right\| + \left\|\left(\mathrm{diag}(\mu)^{-1} - \mathrm{diag}(\tilde{\mu})^{-1}\right)\mathbf{F}\right\| \\
&\leq \left\|\tilde{\mu}^{-1}\right\| \cdot \|\tilde{\mathbf{F}} - \mathbf{F}\| + \|\mathrm{diag}(\mu/\tilde{\mu}) - I\| \cdot \|\mathrm{diag}(\mu)\mathbf{F}\| \\
&\leq \left(\min_i \tilde{\mu}_i\right)^{-1} \cdot \|\tilde{\mathbf{F}} - \mathbf{F}\| + \max_{ij} |\mu_i/\tilde{\mu}_i - 1| \cdot \|\mathbf{P}\| \\
&\overset{(44)(45)}{\leq} C\mu_{\min}^{-1}\sqrt{\frac{\mu_{\max}\tau\log^2(n)}{n}} + C\mu_{\min}^{-1}\sqrt{\frac{\mu_{\max}\tau\log^2(n)}{n}}\|\mathbf{P}\|.
\end{aligned}$$

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

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

Finally, by Wedin's perturbation bound, we have

$$\max\left\{\|\sin\Theta(\hat{\mathbf{U}}_P, \mathbf{U}_P)\|, \|\sin\Theta(\hat{\mathbf{V}}_P, \mathbf{V}_P)\|\right\} \leq \frac{C\|\mathbf{P}\| \cdot \sqrt{(p/n)\cdot\mu_{\max}/(p\mu_{\min}^2)\cdot\tau\log^2(n)}}{\sigma_r(\mathbf{P}) - \sigma_{r+1}(\mathbf{P})}$$

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

$$\mathbb{E}\max\left\{\|\sin\Theta(\hat{\mathbf{U}}_P, \mathbf{U}_P)\|_F, \|\sin\Theta(\hat{\mathbf{V}}_P, \mathbf{V}_P)\|_F\right\} \leq \frac{C\|\mathbf{P}\| \cdot \sqrt{(rp/n)\cdot\mu_{\max}/(p\mu_{\min}^2)\cdot\tau\log^2(n)}}{\sigma_r(\mathbf{P}) - \sigma_{r+1}(\mathbf{P})}.$$

□

5.5 Proof of Theorem 4

We focus on the proof for $\mathbf{U}_P$, $\mathbf{U}_F$ and $r = 2$ first, 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 $\mathcal{P}_{p,r}$. To be specific, let

$$\begin{aligned}
\mathbf{P}^{(k)} &= \frac{1}{p}1_p 1_p^\top \\
&+ \frac{\lambda_P}{p}\begin{bmatrix}
\overbrace{\begin{matrix} 1_{p/4} & \cdots & 1_{p/4} \end{matrix}}^{p/2} & \overbrace{\begin{matrix} -1_{p/4} & \cdots & -1_{p/4} \end{matrix}}^{p/2} \\
\begin{matrix} -1_{p/4} & \cdots & -1_{p/4} \end{matrix} & \begin{matrix} 1_{p/4} & \cdots & 1_{p/4} \end{matrix} \\
\begin{matrix} \delta\beta^{(k)} & \cdots & \delta\beta^{(k)} \end{matrix} & \begin{matrix} -\delta\beta^{(k)} & \cdots & -\delta\beta^{(k)} \end{matrix} \\
\begin{matrix} -\delta\beta^{(k)} & \cdots & -\delta\beta^{(k)} \end{matrix} & \begin{matrix} \delta\beta^{(k)} & \cdots & \delta\beta^{(k)} \end{matrix}
\end{bmatrix}.
\end{aligned}$$

Here $\{\beta^{(k)}\}_{k=1}^m$ are m copies of i.i.d. Rademacher $(p/4)$-dimensional random vectors, $0 < \delta \le 1$ and m are constants to be determined later. It is not hard to check that the singular value decomposition of $\mathbf{P}^{(k)}$ can be written as

$$\mathbf{P}^{(k)} = \left(\frac{1}{\sqrt{p}}1_p\right)\left(\frac{1}{\sqrt{p}}1_p\right)^\top + \sigma^{(k)}u^{(k)}(v^{(k)})^\top, \tag{46}$$

where

$$\sigma^{(k)} = \frac{\lambda_P}{p}\sqrt{\frac{p^2}{2}(1+\delta^2)} \le \lambda_P,$$

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

Namely, $\mathbf{P}^{(k)} \in \mathcal{P}_{p,\lambda_P}$, $k = 1,\ldots,m$. Since $\lambda_P \le 1/2$, $1/2 \le \mathbf{P}_{ij}^{(k)} \le 3/2$.

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

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

then

$$\mathbb{P}\left(\exists k \ne l, \quad \text{s.t.} \quad \frac{1}{p/4}\left|(\beta^{(k)})^\top\beta^{(l)}\right| \ge \frac{1}{2}\right) \le 2\cdot\frac{m(m-1)}{2}\exp\left(-p/28\right) < m^2\exp\left(-p/28\right). \tag{47}$$

If we set $m = \lceil \exp(-p/56) \rceil$, the probability in the right hand side of (47) is strictly less than 1, which means there must exists fixed $\left\{\beta^{(k)}\right\}_{k=1}^m$ such that

$$|(\beta^{(k)})^\top\beta^{(l)}| < p/8, \quad \forall 1 \le k < l \le m. \tag{48}$$

For the rest of the proof we assume (48) always hold. Now, for any $k \ne l$,

$$\begin{aligned}
&\left\|\sin\Theta\left(\mathbf{U}^{(P,k)},\mathbf{U}^{(P,l)}\right)\right\| = \|\sin\Theta(u^{(k)},u^{(l)})\| = \sqrt{1-\left((u^{(k)})^\top v^{(l)}\right)^2} \\
=&\sqrt{1-\left(\frac{p/2+2\delta^2(\beta^{(k)})^\top\beta^{(l)}}{p/2+\delta^2 p/2}\right)^2} \ge \sqrt{1-\left(\frac{p/2+\delta^2 p/4}{p/2+\delta^2 p/2}\right)^2} \\
=&\sqrt{\frac{\delta^2 p/4}{p/2+\delta^2 p/2}\cdot\left(1+\frac{p/2+\delta^2 p/4}{p/2+\delta^2 p/2}\right)} \ge \sqrt{\frac{\delta^2 p/4}{p}} = \frac{\delta}{2}.
\end{aligned}$$

It is easy to verify that $\frac{1}{p}1_p$ is a stationary distribution for all $\mathbf{P}^{(k)}$. Now for each $1 \le k \le m$, 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

$$\begin{aligned} D_{KL}\left(X^{(k)}||X^{(l)}\right) =& \frac{n}{p}\sum_{i=1}^{p} D_{KL}\left(\mathbf{P}_{i\cdot}^{(k)}||\mathbf{P}_{i\cdot}^{(l)}\right) \\ \overset{\text{Lemma 5}}{\leq} & 3n\sum_{i=1}^{p}\|\mathbf{P}_{i\cdot}^{(k)} - \mathbf{P}_{i\cdot}^{(l)}\|_2^2 \leq 3n\cdot\frac{\lambda_P^2}{p^2}\cdot\left(\delta^2\frac{p^2}{2}\right) \leq \frac{3n\lambda_P^2\delta^2}{2}. \end{aligned}$$

Finally we set $\delta = \sqrt{\frac{2\log(m)-\log 2}{3n\lambda_P^2}}$. By generalized Fano's lemma,

$$\begin{aligned} \inf_{\tilde{\mathbf{U}}_P}\sup_{\mathbf{P}\in\{\mathbf{P}^{(1)},\ldots,\mathbf{P}^{(m)}\}}\left\|\sin\Theta(\tilde{\mathbf{U}}_P,\mathbf{U}_P)\right\| \geq & \frac{\delta}{2}\left(1-\frac{3n\lambda_P^2\delta^2+\log 2}{\log m}\right) \geq \frac{\delta}{4} \\ \geq & c\frac{\sqrt{p/n}}{\lambda_P}, \end{aligned}$$

which has finished the proof for the theorem. □

5.6 Proof of Theorem 5

Denote $\mathbf{Z} = \tilde{\mathbf{F}} - \mathbf{F}$. Based on the intermediate step in the proof of Theorem 1, we have

$$\|\mathbf{Z}\| = \left\|\tilde{\mathbf{F}} - \mathbf{F}\right\| \leq C\sqrt{\frac{\mu_{\max}\tau\log^2(n)}{n}}.$$

with probability at least $1-n^{-c}$. Provided that $\sigma_r(\mathbf{F}_1) > \|\mathbf{F}_2\|$, the one-sided perturbation bound (Proposition 1 in [8]) yields,

$$\left\|\sin\Theta\left(\hat{\mathbf{V}}_F,\mathbf{V}_F\right)\right\| \leq \frac{\sigma_r(\tilde{\mathbf{F}}\mathbf{V}_F^\top)\cdot\|P_{(\tilde{\mathbf{F}}\mathbf{V})}\tilde{\mathbf{F}}\mathbf{V}_\perp\|}{\sigma_r^2(\tilde{\mathbf{F}}\mathbf{V}_F^\top)-\sigma_{r+1}^2(\tilde{\mathbf{F}})}$$

Here, $P_{\mathbf{A}} = \mathbf{A}(\mathbf{A}^\top\mathbf{A})^\dagger\mathbf{A}^\top$ is the projection for any matrix $\mathbf{A}$,

$$\sigma_r(\tilde{\mathbf{F}}\mathbf{V}_F) \geq \sigma_r(\mathbf{F}\mathbf{V}_F) - \|\mathbf{Z}\| = \sigma_r(\mathbf{F}_1\mathbf{V}_F + \mathbf{F}_2\mathbf{V}_F) - \|\mathbf{Z}\| \geq \sigma_r(\mathbf{F}_1) - \|\mathbf{Z}\|,$$

$$\sigma_{r+1}(\tilde{\mathbf{F}}) \leq \sigma_{r+1}(\mathbf{F}) + \|\mathbf{Z}\| = \min_{\text{rank}(K)\leq r}\|\mathbf{F}-K\| + \|\mathbf{Z}\| \leq \|\mathbf{F}-\mathbf{F}_1\| \leq \|\mathbf{F}_2\| + \|\mathbf{Z}\|,$$

