Learning Markov models via low-rank optimization

Ziwei Zhu
Department of Statistics, University of Michigan, Ann Arbor, MI 48109, ziweiz@umich.edu
Xudong Li
School of Data Science, Fudan University, Shanghai 200433, lixudong@fudan.edu.cn
Shanghai Center for Mathematical Sciences, Fudan University, Shanghai 200433
Mengdi Wang
Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, mengdiw@princeton.edu
Anru R. Zhang
Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706, anruzhang@stat.wisc.edu
Department of Biostatistics and Bioinformatics, Duke University, 27710

Modeling unknown systems from data is a precursor of system optimization and sequential decision making. In this paper, we focus on learning a Markov model from a single trajectory of states. Suppose that the transition model has a small rank despite of having a large state space, meaning that the system admits a low-dimensional latent structure. We show that one can estimate the full transition model accurately using a trajectory of length that is proportional to the total number of states. We propose two maximum likelihood estimation methods: a convex approach with nuclear-norm regularization and a nonconvex approach with rank constraint. We explicitly derive the statistical rates of both estimators in terms of the Kullback-Leiber divergence and the ℓ_2 error and also establish a minimax lower bound to assess the tightness of these rates. For computing the nonconvex estimator, we develop a novel DC (difference of convex function) programming algorithm that starts with the convex M-estimator and then successively refines the solution till convergence. Empirical experiments demonstrate consistent superiority of the nonconvex estimator over the convex one.

Key words: Markov Model, DC-programming, Non-convex Optimization, Rank Constrained Likelihood

1. Introduction

In engineering and management applications, one often has to collect data from unknown systems, learn their transition functions, and learn to make predictions and decisions. A critical precursor of decision making is to model the system from data. We study how to learn an unknown Markov model of the system from its state-transition trajectories. When the system admits a large number of states, recovering the full model becomes sample expensive.

In this paper, we focus on Markov processes where the transition matrix has a small rank. The small rank implies that the observed process is governed by a low-dimensional latent process which we cannot see in a straightforward manner. It is a property that is (approximately) satisfied in a wide range of practical systems. Despite the large state space, the low-rank property unlocks potential of accurate learning of a full set of transition density functions based on short empirical trajectories.

1.1. Motivating Examples

Practical state-transition processes with a large number of states often exhibit low-rank structures. For example, the sequence of stops made by a taxi turns out to follow a Markov model with approximately low rank structure (Liu et al. 2012, Benson et al. 2017). For another example, random walk on a lumpable network has a low-rank transition matrix (Buchholz 1994, E et al. 2008). The transition kernel with fast decaying eigenvalues has been also observed in molecular dynamics (Rohrdanz et al. 2011), which can be used to find metastable states, coresets and manifold structures of complicated dynamics (Chodera et al. 2007, Coifman et al. 2008).

Low-rank Markov models are also related to dimension reduction for control systems and reinforcement learning. For example, the state aggregation approach for modeling a high-dimensional system can be viewed as a low-rank approximation approach (Bertsekas 1995, Bertsekas and Tsitsiklis 1995, Singh et al. 1995). In state aggregation, one assumes that there exists a latent stochastic process $\{z_t\} \subset [r]$ such that $\mathbb{P}(s_{t+1} \mid s_t) = \sum_z \mathbb{P}(z_t = z \mid s_t)\mathbb{P}(s_{t+1} \mid z_t = z)$, which is equivalent to a factorization model of the transition kernel $\mathbf{P}$. In the context of reinforcement learning, the nonnegative factorization model was referred to as the generalization to the rich-observation model (Azizzadenesheli et al. 2016). The low-rank structure allows us to model and optimize the system using significantly fewer observations and less computation. Effective methods for estimating the low-rank Markov model would pave the way to better understanding of process data and more efficient decision making.

1.2. Our approach

We propose to estimate the low-rank Markov model based on an empirical trajectory of states, whose length is only proportional to the total number of states. We propose two approaches based on the maximum likelihood principle and low-rank optimization. The first approach uses a convex nuclear-norm regularizer to enforce the low-rank structure and a polyhedral constraint to ensure that optimization is over all probabilistic matrices. The second approach is to solve a rank-constrained optimization problem using difference-of-convex (DC) programming. For both approaches, we provide statistical upper bounds for the Kullback-Leibler (KL) divergence between the estimator and the true transition matrix as well as the ℓ_2 risk. We also provide an information-theoretic lower bound to show that the proposed estimators are nearly rate-optimal. Note that the low-rank estimation of the Markov model was considered in Zhang and Wang (2017) where a spectral method with total variation bound is given. In comparison, the novelty of our methods lies in the use of maximum likelihood principle and low-rank optimization, which allows us to obtain the first KL divergence bound for learning low-rank Markov models.

Our second approach involves solving a rank constraint optimization problem over probabilistic matrices, which is a refinement of the convex nuclear-norm approach. Due to the non-convex rank constraint, the optimization problem is difficult - to the best of our knowledge, there is no efficient approach that directly solves the rank-constraint problem. In this paper, we develop a penalty approach to relax the rank constraint and transform the original problem into a DC (difference of convex functions) programming one. Furthermore, we develop a particular DC algorithm to solve the problem by initiating at the solution to the convex problem and successively refining the solution through solving a sequence of inner subproblems. Each subroutine is based on the multi-block alternating direction method of multipliers (ADMM). Empirical experiments show that the successive refinements through DC programming do improve the learning quality. As a byproduct of this research, we develop a new class of DC algorithms and a unified convergence analysis for solving non-convex non-smooth problems, which were not available in the literature to our best knowledge.