$$
\begin{aligned}
\left\|P_{(\tilde{\mathbf{F}}\mathbf{V}_F)}\tilde{\mathbf{F}}\mathbf{V}_{F,\perp}\right\| =& \left\|P_{(\tilde{\mathbf{F}}\mathbf{V}_F)}P_{\mathbf{U}_F}\tilde{\mathbf{F}}\mathbf{V}_{F,\perp} + P_{(\tilde{\mathbf{F}}\mathbf{V}_F)}P_{\mathbf{U}_{F,\perp}}\tilde{\mathbf{F}}\mathbf{V}_{F,\perp}\right\| \\
\leq & \left\|P_{(\tilde{\mathbf{F}}\mathbf{V}_F)}\mathbf{U}_F\mathbf{U}_F^\top\tilde{\mathbf{F}}\mathbf{V}_{F,\perp}\right\| + \left\|P_{(\tilde{\mathbf{F}}\mathbf{V}_F)}\mathbf{U}_{F,\perp}\mathbf{U}_{F,\perp}^\top\tilde{\mathbf{F}}\mathbf{V}_{F,\perp}\right\| \\
\leq & \left\|\mathbf{U}_F^\top\tilde{\mathbf{F}}\mathbf{V}_{F,\perp}\right\| + \left\|(\tilde{\mathbf{F}}\mathbf{V}_F)\left[(\tilde{\mathbf{F}}\mathbf{V}_F)^\top(\tilde{\mathbf{F}}\mathbf{V}_F)\right]^{-1}(\tilde{\mathbf{F}}\mathbf{V}_F)^\top\mathbf{U}_{F,\perp}\mathbf{U}_{F,\perp}^\top\tilde{\mathbf{F}}\mathbf{V}_{F,\perp}\right\| \\
\leq & \left\|\mathbf{U}_F^\top(\mathbf{F}_1+\mathbf{F}_2+\mathbf{Z})\mathbf{V}_{F,\perp}\right\| + \frac{1}{\sigma_{\min}(\tilde{\mathbf{F}}\mathbf{V}_F)}\left\|(\tilde{\mathbf{F}}\mathbf{V}_F)^\top\mathbf{U}_{F,\perp}\right\|\cdot\left\|\mathbf{U}_{F,\perp}^\top\tilde{\mathbf{F}}\mathbf{V}_{F,\perp}\right\| \\
\leq & \left\|\mathbf{U}_F^\top(\mathbf{F}_1+\mathbf{F}_2+\mathbf{Z})\mathbf{V}_{F,\perp}\right\| \\
& + \frac{1}{\sigma_{\min}((\mathbf{F}_1+\mathbf{F}_2+\mathbf{Z})\mathbf{V}_F)}\left\|\mathbf{U}_{F,\perp}^\top(\mathbf{F}_1+\mathbf{F}_2+\mathbf{Z})\mathbf{V}_F\right\|\cdot\left\|\mathbf{U}_{F,\perp}^\top(\mathbf{F}_1+\mathbf{F}_2+\mathbf{Z})\mathbf{V}_{F,\perp}\right\| \\
& (\text{since } \mathbf{U}_F^\top\mathbf{F}_2=0,\ \mathbf{F}_1\mathbf{V}_{F,\perp}=0,\ \mathbf{U}_{F,\perp}^\top\mathbf{F}_1=0) \\
=& \|\mathbf{U}_F^\top\mathbf{Z}\mathbf{V}_{F,\perp}\| + \frac{1}{\sigma_{\min}((\mathbf{F}_1-\mathbf{Z})\mathbf{V}_F)}\|\mathbf{U}_{F,\perp}^\top(\mathbf{F}_2+\mathbf{Z})\mathbf{V}_F\|\cdot\|\mathbf{U}_{F,\perp}^\top(\mathbf{F}_2+\mathbf{Z})\mathbf{V}_{F,\perp}\| \\
\leq & \|\mathbf{Z}\| + \frac{(\|\mathbf{F}_2\|+\|\mathbf{Z}\|)^2}{\sigma_{\min}(\mathbf{F}_1)-\|\mathbf{Z}\|}.
\end{aligned}
$$

Thus,

$$
\|\sin\Theta(\hat{\mathbf{V}}_F,\mathbf{V}_F)\| \leq \left(\frac{\|\mathbf{Z}\|(\sigma_r(\mathbf{F}_1)+\|\mathbf{Z}\|)+(\|\mathbf{F}_2\|+\|\mathbf{Z}\|)^2}{(\sigma_r(\mathbf{F}_1)-\|\mathbf{Z}\|)^2-(\|\mathbf{F}_2\|+\|\mathbf{Z}\|)^2}\right)\wedge 1.
$$

When $\sigma_r(\mathbf{F}_1)-\|\mathbf{F}_2\|\geq 3\|\mathbf{Z}\|$,

$$
\begin{aligned}
&(\sigma_r(\mathbf{F}_1)-\|\mathbf{Z}\|)^2-(\|\mathbf{F}_2\|+\|\mathbf{Z}\|)^2 = (\sigma_r(\mathbf{F}_1)-\|\mathbf{F}_2\|-2\|\mathbf{Z}\|)(\sigma_r(\mathbf{F}_1)+\|\mathbf{F}_2\|) \\
\leq & \frac{1}{3}(\sigma_r(\mathbf{F}_1)-\|\mathbf{F}_2\|)(\sigma_r(\mathbf{F}_1)+\|\mathbf{F}_2\|) = \frac{1}{3}\left(\sigma_r^2(\mathbf{F}_1)-\|\mathbf{F}_2\|^2\right)
\end{aligned}
$$

$$
\begin{aligned}
&\|\mathbf{Z}\|(\sigma_r(\mathbf{F}_1)+\|\mathbf{Z}\|)+(\|\mathbf{F}_2\|+\|\mathbf{Z}\|)^2 = \|\mathbf{Z}\|(\sigma_r(\mathbf{F}_1)+2\|\mathbf{F}_2\|+2\|\mathbf{Z}\|)+\|\mathbf{F}_2\|^2 \\
\leq & 2\|\mathbf{Z}_2\|(\sigma_r(\mathbf{F}_1)+\|\mathbf{F}_2\|)+\|\mathbf{F}_2\|^2.
\end{aligned}
$$

Thus, when $\sigma_r(\mathbf{F}_1)-\|\mathbf{F}_2\|\geq 3\|\mathbf{Z}\|$,

$$
\begin{aligned}
\|\sin\Theta(\hat{\mathbf{V}}_F,\mathbf{V}_F)\| \leq & \left(\frac{6\|\mathbf{Z}\|}{\sigma_r(\mathbf{F}_1)-\|\mathbf{F}_2\|}+\frac{3\|\mathbf{F}_2\|^2}{\sigma_r^2(\mathbf{F}_1)-\|\mathbf{F}_2\|^2}\right)\wedge 1 \\
\leq & \left(\frac{6\|\mathbf{Z}\|}{\sigma_r(\mathbf{F}_1)-\|\mathbf{F}_2\|}+\frac{6\|\mathbf{F}_2\|^2}{\sigma_r^2(\mathbf{F}_1)}\right)\wedge 1 \\
\leq & \left(\frac{C\sqrt{\mu_{\max}\tau\log^2(n)/n}}{\sigma_r(\mathbf{F}_1)-\|\mathbf{F}_2\|}+\frac{C\|\mathbf{F}_2\|^2}{\sigma_r^2(\mathbf{F}_1)}\right)\wedge 1
\end{aligned} \tag{49}
$$

On the other hand when $\sigma_r(\mathbf{F}_1) - \|\mathbf{F}_2\| < 3\|\mathbf{Z}\|$, $\frac{6\|\mathbf{Z}\|}{\sigma_r(\mathbf{F}_1)-\|\mathbf{Z}\|} \geq 1$ then (49) still hold. Similarly, one can show

$$\begin{aligned}\left\|\sin\Theta\left(\hat{\mathbf{U}}_F, \mathbf{U}_F\right)\right\| \leq & \frac{C\|\mathbf{Z}\|}{\sigma_r(\mathbf{F}_1) - \|\mathbf{F}_2\|} + \frac{C\|\mathbf{F}_2\|^2}{\sigma_r^2(\mathbf{F}_1)} \\ \leq & \left(\frac{C}{\sigma_r(\mathbf{F}_1) - \|\mathbf{F}_2\|}\sqrt{\frac{\mu_{\max}\tau\log^2(n)}{n}} + \frac{C\|\mathbf{F}_2\|}{\sigma_r(\mathbf{F}_1)}\right) \wedge 1.\end{aligned}$$

Next we consider the second part on the misclassification rate of r-means. By Proposition 1, $\mathbf{V}_F$ is piece-wise linear with respect to partitions $G_1, \ldots, G_r$. Then we write $\mathbf{V}_F = MX$, where M and X are the membership and probability matrices, respectively:

$$\mathbf{M} \in \mathbb{R}^{p\times r}, \quad \mathbf{M}_{ij} = \begin{cases} 1, & \text{i-th state} \in G_j \\ 0, & \text{i-th state} \notin G_j. \end{cases} \tag{50}$$

$$\mathbf{X} \in \mathbb{R}^{r\times r}, \quad \mathbf{X} \geq 0. \tag{51}$$

Since $\mathbf{V}_F = \mathbf{MX}$ is an orthogonal matrix, $\mathbf{X}^\top\mathbf{M}^\top\mathbf{MX} = \mathbf{I}_r$, note that $\mathbf{MM} = \mathrm{diag}(n_1, \ldots, n_r)$, we have $\mathbf{X}^\top\mathrm{diag}(n_1, \ldots, n_r)\mathbf{X} = \mathbf{I}_r$, then $\mathrm{diag}(n_1^{1/2}, \ldots, n_r^{1/2})\mathbf{X}$ is an orthogonal matrix, which implies

$$\|\mathbf{X}_{[s,:]}\|_2^2 = \frac{1}{|G_s|}, \quad \mathbf{X}_{[s,:]}^\top\mathbf{X}_{[t,:]} = (|G_s|\cdot|G_t|)^{-1/2}\cdot 1_{\{s=t\}}, \quad \forall 1 \leq i, j \leq p.$$

This implies for any two states i, j, if $i \in G_s, j \in G_t$, then

$$\begin{aligned}\|(\mathbf{V}_F)_{[i,:]} - (\mathbf{V}_F)_{[j,:]}\|_2^2 =& \|\mathbf{X}_{[s,:]} - \mathbf{X}_{[t,:]}\|_2^2 = \|\mathbf{X}_{[s,:]}\|_2^2 + \|\mathbf{X}_{[t,:]}\|_2^2 + 2\mathbf{X}_{[s,:]}^\top\mathbf{X}_{[t,:]} \\ =& \frac{1}{|G_s|} + \frac{1}{|G_t|} - 2\left(|G_s|\cdot|G_t|\right)^{-1/2}\cdot 1_{\{s=t\}} \\ =& \begin{cases} 0, & i \text{ and } j \text{ belong to the same group;} \\ \frac{1}{|G_s|} + \frac{1}{|G_t|}, & \text{otherwise.} \end{cases}\end{aligned}$$

Next, we apply the approximation for r-means (see, e.g. Lemma 5.3 in [24]), we have

$$\begin{aligned}\sum_{i=1}^r |S_i|\cdot\frac{1}{|G_i|} \leq & 4(4+2\varepsilon)\min_{O\in\mathbb{O}_r}\|\hat{\mathbf{V}}_F - \mathbf{V}O\|_F^2 \leq Cr\left\|\sin\Theta\left(\hat{\mathbf{V}}_F, \mathbf{V}_F\right)\right\|^2 \\ \leq & C\left(\frac{r\mu_{\max}\tau\log^2(n)}{n(\sigma_r(\mathbf{F}_1) - \|\mathbf{F}_2\|)^2} + \frac{Cr\|\mathbf{F}_2\|^4}{\sigma_r^4(\mathbf{F}_1)}\right) \wedge r.\end{aligned}$$

which has finished the proof for this theorem. □

5.7 Proof of Theorem 6 and Corollary 3

In order to show this result, we introduce the following lemma to characterize the ℓ_2 error of $\tilde{\mathbf{F}}$.