1.3. Contributions and paper outline

The paper provides a full set of solutions for learning low-rank Markov models. The main contributions are: (1) We develop statistical methods for learning low-rank Markov model with rate-optimal Kullback-Leiber divergence guarantee for the first time; (2) We develop low-rank optimization methods that are tailored to the computation problems for nuclear-norm regularized and rank-constrained M-estimation; (3) A byproduct is a generalized DC algorithm that applies to nonsmooth nonconvex optimization with convergence guarantee.

The rest of the paper is organized as follows. Section 2 surveys related literature. Section 3 proposes two maximum likelihood estimators based on low-rank optimization and establishes their statistical properties. Section 4 develops computation methods and establishes convergence of the methods. Section 5 presents the results of our numerical experiments.

2. Related literature

Model reduction for complicated systems has a long history. It traces back to variable-resolution dynamic programming (Moore 1991) and state aggregation for decision process (Sutton and Barto 1998). In the case of Markov process, (Deng et al. 2011, Deng and Huang 2012) considered low-rank reduction of Markov models with explicitly known transition probability matrix, but not the estimation of the reduced models. Low-rank matrix approximation has been proved powerful in analysis of large-scale panel data, with numerous applications including network analysis (E et al. 2008), community detection (Newman 2013), ranking (Negahban et al. 2016), product recommendation (Keshavan et al. 2010) and many more. The main goal is to impute corrupted or missing entries of a large data matrix. Statistical theory and computation methods are well understood in the settings where a low-rank signal matrix is corrupted with independent Gaussian noise or its entries are missed independently.

In contrast, our problem is to estimate the transition density functions from dependent state trajectories, where statistical theory and efficient methods are underdeveloped. When the Markov model has rank 1, it becomes an independent process. In this case, our problem reduces to estimation of a discrete distribution from independent samples (Steinhaus 1957, Lehmann and Casella

2006, Han et al. 2015). For a rank-2 transition matrix, Huang et al. (2016) proposed an estimation method using a small number of independent samples. Very recently there have been some works on minimax learning of Markov chains. Hao et al. (2018) derived the minimax rates of estimating a Markov model in terms of a smooth class of f-divergences. They considered the family of α-minorated Markov chains, i.e., all the transition probabilities are greater than α. Wolfer and Kontorovich (2019b) computed the finite-sample PAC-type minimax sample complexity of recovering the transition matrix from a state trajectory of a Markov chain, up to a tolerance in a total-variation-based (TV-based) metric. This TV-based metric does not belong to the family of the smooth f-divergences in Hao et al. (2018), and their class of Markov models strictly contains the class of the α-minorated ones. Neither of these works considered low-rank Markov models though.

The closest work to ours is Zhang and Wang (2017), in which a spectral method via truncated singular value decomposition was introduced and the upper and lower error bounds in terms of total variation were established. Yang et al. (2017) developed an online stochastic gradient method for computing the leading singular space of a transition matrix from random walk data. To our best knowledge, none of the existing works has analyzed efficient recovery of a low-rank Markov model with Kullback-Leiber divergence guarantee.

Hidden Markov Models (HMMs) are closely related with our low-rank Markov models. Note that the observation trajectory of an HMM is not necessarily Markovian. Therefore, an HMM can be regarded as a relaxed variant of low-rank Markov models. There have been many works on estimating HMM, in particular through spectral approaches, e.g., Hsu et al. (2012) and Anandkumar et al. (2014). A critical difference is: States are not fully observable in HMM, but are fully observable in low-rank Markov models. Although HMM is more general, the low-rank Markov model is more suitable for dynamical processes where the state space is large but fully observable, for which we will establish tighter error bounds.

On the optimization side, we adopt DC programming to handle the rank constraint and replace it with the difference of two convex functions. DC programming was first introduced by Pham Dinh and Le Thi (1997) and has become a prominent tool for handling a class of nonconvex optimization problems (see also Pham Dinh and Le Thi (2005), Le Thi et al. (2012, 2017), Le Thi and Pham Dinh (2018)). In particular, Van Dinh et al. (2015) and Wen et al. (2017) considered the majorized DC algorithm, which motivated the optimization method developed in this paper. However, both Van Dinh et al. (2015) and Wen et al. (2017) used the majorization technique with restricted choices of majorants, and neither considered the introduction of the indefinite proximal terms. In addition, Wen et al. (2017) further assumes the smooth part in the objective to be convex. In comparison with the existing methods, our DC programming method applies to nonsmooth problems and is compatible with a more flexible and possibly indefinite proximal term.

Finally, we would like to mention the probabilistic tools we used to derive the statistical results. Recent years have witnessed many works on measure concentration of dependent random variables, e.g., Marton (1996), Kontorovich (2007), Kontorovich and Ramanan (2008), Paulin (2015), Jiang et al. (2018), etc. Nevertheless, these results do not suffice to establish the desired statistical guarantee, because exploiting low-rank structure requires studying the concentration of a matrix martingale in terms of the spectral norm, as shown in Lemma 2. The matrix Freedman inequality (Tropp 2011, Corollary 1.3) turns out to be the right tool for analyzing the concentration of the matrix martingale. We also used an variant of Bernstein's inequality for general Markov chains (Jiang et al. 2018, Theorem 1.2) to derive an exponential tail bound for the status counts of the Markov chain $\mathcal{X}$.

3. Minimax rate-optimal estimation of low-rank Markov chains

Consider an ergodic Markov chain $\mathcal{X} = \{X_0, X_1, \ldots, X_n\}$ on p states $\mathcal{S} = \{s_j\}_{j=1}^p$ with the transition probability matrix $\mathbf{P} \in \mathbb{R}^{p\times p}$ and stationary distribution π, where $P_{ij} = \mathbb{P}(X_1 = s_j | X_0 = s_i)$ for any $i, j \in [p]$. Let $\pi_{\min} := \min_{j\in[p]} \pi_j$ and $\pi_{\max} := \max_{j\in[p]} \pi_j$. We quantify the distance between two transition matrices $\mathbf{P}$ and $\widehat{\mathbf{P}}$ in Frobenius norm $\|\widehat{\mathbf{P}} - \mathbf{P}\|_{\mathrm{F}} = \left\{\sum_{i,j=1}^p (\widehat{P}_{ij} - P_{ij})^2\right\}^{1/2}$ and Kullback–Leibler divergence $D_{\mathrm{KL}}(\mathbf{P}, \widehat{\mathbf{P}}) = \sum_{i,j=1}^p \pi_i P_{ij} \log(P_{ij}/\widehat{P}_{ij}) 1_{\{P_{ij}\neq 0\}}$. Suppose that the unknown transition matrix $\mathbf{P}$ has a small rank $r \ll p$. Our goal is to estimate the transition matrix $\mathbf{P}$ via a state trajectory of length n.

3.1. Spectral gap of nonreversible Markov chains

We first introduce the *right $\mathcal{L}_2$-spectral gap* of $\mathbf{P}$ (Fill 1991, Jiang et al. 2018), a quantity that determines the convergence speed of the Markov chain $\mathcal{X}$ to its invariant distribution π. Let $\mathcal{L}_2(\pi) := \{h \in \Re^p : \sum_{j\in[p]} h_j^2 \pi_j < \infty\}$ be a Hilbert space endowed with the following inner product:

$$\langle h_1, h_2 \rangle_\pi := \sum_{j\in[p]} h_{1j} h_{2j} \pi_j.$$

The matrix $\mathbf{P}$ induces a linear operator on $\mathcal{L}_2(\pi)$: $h \mapsto \mathbf{P}h$, which we abuse $\mathbf{P}$ to denote. Let $\mathbf{P}^*$ be the adjoint operator of $\mathbf{P}$ with respect to $\mathcal{L}_2(\pi)$:

$$\mathbf{P}^* = \mathrm{Diag}(\pi)^{-1} \mathbf{P}^\top \mathrm{Diag}(\pi).$$

Note that the following four statements are equivalent: (a) $\mathbf{P}$ is self-adjoint; (b) $\mathbf{P}^* = \mathbf{P}$; (c) the detailed balance condition holds: $\pi_i P_{ij} = \pi_j P_{ji}$; (d) the Markov chain is reversible. In our analysis, we do *not* require the Markov chain to be reversible. We therefore introduce the *additive reversiblization of* $\mathbf{P}$: $(\mathbf{P} + \mathbf{P}^*)/2$, which is a self-adjoint operator on $\mathcal{L}_2(\pi)$ and has the largest eigenvalue as 1. The right spectral gap of $\mathbf{P}$ is defined as follows:

Definition 1 (Right $\mathcal{L}_2$-spectral gap). We say the right $\mathcal{L}_2$-spectral gap of $\mathbf{P}$ is $1 - \rho_+$ if

$$\rho_+ := \sup_{\langle h,1\rangle_\pi = 0, \langle h,h\rangle_\pi = 1} \frac{1}{2} \langle (\mathbf{P} + \mathbf{P}^*) h, h \rangle_\pi < 1,$$

where 1 in $\langle h, 1\rangle$ refers to the all-one p-dimensional vector.

Define the ϵ-mixing time of the Markov chain $\mathcal{X}$ as

$$\tau(\epsilon) := \min\{t : \max_{j\in[p]} \|(\mathbf{P}^t)_{j\cdot} - \pi\|_{\mathrm{TV}} \le \epsilon\},$$

where $\|(\mathbf{P}^t)_{j\cdot} - \pi\|_{\mathrm{TV}} := 2^{-1}\|(\mathbf{P}^t)_{j\cdot} - \pi\|_1$ is the total variation distance between $\mathbf{P}^t_{j\cdot}$ and π. For reversible and ergodic Markov chains, Levin and Peres (2017, Theorem 12.3) show that

$$\tau(\epsilon) \le \frac{1}{1-\rho_+} \log\left(\frac{1}{\epsilon \pi_{\min}}\right), \tag{1}$$

which implies that the larger the spectral gap is, the faster the Markov chain converges to the stationary distribution.

3.2. Estimation methods and statistical results

Now we are in position to present our methods and statistical results. Given the trajectory $\{X_1,\ldots,X_n\}$, we count the number of times that the state s_i transitions to s_j:

$$n_{ij} := |\{1\le k\le n \mid X_{k-1}=s_i, X_k=s_j\}|\,.$$

Let $n_i := \sum_{j=1}^p n_{ij}$ for $i=1,\ldots,p$ and $n:=\sum_{i=1}^p n_i$. The averaged negative log-likelihood function of $\mathbf{P}$ based on the state-transition trajectory $\{x_0,\ldots,x_n\}$ is

$$\ell_n(\mathbf{P}) := -\frac{1}{n}\sum_{k=1}^n \log(\langle \mathbf{P},\mathbf{X}_k\rangle) = -\frac{1}{n}\sum_{i=1}^p\sum_{j=1}^p n_{ij}\log(P_{ij}), \tag{2}$$

where $\mathbf{X}_k := e_i e_j^\top \in \Re^{p\times p}$ if $x_k=s_i$ and $x_{k+1}=s_j$. We first impose the following assumptions on $\mathbf{P}$ and π.

Assumption 1. *(i)* $\mathrm{rank}(\mathbf{P})=r$; *(ii) There exist some positive constants* $\alpha,\beta>0$ *such that for any* $1\le j,k\le p$, $P_{jk}\in\{0\}\cup[\alpha/p,\beta/p]$.