Lemma 1. *Suppose* $\mathbf{K}$ *is an order-*$\lfloor\beta\rfloor$ *smooth kernel. Recall that the kernel estimation for* $\mathbf{F}$ *is*

$$\tilde{\mathbf{F}}(a,b) = \sum_{i=1}^{n} \frac{1}{h}\mathbf{K}\left(\frac{a-\mathbf{X}_{k-1}}{h}, \frac{b-\mathbf{X}_k}{h}\right).$$

Provided that $n \geq C\log(n)\tau$*,*

$$\mathbb{E}\left\|\tilde{\mathbf{F}}-\mathbf{F}\right\|_F^2 = \mathbb{E}\int_{\mathbb{R}^2}\left(\tilde{\mathbf{F}}(a,b)-\mathbf{F}(a,b)\right)^2 dadb \leq C\left(\frac{\tau\log(n)}{n}\right)^{\frac{4\beta}{4\beta+2}}.$$

The proof of Lemma 1 is postponed to the appendix. By Lemmas 1 and 4, we have

$$\mathbb{E}\left\|\hat{\mathbf{F}}-\mathbf{F}\right\|_F^2 \leq C\mathbb{E}\left\|\tilde{\mathbf{F}}-\mathbf{F}\right\|_F^2 \leq C\left(\frac{\tau\log(n)}{n}\right)^{\frac{4\beta}{4\beta+2}}.$$

By Cauchy-Schwarz inequality,

$$\begin{aligned}&\mathbb{E}\left\|\hat{\mathbf{F}}-\mathbf{F}\right\|_{\ell_1} = \mathbb{E}\int\left|\hat{\mathbf{F}}(a,b)-\mathbf{F}(a,b)\right| dadb\\ \leq& C\left(\int\left(\hat{\mathbf{F}}(a,b)-\mathbf{F}(a,b)\right) dadb\right)^{1/2} \leq C\left(\frac{\tau\log(n)}{n}\right)^{\frac{2\beta}{4\beta+2}}.\end{aligned}$$

By Lemma 3,

$$\mathbb{E}\|\hat{\mathbf{P}}-\mathbf{P}\|_{\ell_1} = \mathbb{E}\int\|\hat{\mathbf{P}}(a,\cdot)-\mathbf{P}(a,\cdot)\|_{\ell_1}da \leq \mathbb{E}\int\frac{2\|\hat{\mathbf{F}}-\mathbf{F}\|_{\ell_1}}{\mu_{\min}}da \leq \frac{C}{\mu_{\min}}\left(\frac{\tau\log(n)}{n}\right)^{\frac{2\beta}{4\beta+2}}.$$

By Wedin's perturbation bound [49],

$$\begin{aligned}&\mathbb{E}\left\{\left\|\sin\Theta\left(\hat{\mathbf{U}}_F,\mathbf{U}\right)\right\|_F, \left\|\sin\Theta\left(\hat{\mathbf{V}}_F,\mathbf{V}\right)\right\|\right\} \leq \frac{\|\tilde{\mathbf{F}}-\mathbf{F}\|_F}{\sigma_r(\mathbf{F})-\sigma_{r+1}(\mathbf{F})}\wedge\sqrt{r}\\ \leq& \frac{C}{\sigma_r(\mathbf{F})-\sigma_{r+1}(\mathbf{F})}\left(\frac{\tau\log(n)}{n}\right)^{\frac{2\beta}{4\beta+2}}\wedge\sqrt{r}.\end{aligned}$$

□

References

[1] Theodore W Anderson and Leo A Goodman. Statistical inference about markov chains. *The Annals of Mathematical Statistics*, pages 89–110, 1957.

[2] Ery Arias-Castro, Bruno Pelletier, and Pierre Pudlo. The normalized graph cut and cheeger constant: from discrete to continuous. *Advances in Applied Probability*, 44(4):907–937, 2012.

[3] Austin R Benson, David F Gleich, and Lek-Heng Lim. The spacey random walk: A stochastic process for higher-order data. *SIAM Review*, 59(2):321–345, 2017.

[4] Dimitri P Bertsekas, Dimitri P Bertsekas, Dimitri P Bertsekas, and Dimitri P Bertsekas. *Dynamic programming and optimal control*, volume 1. Athena scientific Belmont, MA, 1995.

[5] Dimitri P Bertsekas and John N Tsitsiklis. Neuro-dynamic programming: an overview. In *Decision and Control, 1995., Proceedings of the 34th IEEE Conference on*, volume 1, pages 560–564. IEEE, 1995.

[6] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[7] Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

[8] T Tony Cai and Anru Zhang. Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *The Annals of Statistics*, to appear, 2017.

[9] Emmanuel J Candes, Carlos A Sing-Long, and Joshua D Trzasko. Unbiased risk estimates for singular value thresholding and spectral estimators. *IEEE transactions on signal processing*, 61(19):4643–4657, 2013.

[10] Sourav Chatterjee et al. Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214, 2015.

[11] Jeff Cheeger. A lower bound for the smallest eigenvalue of the laplacian. *Problems in analysis*, pages 195–199, 1970.

[12] Jingchun Chen and Bo Yuan. Detecting functional modules in the yeast protein–protein interaction network. *Bioinformatics*, 22(18):2283–2290, 2006.

[13] Yuxin Chen, Jianqing Fan, Cong Ma, and Kaizheng Wang. Spectral method and regularized mle are both optimal for top-k ranking. *arXiv preprint arXiv:1707.09971*, 2017.

[14] Fan RK Chung. *Spectral graph theory*, volume 92. American Mathematical Soc., 1997.

[15] David Donoho, Matan Gavish, et al. Minimax risk of matrix denoising by singular value thresholding. *The Annals of Statistics*, 42(6):2413–2440, 2014.

[16] Weinan E, Tiejun Li, and Eric Vanden-Eijnden. Optimal partition and effective dynamics of complex networks. *Proceedings of the National Academy of Sciences*, 105(23):7907–7912, 2008.

[17] David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

[18] James Douglas Hamilton. *Time series analysis*, volume 2. Princeton university press Princeton, 1994.

[19] Daniel J Hsu, Aryeh Kontorovich, and Csaba Szepesvári. Mixing time estimation in reversible markov chains from a single sample path. In *Advances in neural information processing systems*, pages 1459–1467, 2015.

[20] Jiantao Jiao, Yanjun Han, and Tsachy Weissman. Minimax estimation of the l 1 distance. In *Information Theory (ISIT), 2016 IEEE International Symposium on*, pages 750–754. IEEE, 2016.

[21] Jiashun Jin and Wanjie Wang. Influential features pca for high dimensional clustering. *The Annals of Statistics*, 44(6):2323–2359, 2016.

[22] Björn H Junker and Falk Schreiber. *Analysis of biological networks*, volume 2. John Wiley & Sons, 2011.

[23] John G Kemeny, James Laurie Snell, et al. *Finite markov chains*, volume 356. van Nostrand Princeton, NJ, 1960.

[24] Jing Lei, Alessandro Rinaldo, et al. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1):215–237, 2015.

[25] David Asher Levin, Yuval Peres, and Elizabeth Lee Wilmer. *Markov chains and mixing times*. American Mathematical Soc., 2009.

[26] Yu Liu, Chaogui Kang, Song Gao, Yu Xiao, and Yuan Tian. Understanding intra-urban trip patterns from taxi trajectory data. *Journal of geographical systems*, 14(4):463–483, 2012.

[27] Rosario N Mantegna. Hierarchical structure in financial markets. *The European Physical Journal B-Condensed Matter and Complex Systems*, 11(1):193–197, 1999.

[28] Nathan McNew. The eigenvalue gap and mixing time. 2011.

[29] Marina Meila and Jianbo Shi. A random walks view of spectral segmentation. 2001.

[30] Milena Mihail. Conductance and convergence of markov chains-a combinatorial treatment of expanders. In *Foundations of computer science, 1989., 30th annual symposium on*, pages 526–531. IEEE, 1989.

[31] Andrew W Moore. Variable resolution dynamic programming: Efficiently learning action maps in multivariate real-valued state-spaces. In *Machine Learning Proceedings 1991*, pages 333–337. Elsevier, 1991.

[32] Sahand Negahban, Sewoong Oh, and Devavrat Shah. Rank centrality: Ranking from pairwise comparisons. *Operations Research*, 65(1):266–287, 2016.

[33] Sridhar Nerur, Riyaz Sikora, George Mangalaraj, and VenuGopal Balijepally. Assessing the relative influence of journals in a citation network. *Communications of the ACM*, 48(11):71–74, 2005.

[34] Mark EJ Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2):404–409, 2001.

[35] Mark EJ Newman. Spectral methods for community detection and graph partitioning. *Physical Review E*, 88(4):042822, 2013.

[36] Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pages 849–856, 2002.

[37] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[38] Phillip James Edwin Peebles. *The large-scale structure of the universe.* Princeton university press, 1980.

[39] Clara Pizzuti. Ga-net: A genetic algorithm for community detection in social networks. *Parallel problem solving from nature–PPSN X*, pages 1081–1090, 2008.

[40] Zhiyuan Ren and BH Krogh. State aggregation in markov decision processes. In *Decision and Control, 2002, Proceedings of the 41st IEEE Conference on*, volume 4, pages 3819–3824. IEEE, 2002.

[41] Karl Rohe, Sourav Chatterjee, and Bin Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, pages 1878–1915, 2011.

[42] Andrey A Shabalin and Andrew B Nobel. Reconstruction of a low-rank matrix in the presence of gaussian noise. *Journal of Multivariate Analysis*, 118:67–76, 2013.

[43] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000.

[44] Satinder P Singh, Tommi Jaakkola, and Michael I Jordan. Reinforcement learning with soft state aggregation. In *Advances in neural information processing systems*, pages 361–368, 1995.

[45] Ho Steinhaus. The problem of estimation. *The Annals of Mathematical Statistics*, 28(3):633–648, 1957.

[46] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

[47] Joel A Tropp. Freedmans inequality for matrix martingales. *Electron. Commun. Probab*, 16:262–270, 2011.

[48] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008.

[49] Per-Ake Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.

[50] Hermann Weyl. Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung). *Mathematische Annalen*, 71(4):441–479, 1912.

[51] Maciej Wilczyński. Minimax estimation for the multinomial and multivariate hypergeometric distributions. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 128–132, 1985.