Remark 1. The entrywise constraints on $\mathbf{P}$ are imposed by our theoretical analysis and may not be necessary in practice. Specifically, the upper and lower bounds for the nonzero entries of $\mathbf{P}$ ensure that (i) the gradient of the log-likelihood $\nabla\ell_n(\mathbf{P})$ is well controlled and exhibits exponential concentration around its population mean (see (EC.9) for the reason we need α there); (ii) the converter between the ℓ_2-risk $\|\widehat{\mathbf{P}}-\mathbf{P}\|_{\mathrm{F}}$ ($\|\widehat{\mathbf{P}}^r-\mathbf{P}\|_{\mathrm{F}}$ resp.) and the KL-divergence $D_{\mathrm{KL}}(\mathbf{P},\widehat{\mathbf{P}})$ ($D_{\mathrm{KL}}(\mathbf{P},\widehat{\mathbf{P}}^r)$ resp.) depends on α and β, as per Lemma EC.1. The entry-wise upper and lower bounds are common in statistical analysis of count data, e.g., Poisson matrix completion (Cao and Xie 2016, Equation (10)), Poisson sparse regression (Jiang et al. 2015, Assumption 2.1), point autoregressive model (Hall et al. 2016, Definition of $\mathcal{A}_s$), etc.

Remark 2. If we remove 0 in the feasible range of P_{jk}, we obtain the (α/p)-minoration condition: $P_{jk}\ge\alpha/p$ for all $j,k\in[p]$. The (α/p)-minoration condition implies strong mixing since combining Brémaud (1999, pp. 237-238) and Kontorovich (2007, Lemma 2.2.2) yields $1-\rho_+\ge\alpha$ and we can deduce that $\tau(\epsilon)\le\alpha^{-1}\log\{(\epsilon\pi_{\min})^{-1}\}$ given (1).

Next we propose and analyze a nuclear-norm regularized maximum likelihood estimator (MLE) of $\mathbf{P}$ defined as follows:

$$\begin{aligned} \widehat{\mathbf{P}} &:= \arg\min\ \ell_n(\mathbf{Q}) + \lambda \|\mathbf{Q}\|_* \\ &\text{s.t. } \mathbf{Q}1_p = 1_p, \quad \alpha/p \le Q_{ij} \le \beta/p, \quad \forall\, 1 \le i,j \le p, \end{aligned} \tag{3}$$

where $\lambda > 0$ is a tuning parameter. Note that we cannot allow $\mathbf{Q}$ to have zero entries as in Assumption 1, because otherwise we may have that $\widehat{P}_{ij} = 0$ and $P_{ij} > 0$ for some (i,j), violating the requirement of the definition of $D_{\mathrm{KL}}(\mathbf{P}, \widehat{\mathbf{P}})$. Our first theorem shows that with an appropriate choice of λ, $\widehat{\mathbf{P}}$ exhibits a sharp statistical rate. For simplicity, from now on we say $a \gtrsim b$ ($a \lesssim b$) if there exists a universal constant $c > 0$ ($C > 0$) such that $a \ge cb$ ($a \le Cb$).

THEOREM 1 **(Statistical guarantee for the nuclear-norm regularized estimator)**. *Suppose the initial state X_0 is drawn from the stationary distribution π and Assumption 1 holds. There exists a universal constant $C_1 > 0$, such that for any $\xi > 1$, if we choose*

$$\lambda = C_1 \left\{ \left(\frac{\xi p^2 \pi_{\max} \log p}{n\alpha} \right)^{1/2} + \frac{\xi p \log p}{n\alpha} \right\},$$

then whenever $n\pi_{\max}(1-\rho_+) \ge \max\{\max(20, \xi^2)\log p, \log n\}$, we have that

$$\mathbb{P}\left(D_{\mathrm{KL}}(\mathbf{P}, \widehat{\mathbf{P}}) \gtrsim \frac{\xi r \pi_{\max} \beta^2 p \log p}{\pi_{\min} \alpha^3 n} + \frac{\xi \pi_{\min}}{r p \pi_{\max} \log p} \right) \lesssim e^{-\xi} + p^{-(\xi-1)} + p^{-10},$$

and that

$$\mathbb{P}\left(\|\widehat{\mathbf{P}} - \mathbf{P}\|_{\mathrm{F}}^2 \gtrsim \frac{\xi r \pi_{\max} \beta^4 \log p}{\pi_{\min}^2 \alpha^4 n} + \frac{\xi \beta^2}{\alpha r p^2 \pi_{\max} \log p} \right) \lesssim e^{-\xi} + p^{-(\xi-1)} + p^{-10}.$$

REMARK 3. When $n \lesssim \{ r p \pi_{\max} (\log p) \beta / (\pi_{\min} \alpha^{3/2}) \}^2$, the second terms of both the KL-Divergence and Frobenius-norm error bounds are dominated by the first terms respectively, so that

$$D_{\mathrm{KL}}(\mathbf{P}, \widehat{\mathbf{P}}) = O_{\mathbb{P}}\left(\frac{r \pi_{\max} \beta^2 p \log p}{\pi_{\min} \alpha^3 n} \right) \text{ and } \|\widehat{\mathbf{P}} - \mathbf{P}\|_{\mathrm{F}}^2 = O_{\mathbb{P}}\left(\frac{r \pi_{\max} \beta^4 \log p}{\pi_{\min}^2 \alpha^4 n} \right). \tag{4}$$

When $\alpha \asymp \beta$ and $\pi_{\max} \asymp \pi_{\min}$, we have that $\alpha, \beta \asymp 1$ and that $\pi_{\max}, \pi_{\min} \asymp 1/p$. Therefore, $D_{\mathrm{KL}}(\mathbf{P}, \widehat{\mathbf{P}}) = O_{\mathbb{P}}(rp \log p / n)$ and $\|\widehat{\mathbf{P}} - \mathbf{P}\|_{\mathrm{F}}^2 = O_{\mathbb{P}}(rp \log p / n)$. These rates are consistent with those derived in the literature of low-rank matrix estimation (Negahban and Wainwright 2011, Koltchinskii et al. 2011). For a big n, the current error bounds are sub-optimal: the second terms of the bounds are independent of n and thus do not converge to zero as n goes to infinity. These terms are due to the requirement of the uniform concentration of $\widetilde{D}_{\mathrm{KL}}(\mathbf{P}, \widehat{\mathbf{P}})$, the empirical counterpart of $D_{\mathrm{KL}}(\mathbf{P}, \widehat{\mathbf{P}})$, to $D_{\mathrm{KL}}(\mathbf{P}, \widehat{\mathbf{P}})$ (see Lemma 3). We eliminate these trailing terms through an alternative proof strategy in Section EC.9, though the resulting statistical rates have worse dependence on α and β and are thus relegated to the appendix.

REMARK 4. When $r = 1$, $\mathbf{P}$ can be written as $1v^\top$ for some vector $v \in \Re^p$, and then estimating $\mathbf{P}$ essentially reduces to estimating a discrete distribution from multinomial count data. The first term of the upper bounds in Theorem 1 nearly matches (up to a log factor) the classical results of discrete distribution estimation ℓ_2 risks (see, e.g., Lehmann and Casella (2006, Pg. 349)).

Next we move on to the second approach – using rank-constrained MLE to estimate $\mathbf{P}$:

$$\begin{aligned} &\widehat{\mathbf{P}}^r := \arg\min \ell_n(\mathbf{Q}) \\ &\quad \text{s.t. } \mathbf{Q}1_p = 1_p, \quad \alpha/p \le Q_{ij} \le \beta/p, \quad \forall\, 1 \le i, j \le p, \quad \operatorname{rank}(\mathbf{Q}) \le r. \end{aligned} \tag{5}$$

Similarly to (3), we cannot allow $\mathbf{Q}$ to have zero entries. In contrast to $\widehat{\mathbf{P}}$, the rank-constrained MLE $\widehat{\mathbf{P}}^r$ enforces the prior knowledge "$\mathbf{P}$ is low-rank" exactly without inducing any additional bias. It requires solving a non-convex and non-smooth optimization problem, for which we will provide an algorithm based on DC programming in Section 4.2. Here we first present its statistical guarantee.

THEOREM 2 **(Statistical guarantee for the rank-constrained estimator)**. *Suppose that Assumption 1 holds and that* $n\pi_{\max}(1-\rho_+) > \max(20\log p, \log n)$. *There exist universal constants* $C_1, C_2 > 0$ *such that for any* $\xi > 0$,

$$\mathbb{P}\left\{ D_{\mathrm{KL}}(\mathbf{P}, \widehat{\mathbf{P}}^r) \ge \max\left(\frac{C_1 r \pi_{\max} \beta^2 p \log p}{\pi_{\min}\alpha^3 n}, \frac{\xi \pi_{\min}}{r p \pi_{\max} \log p} \right) \right\} \le C_2 e^{-\xi},$$

and

$$\mathbb{P}\left\{ \|\widehat{\mathbf{P}}^r - \mathbf{P}\|_{\mathrm{F}}^2 \ge \max\left(\frac{C_1 r \pi_{\max} \beta^4 \log p}{\pi_{\min}^2 \alpha^4 n}, \frac{\xi \beta^2}{\alpha r p^2 \pi_{\max} \log p} \right) \right\} \le C_2 e^{-\xi}.$$

Remark 5. The proof of the rank constrained method requires fewer inequality steps and is more straightforward than the that of the nuclear method. Although our upper bounds of the nuclear norm regularized method and the rank constrained one have the same rate, the difference of their proofs may implicitly suggest the advantage of the rank constrained method in the constant, as futher illustrated by our numerical studies.

To assess the quality of the established statistical guarantee, we further provide a lower bound result below. It shows that when α, β are constants, both estimators $\widehat{\mathbf{P}}$ and $\widehat{\mathbf{P}}^r$ are rate-optimal up to a logarithmic factor. Informally speaking, they are not improvable for estimating the class of rank-r Markov chains.

Theorem 3 **(Minimax error lower bound for estimating low-rank Markov models)**. *Consider the following set of low-rank transition matrices*

$$\Theta := \left\{\mathbf{P} : \forall j,k \in [p], P_{jk} \in \{0\} \cup [\alpha/p, +\infty),\ \mathbf{P}1_p = 1_p,\ \operatorname{rank}(\mathbf{P}) \le r\right\}.$$

There exists a universal constant $c > 0$ such that when $p(r-1) \ge 192 \log 2$, we have

$$\inf_{\widehat{\mathbf{P}}} \sup_{\mathbf{P} \in \Theta} \mathbb{E}\|\widehat{\mathbf{P}} - \mathbf{P}\|_F^2 \ge \frac{cp(r-1)}{n\alpha}.$$

Remark 6. Theorem 3 shows that a smaller α makes the estimation problem harder. It still remains to be an open problem whether β in Assumption 1 should be in this minimax risk or not.

Besides the full transition matrix $\mathbf{P}$, the leading left and right singular vectors of $\mathbf{P}$, denoted by $\mathbf{U}, \mathbf{V} \in \mathbb{O}^{p \times r}$, also play important roles in Markov chain analysis. For example, performing k-means on reliable estimate of $\mathbf{U}$ or $\mathbf{V}$ can give rise to state aggregation of the Markov chain (Zhang and Wang 2017). In the following, we further establish the statistical rate of estimating the singular subspace of the Markov transition matrix, based on the previous results.