[52] Lin F Yang, Vladimir Braverman, Tuo Zhao, and Mengdi Wang. Dynamic partition of complex networks. *arXiv preprint arXiv:1705.07881*, 2017.

[53] Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, pages 1564–1599, 1999.

[54] Bin Yu. Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.

[55] Shan Zeng, Rui Huang, Zhen Kang, and Nong Sang. Image segmentation using spectral clustering of gaussian mixture models. *Neurocomputing*, 144:346–356, 2014.

6 Technical Lemmas

We collect the technical lemmas for the main results in this section. The first Lemma 2 demonstrates a sufficient and necessary condition for being transition and frequency matrices of some ergoic Markov chain.

Lemma 2 (Basic properties of transition and frequency matrix for ergodic Markov process)**.** $\mathbf{P}, \mathbf{F} \in \mathbb{R}^{p\times p}$ *are the transition matrix and frequency matrix of some ergodic Markov process if and only if*

$$\mathbf{P} \in \mathcal{P}, \quad \mathcal{P} = \left\{ \mathbf{P}: \begin{array}{l} 0 \leq \mathbf{P}_{ij} \leq 1; \forall 1 \leq i \leq p, \sum_{j=1}^p \mathbf{P}_{ij} = 1, \\ \forall I \subseteq \{1, \ldots, p\}, \mathbf{P}_{[I, I^c]} \neq 0 \end{array} \right\}, \tag{52}$$

$$\textit{and} \quad \mathbf{F} \in \mathcal{F}, \quad \mathcal{F} = \left\{ \mathbf{F} \in \mathbb{R}^{p\times p}: \begin{array}{l} \mathbf{F}1_p = \mathbf{F}^\top 1_p, \quad 1_p^\top \mathbf{F} 1_p = 1 \\ \forall I \subseteq \{1, \ldots, p\}, \mathbf{F}_{[I, I^c]} \neq 0 \end{array} \right\}. \tag{53}$$

Proof of Lemma 2. We first 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 $\mathbf{F} = \mathrm{diag}(\mu)\mathbf{P}$, where μ and $\mathbf{P}$ are the corresponding stationary distribution and stochastic matrix. Then

$$\mathbf{F}1_p = \mathrm{diag}(\mu)\mathbf{P}1_p = \mathrm{diag}(\mu)1_p = \mu;$$

$$\mathbf{F}^\top 1_p = \mathbf{P}^\top \mathrm{diag}(\mu) 1_p = \mathbf{P}^\top \mu = \mu = \mathbf{F}1_p$$

$$1_p^\top \mathbf{F} 1_p = 1_p^\top \mu = 1.$$

Here we used the fact that $\mu^\top \mathbf{P} = \mu^\top$ and $\mathbf{P}1_p = 1_p$. Next, since the Markov is ergodic, $\mu_i > 0$ for any i. Thus for any $I \subseteq \{1, \ldots, p\}$, $\mathbf{F}_{[I, I^c]} = \mathrm{diag}(\mu_I) \cdot \mathbf{P}_{[I, I^c]} \neq 0$. This implies $\mathbf{F} \in \mathcal{F}$.

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

$$\begin{aligned} & 1_p^\top \mu = 1_p^\top \mathbf{F} 1_p = 1, \quad \mathbf{P}_{ij} \geq 0, \quad \mathbf{P}1_p = \mathrm{diag}(\mu^{-1})\mathbf{F}1_p = \mathrm{diag}(\mu^{-1})\mu = 1_p, \\ & \mu^\top \mathbf{P} = \mu^\top \mathrm{diag}(\mu)\mathbf{F} = 1_p^\top \mathbf{F} = (\mathbf{F}^\top 1_p)^\top = (\mathbf{F}1_p)^\top = \mu, \\ & \forall I \subseteq \{1, \ldots, p\}, \mathbf{P}_{[I, I^c]} = \mathrm{diag}(\mu_I^{-1}) \cdot \mathbf{F}_{[I, I^c]} \neq 0. \end{aligned} \tag{54}$$

By comparing above properties with the definition of ergodic transition matrix (52), we can see $\mathbf{F}$ is indeed a frequency matrix of some ergodic Markov process.

The proof for the transition matrix (52) is similar and more straightforward. Thus, we have finished the proof of this lemma. □

The next Lemma 3 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 3. *Suppose $u, v \neq 0$ are two vectors of the same dimension, then*

$$\left\| \frac{u}{\|u\|_1} - \frac{v}{\|v\|_1} \right\|_1 \leq \frac{2\|u - v\|_1}{\max\{\|u\|_1, \|v\|_1\}}. \tag{55}$$

If one replace u, v by univariate function, the similar result still holds.

Proof of Lemma 3.

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

Similarly, $\left\| \frac{u}{\|u\|_1} - \frac{v}{\|v\|_1} \right\|_1 \leq \frac{2\|u-v\|_1}{\|v\|_1}$, which implies (55). $\square$

The following Lemma 4 demonstrate the error for truncated singular value decomposition.

Lemma 4. *Suppose $\tilde{\mathbf{A}}$ and $\mathbf{A}$ are any two matrices of the same dimension, $\hat{\mathbf{A}} = \tilde{\mathbf{A}}_{\max(r)}$. Then the following inequality holds,*

$$\left\| \hat{\mathbf{A}} - \mathbf{A} \right\|_F \leq 2\sqrt{2r} \left\| \tilde{\mathbf{A}} - \mathbf{A} \right\| + 2\sqrt{2r} \|\mathbf{A}_{-\max(r)}\| + \|\mathbf{A}_{-\max(r)}\|_F. \tag{56}$$

If $\operatorname{rank}(\mathbf{A}) \leq r$*, we also have*

$$\left\| \hat{\mathbf{A}} - \mathbf{A} \right\|_F \leq 2\|\tilde{\mathbf{A}} - \mathbf{A}\|_F. \tag{57}$$

The results still hold if $\mathbf{A}$ and $\mathbf{A}$ are two bivariate operators.

Proof of Lemma 4. Note that $\hat{\mathbf{A}}$ and $\mathbf{A}_{\max(r)}$ are both of rank-r, thus $\hat{\mathbf{A}} - \mathbf{A}_{\max(r)}$ is of rank at most $2r$, and $\|\hat{\mathbf{A}} - \mathbf{A}_{\max(r)}\|_F \leq \sqrt{2r} \|\hat{\mathbf{A}} - \mathbf{A}_{\max(r)}\|$. By Weyl's inequality [50], $\sigma_{r+1}(\tilde{\mathbf{A}}) \leq \sigma_{r+1}(\mathbf{A}) + \|\mathbf{A} - \tilde{\mathbf{A}}\|$. Therefore,

$$\begin{aligned}\|\hat{\mathbf{A}} - \mathbf{A}\|_F \leq & \|\hat{\mathbf{A}} - \mathbf{A}_{\max(r)}\|_F + \|\mathbf{A}_{-\max(r)}\|_F \leq \sqrt{2r} \|\hat{\mathbf{A}} - \mathbf{A}_{\max(r)}\| + \|\mathbf{A}_{-\max(r)}\|_F \\ \leq & \sqrt{2r} \left(\|\tilde{\mathbf{A}} - \mathbf{A}\| + \|\tilde{\mathbf{A}}_{-\max(r)}\| + \|\mathbf{A}_{-\max(r)}\| \right) + \|\mathbf{A}_{-\max(r)}\|_F \\ \leq & \sqrt{2r} \left(\|\tilde{\mathbf{A}} - \mathbf{A}\| + \sigma_{r+1}(\tilde{\mathbf{A}}) + \sigma_{r+1}(\mathbf{A}) \right) + \|\mathbf{A}_{-\max(r)}\|_F \\ \overset{\text{Weyl's inequality}}{\leq} & \sqrt{2r} \left(\|\tilde{\mathbf{A}} - \mathbf{A}\| + 2\sigma_{r+1}(\mathbf{A}) + \|\tilde{\mathbf{A}} - \mathbf{A}\| \right) + \|\mathbf{A}_{-\max(r)}\|_F \\ = & 2\sqrt{2r} \|\tilde{\mathbf{A}} - \mathbf{A}\| + 2\sqrt{2r} \|\mathbf{A}_{-\max(r)}\| + \|\mathbf{A}_{-\max(r)}\|_F,\end{aligned}$$

which yields (56). For (57), we have

$$\begin{aligned}\left\| \hat{\mathbf{A}} - \mathbf{A} \right\|_F \leq & \|\hat{\mathbf{A}} - \tilde{\mathbf{A}}\|_F + \|\tilde{\mathbf{A}} - \mathbf{A}\|_F = \min_{\operatorname{rank}(\mathbf{M}) \leq r} \|\tilde{\mathbf{A}} - \mathbf{M}\|_F + \|\tilde{\mathbf{A}} - \mathbf{A}\|_F \\ \leq & \|\tilde{\mathbf{A}} - \mathbf{A}\|_F + \|\tilde{\mathbf{A}} - \mathbf{A}\|_F = 2\|\tilde{\mathbf{A}} - \mathbf{A}\|_F.\end{aligned}$$

□

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

Lemma 5. *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 $a, b > 0$ 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,*

$$\frac{a}{2b^2}\|u-v\|_2^2 \leq D_{KL}(u||v) \leq \frac{b}{2a^2}\|u-v\|_2^2, \tag{58}$$

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 5. 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. □

Lemma 6 (Exponential Decay of Markov Mixing Rate)**.** *Suppose $\tau(\varepsilon)$ is either the discrete or continuous Markov mixing time defined in* (5)*, $\varepsilon \leq \delta < 1/2$, then*

$$\tau(\varepsilon) \leq \tau(\delta) \cdot \left(\left\lceil \frac{\log(\varepsilon/\delta)}{\log(2\delta)} \right\rceil + 1\right). \tag{59}$$

Proof of Lemma 6. 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.

$$(\theta_+)_j = \min\{\theta_j, 0\}, \quad (\theta_-)_j = -\max\{\theta_j, 0\}, \quad 1 \leq j \leq p. \tag{60}$$

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

$$\frac{1}{2}\left\|\mathbf{P}^{k'}\theta - \mu\right\|_1 = \frac{1}{2}\left\|\sum_{i=1}^{p}\mathbf{P}^{k'}\theta_i e^{(i)} - \mu\right\|_1 \leq \sum_{i=1}^{p}|\theta_i| \cdot \frac{1}{2}\left\|\mathbf{P}^{k'}e^{(i)} - \mu\right\|_1 \leq \sum_{i=1}^{p}|\theta_i| \cdot \delta = \delta. \quad (61)$$

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

$$\left\|\left(\mathbf{P}^k\theta - \mu\right)_+\right\|_1 = \left\|\left(\mathbf{P}^{k'}\theta - \mu\right)_-\right\|_1 = \frac{1}{2}\left\|\mathbf{P}^k\theta - \mu\right\|_1. \quad (62)$$