Theorem 4. *Under the setting of Theorem 1, let $\widehat{\mathbf{U}}, \widehat{\mathbf{V}} \in \mathbb{O}^{p \times r}$ be the left and right singular vectors of $\widehat{\mathbf{P}}$ respectively. Then there exist universal constants C_1, C_2, such that for any $\xi > 0$, we have that*

$$\max\left\{\|\sin\Theta(\widehat{\mathbf{U}}, \mathbf{U})\|_F^2, \|\sin\Theta(\widehat{\mathbf{V}}, \mathbf{V})\|_F^2\right\} \le \min\left\{\max\left(\frac{Cr\pi_{\max}\beta^4 \log p}{\pi_{\min}^2 \alpha^4 n \sigma_r^2(\mathbf{P})}, \frac{\xi\beta^2}{\alpha r p^2 \pi_{\max} (\log p) \sigma_r^2(\mathbf{P})}\right), r\right\}$$

with probability at least $1-C_2(e^{-\xi}+p^{-(\xi-1)}+p^{-10})$. *Here,* $\sigma_r(\mathbf{P})$ *is the* r*-th largest singular value of* $\mathbf{P}$ *and* $\|\sin\Theta(\widehat{\mathbf{U}},\mathbf{U})\|_F := (r-\|\widehat{\mathbf{U}}^\top\mathbf{U}\|_F^2)^{1/2}$ *is the Frobenius norm* $\sin\Theta$ *distance between* $\widehat{\mathbf{U}}$ *and* $\mathbf{U}$.

3.3. Proof outline of Theorems 1, 2

In this section, we elucidate the roadmap to proving Theorems 1 and 2. Complete proofs are deferred to the supplementary materials. We mainly focus on Theorem 1 for the nuclear-norm penalized MLE $\widehat{\mathbf{P}}$, as we use similar strategies to prove Theorem 2.

We first show in the forthcoming Lemma 1 that when the regularization parameter λ is sufficiently large, the statistical error $\widehat{\boldsymbol{\Delta}} := \widehat{\mathbf{P}} - \mathbf{P}$ falls in a restricted nuclear-norm cone. This cone structure is crucial to establishing strong statistical guarantee for estimation of low-rank matrices with high-dimensional scaling (Negahban and Wainwright 2011). Define a linear subspace $\mathcal{N} := \{\mathbf{Q} : \mathbf{Q}1_p = 1_p\}$ and denote the corresponding projection operator by $\Pi_{\mathcal{N}}$. In other words, for any $\mathbf{Q}\in\mathcal{N}$ and any $j=1,\ldots,p$, the summation of all the entries in the jth row of $\mathbf{Q}$ equals one. One can verify that for any $\mathbf{Q}\in\Re^{p\times p}$, $\Pi_{\mathcal{N}}(\mathbf{Q}) = \mathbf{Q}-\mathbf{Q}11^\top/p$. Let $\mathbf{P}=\mathbf{U}\mathbf{D}\mathbf{V}^\top$ be an SVD of $\mathbf{P}$, where $\mathbf{U},\mathbf{V}\in\Re^{p\times r}$ are orthonormal and the diagonals of $\mathbf{D}$ are in the non-increasing order. Define

$$\mathcal{M} := \{\mathbf{Q}\in\Re^{p\times p} \mid \mathrm{row}(\mathbf{Q})\subseteq\mathrm{col}(\mathbf{V}), \mathrm{col}(\mathbf{Q})\subseteq\mathrm{col}(\mathbf{U})\},$$

$$\overline{\mathcal{M}}^\perp := \{\mathbf{Q}\in\Re^{p\times p} \mid \mathrm{row}(\mathbf{Q})\perp\mathrm{col}(\mathbf{V}), \mathrm{col}(\mathbf{Q})\perp\mathrm{col}(\mathbf{U})\},$$

where $\mathrm{col}(\cdot)$ and $\mathrm{row}(\cdot)$ denote the column space and row space respectively. We can write any $\boldsymbol{\Delta}\in\Re^{p\times p}$ as

$$\boldsymbol{\Delta} = [\mathbf{U},\mathbf{U}^\perp]\begin{bmatrix}\boldsymbol{\Gamma}_{11} & \boldsymbol{\Gamma}_{12}\\ \boldsymbol{\Gamma}_{21} & \boldsymbol{\Gamma}_{22}\end{bmatrix}[\mathbf{V},\mathbf{V}^\perp]^\top.$$

Define $\boldsymbol{\Delta}_{\mathcal{W}}$ as the projection of $\boldsymbol{\Delta}$ onto any Hilbert space $\mathcal{W}\subseteq\Re^{p\times p}$. Then,

$$\boldsymbol{\Delta}_{\mathcal{M}} = \mathbf{U}\boldsymbol{\Gamma}_{11}\mathbf{V}^\top,\quad \boldsymbol{\Delta}_{\overline{\mathcal{M}}^\perp} = \mathbf{U}^\perp\boldsymbol{\Gamma}_{22}(\mathbf{V}^\perp)^\top,\quad \boldsymbol{\Delta}_{\overline{\mathcal{M}}} = [\mathbf{U},\mathbf{U}^\perp]\begin{bmatrix}\boldsymbol{\Gamma}_{11} & \boldsymbol{\Gamma}_{12}\\ \boldsymbol{\Gamma}_{21} & \mathbf{0}\end{bmatrix}[\mathbf{V},\mathbf{V}^\perp]^\top. \tag{6}$$

The lemma below shows that $\widehat{\boldsymbol{\Delta}} := \widehat{\mathbf{P}} - \mathbf{P}$ falls in a nuclear-norm cone if λ is sufficiently large.