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

$$\begin{aligned}
&\frac{1}{2}\left\|\mathbf{P}^{k'}\theta - \mu\right\|_1 = \frac{1}{2}\left\|\mathbf{P}^{k'-k}(\mathbf{P}^k\theta - \mu)\right\|_1 = \frac{1}{2}\left\|\mathbf{P}^{k'-k}\left[(\mathbf{P}^k\theta - \mu)_+ - (\mathbf{P}^k e_i - \mu)_-\right]\right\|_1 \\
\leq & \frac{1}{2}\left\|\mathbf{P}^{k'-k}\frac{(\mathbf{P}^k\theta - \mu)_+}{\|(\mathbf{P}^k\theta - \mu)_+\|_1} - \mu\right\|_1 \cdot \|(\mathbf{P}^k\theta - \mu)_+\|_1 \\
& + \frac{1}{2}\left\|\mathbf{P}^{k'-k}\frac{(\mathbf{P}^k\theta - \mu)_-}{\|(\mathbf{P}^k\theta - \mu)_-\|_1} - \mu\right\|_1 \cdot \|(\mathbf{P}^k\theta - \mu)_-\|_1 \\
\overset{(61)(62)}{\leq} & \delta\left(\left\|(\mathbf{P}^k\theta - \mu)_+\right\|_1 + \left\|(\mathbf{P}^k\theta - \mu)_-\right\|_1\right) \leq \delta\|\mathbf{P}^k\theta - \mu\| \leq \frac{1}{2}(2\delta)^2.
\end{aligned}$$

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

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

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

$$\frac{1}{2}\left\|\mathbf{P}^{k'}\theta - \mu\right\|_1 \leq \frac{1}{2}(2\delta)^l \leq \frac{1}{2}(2\delta)^{\frac{\log(\varepsilon/\delta)}{\log(2\delta)}+1} = \frac{1}{2}2\delta \cdot (\varepsilon/\delta) = \varepsilon, \quad (63)$$

which implies $\tau(\varepsilon) \leq kl = \tau(\delta) \cdot (\lceil\log(\varepsilon/\delta)/\log(2\delta)\rceil + 1)$, and complete the proof for (59). The proof for continuous case is similar after replacing $\mathbf{P}$ by $\mathbf{P}$. Thus we have finished the proof for Lemma 6. □

The next Lemmas 7 and 8 relates the Markov mixing time to Cheeger's constant and eigen-gap condition. The detailed proofs were given in [30] and [28].

Lemma 7 (Markov Mixing Time and Cheeger's Constant)**.** *Suppose $\mathbf{P} \in \mathbb{R}^{p\times p}$ is an ergodic Markov chain transition matrix with p states and stationary distribution μ. Let*

$$\Phi = \min_{\Omega\subseteq\{1,\ldots,p\},\mu(\Omega)} \frac{\sum_{i\in\Omega, j\in\Omega^c}\mu_i\mathbf{P}_{ij}}{\sum_{i\in\Omega}\mu_i} \quad (64)$$

be the Cheeger's constant, then the mixing time, defined as (5)*, satisfies the following upper bound,*

$$\tau(\varepsilon) \leq \frac{2}{\Phi^2}\log\left(\frac{1}{2\varepsilon\mu_{\min}}\right). \quad (65)$$

Lemma 8 (Markov Mixing Time and Eigen-gap Condition)**.** *Suppose* $\mathbf{P} \in \mathbb{R}^{p\times p}$ *is an ergodic and reversible Markov chain transition matrix with* p *states and stationary distribution* μ*. Suppose* λ_2 *is its second largest eigen-value, then*

$$\tau(\varepsilon) \leq \frac{1}{1-\lambda_2}\log\left(\frac{1}{2\varepsilon\mu_{\min}}\right). \tag{66}$$

Lemma 9 (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 stationary distribution* μ *and the Markov mixing time* $\tau(\varepsilon)$ *defined as* (5)*. Recall the frequency matrix is defined as as* $\mathbf{F} = \mathrm{diag}(\mu)\mathbf{P}$*. Given a Markov chain with* $(n+1)$ *observable states* $X = \{x_0, x_1, \ldots, x_n\}$*, we introduce the empirical stationary distribution* $\tilde{\mu}$ *and empirical frequency matrix as*

$$\tilde{\mu} = \frac{1}{n}\sum_{k=1}^{n} e_{x_k}, \quad \text{where } e_{x_k} \text{is the indicator for } x_k, \text{ i.e.,} \quad (e_{x_k})_i = \begin{cases} 1, & x_k = i; \\ 0, & x_k \neq i; \end{cases} \tag{67}$$

$$\tilde{\mathbf{F}} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{E}_k, \quad \text{where } \mathbf{E}_k \in \mathbb{R}^{p\times p}, \quad (\mathbf{E}_k)_{ij} = \begin{cases} 1, & (x_{k-1}, x_k) = (i,j) \\ 0, & \text{otherwise.} \end{cases} \tag{68}$$

When $n \geq C\mu_{\max}/\alpha$*,* $\alpha = \tau\left(\min\{t/2, \mu_{\max}\}\right) + 1$*, we have*

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

$$\mathbb{P}\left(\left\|\tilde{\mathbf{F}} - \mathbf{F}\right\| \geq C\sqrt{\frac{\mu_{\max}\tau(\sqrt{\mu_{\max}/n})\log(n)}{n}}\right) \leq n^{-c_0} \tag{70}$$

$$P\left(\|\tilde{\mu} - \mu\|_\infty \geq t\right) \leq 2\alpha \exp\left(-\frac{(tn/\alpha)^2/8}{2n\mu_{\max}/\alpha + tn/(6\alpha)}\right) \tag{71}$$

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

for some constants $C, c, c_0 > 0$*. Furthermore,*

- *if* $\mathbf{P}$ *has Cheeger's constant* Φ *defined as* (64)*,*

$$\mathbb{P}\left(\left\|\tilde{\mathbf{F}} - \mathbf{F}\right\| \geq C\sqrt{\frac{\mu_{\max}\log(1/\mu_{\min})\log^2(n)}{n\Phi^2}}\right) \leq n^{-c_0}, \tag{73}$$

$$\mathbb{P}\left(\|\tilde{\mu} - \mu\|_\infty \geq C\sqrt{\frac{\mu_{\max}\log(1/\mu_{\min})\log^2(n)}{n\Phi^2}}\right) \leq n^{-c_0}, \tag{74}$$

  for some uniform constants $C, c, c_0 > 0$*.*

- *if* $\mathbf{P}$ *is reversible with second largest eigenvalue* $\lambda_2 < 1$, *then*

$$\mathbb{P}\left(\left\|\tilde{\mathbf{F}} - \mathbf{F}\right\| \geq C\sqrt{\frac{\mu_{\max}\log(1/\mu_{\min})\log^2(n)}{n(1-\lambda_2)}}\right) \leq n^{-c_0} \tag{75}$$

$$\mathbb{P}\left(\|\tilde{\mu} - \mu\|_\infty \geq C\sqrt{\frac{\mu_{\max}\log(1/\mu_{\min})\log^2(n)}{n(1-\lambda_2)}}\right) \leq n^{-c_0} \tag{76}$$

for some uniform constantc $C, c, c_0 > 0$.

Proof of Lemma 9. Let $n_0 = \lfloor n/\alpha \rfloor$, and without loss of generality, assume n is a multiple of α. The more general case where n is not a multiple of α can be derived similarly.

We introduce the "thin" matrix sequences, $e_{\tilde{x}_k^{(l)}}$ and $\tilde{\mathbf{E}}_k^{(l)}$, as

$$\tilde{e}_k^{(l)} = e_{x_{k\alpha+l}} - \mathbb{E}\left(e_{x_{k\alpha+l}} \middle| e_{x_{(k-1)\alpha+l}}\right), \quad l = 1, \ldots, \alpha; k = 1, \ldots, n_0;$$

$$\tilde{\mathbf{E}}_k^{(l)} = \mathbf{E}_{k\alpha+l} - \mathbb{E}\left(\mathbf{E}_{k\alpha+l} | \mathbf{E}_{(k-1)\alpha+l}\right), \quad l = 1, \ldots, \alpha; k = 1, \ldots, n_0. \tag{77}$$

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

$$\left\|\mathbb{E}\left(e_{x_{k\alpha+l}} \middle| e_{x_{(k-1)\alpha+l}}\right)\right\|_2 \leq \mathbb{E}\|e_{x_{k\alpha+l}}\|_2 \leq 1, \quad \left\|\mathbb{E}\left(\mathbf{E}_{k\alpha+l} | \mathbf{E}_{(k-1)\alpha+l}\right)\right\| \leq \mathbb{E}\|\mathbf{E}_{k\alpha+l}\| \leq 1, \tag{78}$$

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 for any fixed l. Note that for any given $\tilde{\mathbf{E}}_{k-1}^{(l)}$ and $e_{\tilde{x}_{k-1}^{(l)}}$, 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 $\tilde{\mu} = e_{x_{(k-1)\alpha+l}\mathbf{P}^{\alpha-1}}$. By our mixing time assumption,

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

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

$$\tilde{\mathbf{E}}_k^{(l)} = \mathbf{E}_{k\alpha+l} - \mathbb{E}\left(\mathbf{E}_{k\alpha+l} \middle| x_{(k-1)\alpha+l}\right), \quad \text{where} \quad \mathbf{E}_{k\alpha+l} = e_{x_{k\alpha+l-1}} \cdot e_{x_{k\alpha+l}}^\top, \tag{81}$$

$$\begin{aligned} \mathbb{P}\left(\mathbf{E}_{k\alpha+l} = e_i e_j^\top \middle| x_{(k-1)\alpha+l}\right) &= \mathbb{P}\left((x_{k\alpha+l-1}, x_{k\alpha+l}) = (i,j) \middle| x_{(k-1)\alpha+l}\right) \\ &= \left(e_{x_{(k-1)\alpha+l}}^\top \mathbf{P}^{\alpha-1}\right)_i \cdot \mathbf{P}_{ij} = \tilde{\mu}_i \mathbf{P}_{ij}, \end{aligned} \tag{82}$$

we can further calculate that

$$
\begin{aligned}
&\mathbb{E}\left(\mathbf{E}_{k\alpha+l}\mathbf{E}_{k\alpha+l}^\top \middle| x_{(k-1)\alpha+l}\right) = \sum_{i=1}^p \sum_{j=1}^p e_i e_i^\top \tilde{\mu}_i \mathbf{P}_{ij} = \sum_{i=1}^p e_i e_i^\top \tilde{\mu}_i \\
=&\mathrm{diag}\left(\tilde{\mu}\right) = \mathrm{diag}(\mu) + \mathrm{diag}\left(\tilde{\mu} - \mu\right) \\
\preceq&\mu_{\max} I_p + \|\tilde{\mu} - \mu\|_1 \cdot I_p \preceq 2\mu_{\max} I_p;
\end{aligned}
\tag{83}
$$

$$
\begin{aligned}
&\mathbb{E}\left(\mathbf{E}_{k\alpha+l}^\top \mathbf{E}_{k\alpha+l} \middle| x_{(k-1)\alpha+l}\right) = \sum_{i=1}^p \sum_{j=1}^p e_j e_j^\top \left\{\tilde{\mu}_i \mathbf{P}_{ij}\right\} \\
=&\sum_{i=1}^p \sum_{j=1}^p e_j e_j^\top \left\{\mu_i \mathbf{P}_{ij}\right\} + \sum_{i=1}^p \sum_{j=1}^p e_j e_j^\top \left\{(\tilde{\mu}_i - \mu)_i \mathbf{P}_{ij}\right\} \\
\preceq&\sum_{j=1}^p e_j e_j^\top \mu_j + \sum_{j=1}^p e_j e_j^\top \|\tilde{\mu} - \mu\|_1 \cdot \max_{ij} \mathbf{P}_{ij} \quad (\text{since } \mu^\top \mathbf{P} = \mu) \\
\preceq&\mu_{\max} I_p + \|\tilde{\mu} - \mu\|_1 \cdot I_p \preceq 2\mu_{\max} I_p.
\end{aligned}
\tag{84}
$$

Therefore,

$$
\begin{aligned}
0 \preceq&\mathbb{E}\left(\tilde{\mathbf{E}}_k^{(l)} (\tilde{\mathbf{E}}_k^{(l)})^\top \middle| \tilde{\mathbf{E}}_{k-1}^{(l)}\right) \\
=&\mathbb{E}\left\{\left(\mathbf{E}_{k\alpha+l} - \mathbb{E}(\mathbf{E}_{k\alpha+l} | x_{(k-1)\alpha+l})\right)\left(\mathbf{E}_{k\alpha+l} - \mathbb{E}(\mathbf{E}_{k\alpha+l} | x_{(k-1)\alpha+l})\right)^\top \middle| x_{(k-1)\alpha+l}\right\} \\
=&\mathbb{E}\left\{\mathbf{E}_{k\alpha+1}\mathbf{E}_{k\alpha+1}^\top \middle| x_{(k-1)\alpha+l}\right\} - \mathbb{E}\left\{\mathbf{E}_{k\alpha+l} \middle| x_{(k-1)\alpha+l}\right\}\mathbb{E}\left\{\mathbf{E}_{k\alpha+l}^\top \middle| x_{(k-1)\alpha+l}\right\} \\
\preceq&\mathbb{E}\left\{\mathbf{E}_{k\alpha+1}\mathbf{E}_{k\alpha+1}^\top \middle| x_{(k-1)\alpha+l}\right\} \preceq 2\mu_{\max} I_p.
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
0 \preceq&\mathbb{E}\left((\tilde{\mathbf{E}}_k^{(l)})^\top \tilde{\mathbf{E}}_k^{(l)} \middle| \tilde{\mathbf{E}}_{k-1}^{(l)}\right) \\
=&\mathbb{E}\left\{\left(\mathbf{E}_{k\alpha+l} - \mathbb{E}(\mathbf{E}_{k\alpha+l} | x_{(k-1)\alpha+l})\right)^\top\left(\mathbf{E}_{k\alpha+l} - \mathbb{E}(\mathbf{E}_{k\alpha+l} | x_{(k-1)\alpha+l})\right) \middle| x_{(k-1)\alpha+l}\right\} \\
=&\mathbb{E}\left\{\mathbf{E}_{k\alpha+1}^\top \mathbf{E}_{k\alpha+1} \middle| x_{(k-1)\alpha+l}\right\} - \mathbb{E}\left\{\mathbf{E}_{k\alpha+l}^\top \middle| x_{(k-1)\alpha+l}\right\}\mathbb{E}\left\{\mathbf{E}_{k\alpha+l} \middle| x_{(k-1)\alpha+l}\right\} \\
\preceq&\mathbb{E}\left\{\mathbf{E}_{k\alpha+1}^\top \mathbf{E}_{k\alpha+1} \middle| x_{(k-1)\alpha+l}\right\} \preceq 2\mu_{\max} I_p,
\end{aligned}
$$

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

$$
\max\left\{\left\|\mathbb{E}\left((\tilde{\mathbf{E}}_k^{(l)})^\top \tilde{\mathbf{E}}_k^{(l)} \middle| \tilde{\mathbf{E}}_{k-1}^{(l)}\right)\right\|, \ \left\|\mathbb{E}\left(\tilde{\mathbf{E}}_k^{(l)} (\tilde{\mathbf{E}}_k^{(l)})^\top \middle| \tilde{\mathbf{E}}_{k-1}^{(l)}\right)\right\|\right\} \leq 2\mu_{\max} I_p. \tag{85}
$$

Next, the predictable quadratic variation process of the martingale satisfies

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

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

Now by matrix Freedman's inequality (Corollary 1.3 in [47]), we know

$$\mathbb{P}\left(\left\|\frac{1}{n_0}\sum_{k=1}^{n_0}\tilde{\mathbf{E}}_k^{(l)}\right\|\geq t/2\right)\leq 2p\exp\left(-\frac{(tn_0)^2/8}{2n_0\mu_{\max}+tn_0/6}\right). \tag{86}$$

Next, we shall note that

$$\begin{aligned}
&\mathbb{E}\left(\mathbf{E}_{k\alpha+l}\middle|x_{(k-1)\alpha+l}\right)-\mathrm{diag}(\mu)\mathbf{P}\\
=&\sum_{i=1}^p\sum_{j=1}^p e_i\left(e_{x_{(k-1)\alpha+l}}^\top\mathbf{P}^{\alpha-1}\right)_i\mathbf{P}_{ij}e_j^\top-\mathrm{diag}(\mu)\mathbf{P}\\
=&\mathrm{diag}\left(e_{x_{(k-1)\alpha+l}}^\top\mathbf{P}^{\alpha-1}\right)\mathbf{P}-\mathrm{diag}(\mu)\mathbf{P},
\end{aligned}$$

thus

$$\begin{aligned}
&\left\|\mathbb{E}\left(\mathbf{E}_{k\alpha+l}\middle|x_{(k-1)\alpha+l}\right)-\mathrm{diag}(\mu)\mathbf{P}\right\|\leq\|(\tilde{\mu}-\mu)\mathbf{P}\|=\max_{\substack{u,v\in\mathbb{R}^p\\\|u\|_2=\|v\|_2=1}}u^\top\mathrm{diag}(\tilde{\mu}-\mu)\mathbf{P}v\\
\leq&\max_{\substack{u,v\in\mathbb{R}^p\\\|u\|_2=\|v\|_2=1}}\sum_{i=1}^p|u_i(\tilde{\mu}_i-\mu_i)\mathbf{P}_{ij}v_j|\leq\sum_{i=1}^p\sum_{j=1}^p|(\tilde{\mu}_i-\mu_i)\mathbf{P}_{ij}|\leq\|\tilde{\mu}-\mu\|_1\overset{(80)}{\leq}t/2.
\end{aligned}\tag{87}$$

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 (77), (86), and (87), we have for any $l=1,\ldots,\alpha$,

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

Finally, we only need to combine these "thin" summation sequences as follows,

$$\begin{aligned}
&\mathbb{P}\left(\|\tilde{\mathbf{F}}-\mathbf{F}\|\geq t\right)=\mathbb{P}\left(\left\|\frac{1}{\alpha}\sum_{l=1}^{\alpha}\frac{1}{n_0}\sum_{k=1}^{n_0}\mathbf{E}_{k\alpha+l}-\mathbf{F}\right\|\geq t\right)\\
\leq&\mathbf{P}\left(\max_{1\leq l\leq\alpha}\left\|\sum_{k=1}^{n_0}\frac{1}{n_0}\mathbf{E}_{k\alpha+l}-\mathbf{F}\right\|\geq t\right)\leq\alpha\max_{1\leq l\leq p}\mathbb{P}\left(\left\|\sum_{k=1}^{n_0}\frac{1}{n_0}\mathbf{E}_{k\alpha+l}-\mathbf{F}\right\|\geq t\right)\\
\leq&2\alpha p\exp\left(-\frac{(tn_0)^2/8}{2n_0\mu_{\max}+tn_0/6}\right),
\end{aligned}$$

which has finished the proof for the error bound of $\tilde{\mathbf{F}}$.

The proof for $\|\tilde{\mu}-\mu\|_\infty$ is similar. Note that for any index $j\in\{1,\ldots,p\}$,

$$\begin{aligned}
\left(\tilde{e}_k^{(l)}\right)_j=&\left(e_{x_{k\alpha+l}}\right)_j-\mathbb{E}\left(\left(e_{x_{k\alpha+l}}\right)_j\middle|e_{x_{(k-1)\alpha+l}}\right)\\
=&1_{\{x_{k\alpha+l}=j\}}-\mathbb{E}\left(1_{\{x_{k\alpha+l}=j\}}\middle|x_{(k-1)\alpha+l}\right).
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{E}\left(\tilde{e}_k^{(l)}\right)_j^2 = \mathrm{Var}\left(1_{\{x_{k\alpha+l}=j\}}\middle|x_{(k-1)\alpha+l}\right) \le \mathbb{E}\left(1^2_{\{x_{k\alpha+l}=j\}}\right) = \left(e^\top_{x_{(k-1)\alpha+l}}\mathbf{P}^\alpha\right)_j \\
\le&\mu_j + \left(e^\top_{x(k-1)\alpha+l}\mathbf{P}^\alpha - \mu^\top\right)_j \le 2\mu_{\max}.
\end{aligned}$$

By Freedman's inequality (e.g. Theorem 1.6 in [17] and Theorem 1.1 in [47]), for any $1 \le j \le p$,

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

On the other hand,

$$\left\|\mathbb{E}\left(e_{x_{k\alpha+l}}\middle|e_{x_{(k-1)\alpha+l}}\right) - \mu\right\|_\infty = \left\|e^\top_{x_{(k-1)\alpha+l}}\mathbf{P}^\alpha - \mu^\top\right\|_\infty \le \left\|e^\top_{x_{(k-1)\alpha+l}}\mathbf{P}^\alpha - \mu^\top\right\|_1 \le \frac{t}{2}\wedge\mu_{\max}.$$

Combining the two inequality above, we have

$$\mathbb{P}\left(\left|\sum_{k=1}^{n_0}(\tilde{e}_k^{(l)})_j\right| \ge t\right) \le 2\exp\left(\frac{-t^2/8}{2n_0\mu_{\max} + t/6}\right). \tag{88}$$

Therefore,

$$\begin{aligned}
&\mathbb{P}\left(\|\tilde{\mu} - \mu\|_\infty \ge t\right) = \mathbb{P}\left(\left\|\frac{1}{\alpha}\sum_{l=1}^{\alpha}\frac{1}{n_0}e_{x_{k\alpha+l}}\right\| \ge t\right) \\
\le&\mathbb{P}\left(\max_{1\le l\le\alpha}\left\|\sum_{k=1}^{n_0}\frac{1}{n_0}e_{x_{k\alpha+l}}\right\| \ge t\right) \le 2\alpha\exp\left(\frac{-(tn_0)^2/8}{2n_0\mu_{\max} + tn_0/6}\right),
\end{aligned} \tag{89}$$

which has developed the ℓ_∞ upper bound for $\tilde{\mu} - \mu$. □

Proof of Lemma 1. We first define

$$\mathbf{E}_k : \mathbb{R}^2 \to \mathbb{R}, \quad \mathbf{E}_k(a,b) = \frac{1}{h^2}\mathbf{K}\left(\frac{a - x_{k-1}}{h}, \frac{b - x_k}{h}\right) - \mathbf{F}(a,b).$$

Then $\tilde{\mathbf{F}} - \mathbf{F} = \frac{1}{n}\sum_{k=1}^n \mathbf{E}_k$. Let $\alpha = \tau(\varepsilon) + 1$ be an integer, where $0 < \varepsilon < 1/2$ is a constant to be determined later, and $n_0 = \lfloor n/\alpha \rfloor$. We first assume n is a multiple of α without loss of generality. The proof for the case when n is not a multiple of n_0 can be derived similarly.