Lemma 1. *If $\lambda \geq 2\|\Pi_{\mathcal{N}}(\nabla \ell_n(\mathbf{P}))\|_2$ in (3), then we have that*

$$\|\widehat{\mathbf{\Delta}}_{\overline{\mathcal{M}}^{\perp}}\|_* \leq 3\|\widehat{\mathbf{\Delta}}_{\overline{\mathcal{M}}}\|_* + 4\|\mathbf{P}_{\mathcal{M}^{\perp}}\|_*.$$

In particular, when $\mathbf{P} \in \mathcal{M}$, we have that $\|\mathbf{P}_{\mathcal{M}^{\perp}}\|_* = 0$ and that

$$\|\widehat{\mathbf{\Delta}}\|_* \leq \|\mathbf{\Delta}_{\overline{\mathcal{M}}^{\perp}}\|_* + \|\mathbf{\Delta}_{\overline{\mathcal{M}}}\|_* \leq 4\|\widehat{\mathbf{\Delta}}_{\overline{\mathcal{M}}}\|_* \leq 4(2r)^{1/2}\|\widehat{\mathbf{\Delta}}\|_{\mathrm{F}}. \tag{7}$$

Lemma 1 implies that the converting factor between the nuclear and Frobenius norms of $\widehat{\mathbf{\Delta}}$ is merely $4(2r)^{1/2}$ when $\mathbf{P} \in \mathcal{M}$, which is much smaller than the worst-case factor $p^{1/2}$ between nuclear and Frobenius norms of general p-by-p matrices. This property of $\widehat{\mathbf{\Delta}}$ is one cornerstone for establishing Theorem 1.

Next, we derive the rate of $\|\Pi_{\mathcal{N}}(\nabla \ell_n(\mathbf{P}))\|_2$ to determine the order of λ that ensures the condition of Lemma 1 to hold.

Lemma 2. *Under Assumption 1, whenever $n\pi_{\max}(1-\rho_+) \geq 2\log p$, for any $\xi > 1$,*

$$\mathbb{P}\left\{\|\Pi_{\mathcal{N}}(\nabla \ell_n(\mathbf{P}))\|_2 \gtrsim \left(\frac{\xi p^2 \pi_{\max} \log p}{n\alpha}\right)^{1/2} + \frac{\xi p \log p}{n\alpha}\right\} \leq 4p^{-(\xi-1)} + \exp\left(-\frac{n\pi_{\max}(1-\rho_+)}{2}\right).$$

Remark 7. Lemma 2 is essentially due to concentration of a matrix martingale. Many existing results on measure concentration of dependent random variables (Marton 1996, Kontorovich 2007, Kontorovich and Ramanan 2008, Paulin 2015) are not directly applicable because of the matrix structure of $\nabla \ell_n(\mathbf{P})$. The main probabilistic tool we use here is the matrix Freedman inequality (Tropp 2011, Corollary 1.3) that characterizes concentration behavior of a matrix martingale (See (EC.18) for details). We notice two recent works, Wolfer and Kontorovich (2019b) and Wolfer and Kontorovich (2019a), that use the same matrix Freedman inequality. Specifically, Wolfer and Kontorovich (2019a) applied the matrix Freedman inequality to derive a confidence interval for the mixing time of a Markov chain based on its single trajectory, and Wolfer and Kontorovich (2019b) used the same inequality to establish an upper bound for the sample complexity of learning a Markov chain. Finally, we also use an variant of Bernstein's inequality for general Markov chains (Jiang et al. 2018, Theorem 1.2) to derive an exponential tail bound for the status counts of the Markov chain $\mathcal{X}$ (See (EC.16) for details).

Let $\mathcal{C} := \{\mathbf{Q} \in \mathbb{R}^{p\times p} : \|\mathbf{Q}-\mathbf{P}\|_* \le 4\times 2^{1/2}\|\mathbf{Q}-\mathbf{P}\|_{\mathrm{F}}, \mathbf{Q}1_p = 1_p, \alpha/p \le Q_{jk} \le \beta/p, \forall (j,k)\in[p]\times[p]\}$. For any $\mathbf{Q}\in\mathcal{C}$, define $\mathcal{L}(\mathbf{Q}) := \mathbb{E}\{-\log(\langle \mathbf{Q}, \mathbf{X}_i\rangle)\}$ and $\ell_n(\mathbf{Q}) := n^{-1}\sum_{i=1}^n -\log(\langle \mathbf{Q},\mathbf{X}_i\rangle)$. Recall that $D_{\mathrm{KL}}(\mathbf{P},\mathbf{Q}) = \mathcal{L}(\mathbf{Q}) - \mathcal{L}(\mathbf{P}) = \sum_{i=1}^p \pi_i D_{\mathrm{KL}}(P_{i\cdot}, Q_{i\cdot}) = \sum_{i=1}^p\sum_{j=1}^p \pi_i P_{ij}\log(P_{ij}/Q_{ij})$. Define the empirical KL divergence of $\mathbf{Q}$ from $\mathbf{P}$ as

$$\widetilde{D}_{\mathrm{KL}}(\mathbf{P},\mathbf{Q}) := \frac{1}{n}\sum_{i=1}^n \langle \log(\mathbf{P}) - \log(\mathbf{Q}), \mathbf{X}_i\rangle = \ell_n(\mathbf{Q}) - \ell_n(\mathbf{P}).$$

The final ingredient of the analysis is the uniform converegence of $\widetilde{D}_{\mathrm{KL}}(\mathbf{P},\mathbf{Q})$ to $D_{\mathrm{KL}}(\mathbf{P},\mathbf{Q})$ when $D_{\mathrm{KL}}(\mathbf{P},\mathbf{Q})$ is large.