We introduce the following lemma to characterize the second moments among the matrices $\{\mathbf{E}_k\}_{k=1}^n$.

Lemma 10. *Under the setting of Lemma 1, if $\alpha \le k$, $\alpha \le k_1 \le k_1 + \alpha \le k_2$, then the following upper bound for the second moments of $\mathbf{E}_{k_1}$ and $\mathbf{E}_{k_2}$ hold,*

$$\mathbb{E}\int \mathbf{E}_k(a,b)^2 da db \le Ch^{4\beta} + \frac{C}{h^2}; \tag{90}$$

$$\mathbb{E}\int \mathbf{E}_{k_1}(a,b)\mathbf{E}_{k_2}(a,b)dadb \le Ch^{4\beta} + \frac{C\varepsilon}{h^2}. \tag{91}$$

Furthermore, for any general k,

$$\mathbb{E}\int \mathbf{E}_k(a,b)^2 dadb \le C + \frac{C}{h^2}. \tag{92}$$

Here, $\mathbf{F}(a,b) = \mu(a)\mathbf{P}(a,b)$. C is a constant only depending on the kernel $\mathbf{K}$.

Next we introduce the following "thin" sequence of the original Markov chain $\{\mathbf{E}_k(a,b)\}_{k=1}^n$,

$$\mathbf{E}_k^{(l)} = \mathbf{E}_{k\alpha+l}, \quad l = 1,\ldots,\alpha, k = 1,\ldots,n_0. \tag{93}$$

By Lemma 10, we have

$$\begin{aligned}
\mathbb{E}\left\|\mathbf{F}-\mathbf{F}\right\|_F^2 =& \mathbb{E}\int\left(\frac{1}{n}\sum_{i=1}^n \mathbf{E}_k(a,b)\right)^2 dadb = \frac{1}{n^2}\int\left(\sum_{k=1}^{\alpha}\mathbf{E}_k(a,b) + \sum_{l=1}^{\alpha}\sum_{k=2}^{n_0}\mathbf{E}_k\right)^2 dadb \\
\le& \frac{2\alpha}{n^2}\int\left\{\sum_{k=1}^{\alpha}\mathbf{E}_k^2(a,b) + \sum_{l=1}^{\alpha}\left(\sum_{k=2}^{n_0}\mathbf{E}_k^{(l)}(a,b)\right)^2\right\} dadb \\
\overset{\text{Lemma 10}}{\le}& \frac{2\alpha}{n^2}\left\{\alpha\left(C+\frac{C}{h^2}\right) + \alpha\left(n_0\left(Ch^{4\beta}+\frac{C}{h^2}\right) + n_0^2\left(Ch^{4\beta}+\frac{4\varepsilon}{h^2}\right)\right)\right\} \\
\le& \frac{4\alpha^2 n_0^2}{n^2}\cdot Ch^{4\beta} + \frac{C\alpha^2 n_0}{n^2h^2} + \frac{C\alpha^2 n_0^2\varepsilon}{n^2h^2} \\
\le& Ch^{4\beta} + \frac{C\alpha}{nh^2} + \frac{\varepsilon}{h^2}.
\end{aligned} \tag{94}$$

Let $\varepsilon = 1/n$, by Lemma 6, $\alpha = \tau(1/n) + 1 \le C\tau\log(n)$. We thus have

$$\mathbb{E}\left\|\tilde{\mathbf{F}}-\mathbf{F}\right\|_F^2 \le Ch^{4\beta} + \frac{C\tau\log(n)}{nh^2}. \tag{95}$$

It is easy to see that $h = n^{-\frac{1}{4\beta+2}}$ achieves the minimum of the right hand side of (95), that is

$$\mathbb{E}\left\|\tilde{\mathbf{F}}-\mathbf{F}\right\|_F^2 \le C\left(\frac{\tau\log(n)}{n}\right)^{\frac{4\beta}{4\beta+2}}. \tag{96}$$

This has finished the proof for Lemma 1. □

Proof of Lemma 10. For convenience, we denote

$$K_q = \left(\int |\mathbf{K}(a,b)|^q dadb\right)^{1/q}, \quad \forall q \ge 1.$$

as the ℓ_q norm of kernel $\mathbf{K}$. In particular, $K_\infty = \max_{a,b} |\mathbf{K}(a,b)|$. Let $\tilde{\mu}_k$ be the marginal density of x_k, then $\tilde{\mathbf{F}}_k(a,b) = \tilde{\mu}_k(a)\mathbf{P}(a,b)$ is the joint density of (x_{k-1}, x_k). When $\alpha \le k_1$, by definition of α, we must have

$$\|\tilde{\mu}_k - \mu\|_{\ell_1} = \int |\tilde{\mu}_k(a) - \mu(a)|\, da \le \varepsilon.$$

Then the ℓ_1 distance between $\tilde{\mathbf{F}}_k$ and $\mathbf{F}$ are upper bounded as

$$\begin{aligned} \left\|\tilde{\mathbf{F}}_k - \mathbf{F}\right\|_{\ell_1} &= \int |(\tilde{\mu}_k(a) - \mu(a))\, \mathbf{F}(a,b)|\, dadb \\ &\le \int |\tilde{\mu}_k(a) - \mu(a)|\mu(a) da \le \mu_{\max}\varepsilon. \end{aligned} \tag{97}$$

Thus, $\mathbb{E}\mathbf{E}_{k_1}^2$ has the following variance-bias decomposition,

$$\begin{aligned} \mathbb{E}\,\|\mathbf{E}_k\|_F^2 &= \int \mathbb{E}\mathbf{E}_k^2(a,b) dadb \\ &\le \int \left(\mathbb{E}\mathbf{E}_k(a,b)\right)^2 dadb + \int \mathrm{Var}\left(\mathbf{E}_k(a,b)\right) dadb, \end{aligned} \tag{98}$$

which is a typical bias-variance decomposition. Next, we analyze these two terms respectively. For the variance,

$$\begin{aligned} &\int \mathrm{Var}(\mathbf{E}_k(a,b)) dadb = \int \mathrm{Var}\left(\frac{1}{h^2}\mathbf{K}\left(\frac{a-x_{k-1}}{h}, \frac{b-x_k}{h}\right)\right) \\ \le &\int \mathbb{E}\frac{1}{h^4}\mathbf{K}^2\left(\frac{a-x_{k-1}}{h}, \frac{b-x_k}{h}\right) dadb = \mathbb{E}\int \frac{1}{h^4}\mathbf{K}^2\left(\frac{a-x_{k-1}}{h}. \frac{b-x_k}{h}\right) d\left(\frac{a}{h}\right) d\left(\frac{b}{h}\right) \\ = &\frac{1}{h^2}\mathbb{E}\int_{\mathbb{R}^2} \mathbf{K}^2(a,b) dadb = \frac{K_2^2}{h^2}. \end{aligned} \tag{99}$$

Next, we consider the bias. For each fixed $a, b \in \mathbb{R}$,

$$\begin{aligned} \mathbb{E}\mathbf{E}_k = &\frac{1}{h^2}\mathbb{E}\mathbf{K}\left(\frac{a-x_{k-1}}{h}, \frac{b-x_k}{h}\right) - \mathbf{F}(a,b) \\ = &\int \frac{1}{h^2}\mathbf{K}\left(\frac{a-x_{k-1}}{h}, \frac{b-x_k}{h}\right)\left(\tilde{\mathbf{F}}_k(x_{k-1}, x_k) - \mathbf{F}(a,b)\right) dx_{k-1}dx_k \\ = &\int \frac{1}{h^2}\mathbf{K}\left(\frac{a-x_{k-1}}{h}, \frac{b-x_k}{h}\right)\left(\tilde{\mathbf{F}}_k(x_{k-1}, x_k) - \mathbf{F}(a,b)\right) dx_{k-1}dx_k \\ = &\int \frac{1}{h^2}\mathbf{K}\left(\frac{a-x_{k-1}}{h}, \frac{b-x_k}{h}\right)\left(\tilde{\mathbf{F}}_k(x_{k-1}, x_k) - \mathbf{F}(x_{k-1}, x_k)\right) dx_{k-1}dx_k \\ &+ \int \frac{1}{h^2}\mathbf{K}\left(\frac{a-x_{k-1}}{h}, \frac{b-x_k}{h}\right)\left(\mathbf{F}(x_{k-1}, x_k) - \mathbf{F}(a,b)\right) dx_{k-1}dx_k. \end{aligned} \tag{100}$$

We analyze the squared loss for the two terms respectively as follows. Note that for all $a, b \in \mathbb{R}$,

$$\begin{aligned}
&\left|\frac{1}{h^2}\int \mathbf{K}\left(\frac{a-x_{k-1}}{h}, \frac{b-x_k}{h}\right)\left(\tilde{\mathbf{F}}_k(x_{k-1},x_k)-\mathbf{F}(x_{k-1},x_k)\right)dx_{k-1}dx_k\right| \\
\leq &\frac{1}{h^2}\int K_\infty \left|\tilde{\mathbf{F}}_k(x_{k-1},x_k)-\mathbf{F}(x_{k-1},x_k)\right|dx_{k-1}dx_k \leq \frac{\varepsilon K_\infty}{h^2}.
\end{aligned}$$

Thus,

$$\begin{aligned}
&\int \frac{1}{h^4}\left\{\int \mathbf{K}\left(\frac{a-x_{k-1}}{h}, \frac{b-x_k}{h}\right)\left(\tilde{\mathbf{F}}(x_{k-1},x_k)-\mathbf{F}(x_{k-1},x_k)\right)dx_{k-1}dx_k\right\}^2 dadb \\
\leq &\frac{K_\infty\varepsilon}{h^2}\cdot\int \frac{1}{h^2}\left|\mathbf{K}\left(\frac{a-x_{k-1}}{h}, \frac{b-x_k}{h}\right)\right|\left|\tilde{\mathbf{F}}(x_{k-1},x_k)-\mathbf{F}(x_{k-1},x_k)\right|dx_{k-1}dx_k dadb \\
\leq &\frac{K_\infty\varepsilon}{h^2}\int |\mathbf{K}(a,b)|\,dadb\cdot\int\left|\tilde{\mathbf{F}}_k(x_{k-1},x_k)-\mathbf{F}(x_{k-1},x_k)\right|dx_{k-1}dx_k \\
\leq &\frac{K_\infty K_1\varepsilon^2}{h^2}.
\end{aligned} \tag{101}$$

Next, since $\mathbf{K}$ is order-$\lfloor\beta\rfloor$ kernel and $\mathbf{F}\in\mathcal{H}(\beta,L)$, similarly to the argument in nonparametric kernel density estimation (see, e.g., Proposition 1.5 in [48]), given that $\mathbf{F}$ belongs to the Nikol'ski class $\mathcal{H}(\beta,L)$, we have

$$\int\left\{\int \frac{1}{h^2}\mathbf{K}\left(\frac{a-x_{k-1}}{h}, \frac{b-x_k}{h}\right)\left(\mathbf{F}(x_{k-1},x_k)-\mathbf{F}(a,b)\right)dx_{k-1}dx_k\right\}^2 dadb \leq C_0 h^{4\beta}, \tag{102}$$

where $C_0 = \frac{L}{\lfloor\beta\rfloor}\int |u|^{\lfloor\beta\rfloor}\cdot|\mathbf{K}(u)|du$. Thus, (100), (101), and (102) together yield

$$\iint_{\mathbb{R}^2}\left(\mathbb{E}\mathbf{E}_k(a,b)\right)^2 dadb \leq C_0 h^{4\beta} + \frac{K_\infty K_1\varepsilon^2}{h^2}. \tag{103}$$

With (98), (99), and (103), and $\varepsilon\leq 2$, we have

$$\mathbb{E}\int_{\mathbb{R}^2}\mathbf{E}_k^2(a,b)dadb \leq Ch^{4\beta}+\frac{C}{h^2},$$

where the constant C only depends on $\mathbf{K}$ and L, but not β and h. Then we have finished the proof for the first part of this lemma.

We consider (91) next. The key for proving (91) is the fact that $\mathbf{E}_{k_1}$ and $\mathbf{E}_{k_2}$ are nearly independent when $k_1+\alpha\leq k_2$ hold. First, note that the joint distribution for $x_{k-1}, x_k, x_{l-1}, x_l$ in the given setting can be written as

$$\mathbf{f}(x_{k_1-1},x_{k_1},x_{k_2-1},x_{k_2}) = \tilde{\mu}_{k_1-1}(x_{k_1-1})\mathbf{P}(x_{k_1-1},x_{k_1})\mathbf{P}^{k_2-k_1-1}(x_{k_1},x_{k_2-1})\mathbf{P}(x_{k_2-1},x_{k_2}),$$

which satisfies the following inequality

$$
\begin{aligned}
&\|\mathbf{f}-\mathbf{F}\|_{\ell_1}\\
&\Big\|\tilde{\mu}_{k_1-1}(x_{k_1-1})\mathbf{P}(x_{k_1-1},x_{k_1})\mathbf{P}^{k_2-k_1-1}(x_{k_1},x_{k_2-1})\mathbf{P}(x_{k_2-1},x_{k_2})\\
&\qquad -\mathbf{F}(x_{k_1-1},x_{k_1})\mathbf{F}(x_{k_2-1},x_{k_2})\Big\|_{\ell_1}\\
\le&\Big\|\{\tilde{\mu}_{k_1-1}(x_{k_1-1})-\mu(x_{k_1-1})\}\,\mathbf{P}(x_{k_1-1},x_{k_1})\mathbf{P}^{k_2-k_1-1}(x_{k_1},x_{k_2-1})\mathbf{P}(x_{k_2-1},x_{k_2})\Big\|_{\ell_1}\\
&+\Big\|\tilde{\mu}_{k_1-1}(x_{k_1-1})\mathbf{P}(x_{k_1-1},x_{k_1})\left\{\mathbf{P}^{k_2-k_1-1}(x_{k_1},x_{k_2-1})-\mu(x_{k_2-1})\right\}\mathbf{P}(x_{k_2-1},x_{k_2})\Big\|_{\ell_1}\\
\le&\|\tilde{\mu}_{k-1}-\mu\|_{\ell_1}+\|\mathbf{P}^{k_2-k_1-1}(x_{k_1},\cdot)-\mu\|_{\ell_1}\le 2\varepsilon.
\end{aligned}
\tag{104}
$$

In other words, two densities $\mathbf{f}(x_{k-1},x_k,x_{l-1},x_l)$ and $\mathbf{F}(x_{k-1},x_k)\mathbf{F}(x_{l-1},x_l)$ are very close in ℓ_1 norm.

Therefore, we have

$$
\begin{aligned}
&\mathbb{E}\int \mathbf{E}_{k_1}(a,b)\mathbf{E}_{k_2}(a,b)dadb\\
=&\int \mathbf{E}_{k_1}(a,b)\mathbf{E}_{k_2}(a,b)\mathbf{f}(x_{k_1-1},x_{k_1},x_{k_2-1},x_{k_2})dx_{k_1-1}dx_{k_1}dx_{k_2-1}dx_{k_2}dadb\\
=&\int \mathbf{E}_{k_1}(a,b)\mathbf{E}_{k_2}(a,b)\mathbf{F}(x_{k_1-1},x_{k_1})\mathbf{F}(x_{k_2-1},x_{k_2})dx_{k_1-1}dx_{k_1}dx_{k_2-1}dx_{k_2}dadb\\
&+\int \mathbf{E}_{k_1}(a,b)\mathbf{E}_{k_2}(a,b)dadb\cdot\{\mathbf{f}-\mathbf{F}\}(x_{k_1-1},x_{k_1},x_{k_2-1},x_{k_2})dx_{k_1-1}dx_{k_1}dx_{k_2-1}dx_{k_2}dadb.
\end{aligned}
$$

We analyze the two terms separately.

$$
\begin{aligned}
&\int \mathbf{E}_{k_1}(a,b)\mathbf{E}_{k_2}(a,b)\mathbf{F}(x_{k_1-1},x_{k_1})\mathbf{F}(x_{k_2-1},x_{k_2})dx_{k_1-1}dx_{k_1}dx_{k_2-1}dx_{k_2}dadb\\
=&\int_{\mathbb{R}^2}\left\{\int_{\mathbb{R}^2}\left(\frac{1}{h^2}\mathbf{K}\left(\frac{a-x}{h},\frac{b-y}{h}\right)-\mathbf{F}(a,b)\right)\mathbf{F}(x,y)dxdy\right\}^2 dadb\\
=&\int_{\mathbb{R}^2}\left\{\int_{\mathbb{R}^2}\left(\frac{1}{h^2}\mathbf{K}\left(\frac{a-x}{h},\frac{b-y}{h}\right)-\mathbf{F}(a,b)\right)\mathbf{F}(x,y)dxdy\right\}^2 dadb
\end{aligned}
$$

For fixed $a,b\in\mathbb{R}$,

$$
\begin{aligned}
&\int_{\mathbb{R}^2}\left(\frac{1}{h^2}\mathbf{K}\left(\frac{a-x}{h},\frac{b-y}{h}\right)-\mathbf{F}(a,b)\right)\mathbf{F}(x,y)dxdy\\
=&\int_{\mathbb{R}^2}\frac{1}{h^2}\mathbf{K}\left(\frac{a-x}{h},\frac{b-y}{h}\right)\mathbf{F}(x,y)dxdy-\mathbf{F}(a,b)\\
=&\int_{\mathbb{R}^2}\frac{1}{h^2}\mathbf{K}\left(\frac{a-x}{h},\frac{b-h}{h}\right)(\mathbf{F}(x,y)-\mathbf{F}(a,b))\,dxdy.
\end{aligned}
$$

Then (102) implies

$$
\int\left\{\int_{\mathbb{R}^2}\frac{1}{h^2}\mathbf{K}\left(\frac{a-x}{h},\frac{b-y}{h}\right)(\mathbf{F}(x,y)-\mathbf{F}(a,b))\,dxdy\right\}^2 dadb\le Ch^{4\beta}.
$$

Next, for fixed $x_{k_1-1}, x_{k_1}, x_{k_2-1}, x_{k_2} \in \mathbb{R}$, the following upper bound holds for the integral on a, b,

$$
\begin{aligned}
&\int |\mathbf{E}_{k_1}(a,b)\mathbf{E}_{k_2}(a,b)|\, dadb \le \int \left[\mathbf{E}_{k_1}^2(a,b) + \mathbf{E}_{k_2}^2(a,b)\right] dadb \\
\le & \int \left[\frac{2}{h^4}\mathbf{K}^2\left(\frac{a-x_{k_1-1}}{h}, \frac{b-x_{k_1}}{h}\right) + \frac{2}{h^4}\mathbf{K}^2\left(\frac{a-x_{k_2-1}}{h}, \frac{b-x_{k_2}}{h}\right) + 4\mathbf{F}(a,b)^2\right] dadb \\
\le & \frac{4}{h^2}\|\mathbf{K}\|_{\ell_2}^2 + 4\|\mathbf{F}\|_{\ell_2}^2.
\end{aligned}
\tag{105}
$$

Thus,

$$
\begin{aligned}
&\Big|\int \mathbf{E}_{k_1}(a,b)\mathbf{E}_{k_2}(a,b)\cdot\{\mathbf{f}-\mathbf{F}\}(x_{k_1-1}, x_{k_1}, x_{k_2-1}, x_{k_2}) dx_{k_1-1}dx_{k_1}dx_{k_2-1}dx_{k_2}dadb\Big| \\
\le & \int \left(\frac{4}{h^2}\|\mathbf{K}\|_{\ell_2}^2 + 4\|\mathbf{F}\|_{\ell_2}^2\right) |\{\mathbf{f}-\mathbf{F}\}(x_{k_1-1}, x_{k_1}, x_{k_2-1}, x_{k_2})|\, dx_{k_1-1}dx_{k_1}dx_{k_2-1}dx_{k_2} \\
\overset{(104)}{\le} & 8\varepsilon\left(\frac{4}{h^2}\|\mathbf{K}\|_{\ell_2}^2 + 4\|\mathbf{F}\|_{\ell_2}^2\right).
\end{aligned}
$$

To sum up,

$$
\mathbb{E}\int \mathbf{E}_{k_1}(a,b)\mathbf{E}_{k_2}(a,b) \le Ch^{4\beta} + \frac{C\varepsilon}{h^2}.
$$

where C only relies on $\mathbf{K}$, L, and $\|\mathbf{F}\|_{\ell_2}$. $\square$