LEMMA 3. *Suppose that* $n\pi_{\max}(1-\rho_+) \ge \max(20\log p, \log n)$*. For any* $\eta > \pi_{\min}/(2\pi_{\max} rp\log p)$*, define* $\mathcal{C}(\eta) := \{\mathbf{Q}\in\mathcal{C} : D_{\mathrm{KL}}(\mathbf{P},\mathbf{Q}) \ge \eta\}$*. Then there exist universal constants* $C_1, C_2 > 0$ *such that*

$$\begin{aligned}\mathbb{P}\bigg\{\forall \mathbf{Q}\in\mathcal{C}(\eta),\ |\widetilde{D}_{\mathrm{KL}}(\mathbf{P},\mathbf{Q}) - D_{\mathrm{KL}}(\mathbf{P},\mathbf{Q})| \le \frac{1}{2}D_{\mathrm{KL}}(\mathbf{P},\mathbf{Q}) + \frac{C_1\pi_{\max}\beta^2 rp\log p}{\pi_{\min}\alpha^3 n}\bigg\} \\ \ge 1 - C_2\exp\bigg(-\frac{\eta\pi_{\max} rp\log p}{\pi_{\min}}\bigg).\end{aligned} \tag{8}$$

Theorem 1 follows immediatley after one combines Lemmas 1, 2 and 3. As for the rank-constrained MLE $\widehat{\mathbf{P}}^r$, let $\widehat{\mathbf{\Delta}}(r) := \widehat{\mathbf{P}}^r - \mathbf{P}$. Note that the rank constraint in (5) implies that $\mathrm{rank}(\widehat{\mathbf{\Delta}}(r)) \le 2r$. Thus, $\|\widehat{\mathbf{\Delta}}(r)\|_* \le (2r)^{1/2}\|\widehat{\mathbf{\Delta}}(r)\|_{\mathrm{F}}$ and Lemma 3 remains applicable in the statistical anlaysis of $\widehat{\mathbf{P}}^r$.

4. Computing Markov models using low-rank optimization

In this section we develop efficient optimization methods to compute the proposed estimators for the low-rank Markov model. From now on, we drop the constraint that $\alpha/p \le Q_{ij} \le \beta/p$, which is used only to derive the statistical guarantees. In other words, α and β are motivated by statistical theory, and do not need to be taken into account in the optimization.

4.1. Optimization methods for the nuclear-norm regularized likelihood problem

We first consider the nuclear-norm regularized likelihood problem (3). It is a special case of the following linearly constrained optimization problem:

$$\min\left\{g(\mathbf{X}) + c\|\mathbf{X}\|_* \mid \mathcal{A}(\mathbf{X}) = b\right\}, \tag{9}$$

where $g : \Re^{p\times p} \to (-\infty, +\infty]$ is a closed, convex, but possibly non-smooth function, $\mathcal{A} : \Re^{p\times p} \to \Re^m$ is a linear map, $b \in \Re^m$ and $c > 0$ are given data. If we take $\alpha = 0, \beta = p$ in problem (3), it becomes a special case of the general problem (9) with $g(\mathbf{X}) = -\ell_n(\mathbf{X}) + \delta(\mathbf{X} \geq 0)$, $\mathcal{A}(\mathbf{X}) = \mathbf{X}\mathbf{1}_p$, $b = \mathbf{1}_p$, and $\delta(\cdot)$ being the indicator function.

Despite of its convexity, problem (9) is highly nontrivial due to the nonsmoothness of g and the presence of the nuclear norm regularizer. Here, we propose to solve it via the dual approach. The dual of problem (9) is

$$\begin{aligned} &\min\ g^*(-\mathbf{\Xi}) - \langle b, y\rangle \\ &\text{s.t. } \mathbf{\Xi} + \mathcal{A}^*(y) + \mathbf{S} = 0, \quad \|\mathbf{S}\|_2 \leq c, \end{aligned} \tag{10}$$

where $\|\cdot\|_2$ denotes the spectral norm, and g^* is the conjugate function of g given by

$$g^*(\mathbf{\Xi}) = \sum_{(i,j)\in\Omega} \frac{n_{ij}}{n}(\log\frac{n_{ij}}{n} - 1 - \log(-\Xi_{ij})) + \delta(\Xi \leq 0) \quad \forall \mathbf{\Xi} \in \Re^{p\times p}$$

with $\Omega = \{(i,j) \mid n_{ij} \neq 0\}$ and $\overline{\Omega} = \{(i,j) \mid n_{ij} = 0\}$. Given $\sigma > 0$, the augmented Lagrangian function $\mathcal{L}_\sigma$ associated with (10) is

$$\mathcal{L}_\sigma(\mathbf{\Xi}, y, \mathbf{S}; \mathbf{X}) = g^*(-\mathbf{\Xi}) - \langle b, y\rangle + \frac{\sigma}{2}\|\mathbf{\Xi} + \mathcal{A}^*(y) + \mathbf{S} + \mathbf{X}/\sigma\|^2 - \frac{1}{2\sigma}\|\mathbf{X}\|^2.$$

We consider popular ADMM type methods for solving problem (10) (A comprehensive numerical study has been conducted in (Li et al. 2016b) and justifies our procedure). Since there are three separable blocks in (10) (namely $\mathbf{\Xi}$, y, and $\mathbf{S}$), the direct extended ADMM is not applicable. Indeed, it has been shown in (Chen et al. 2016) that the direct extended ADMM for multi-block convex minimization problem is not necessarily convergent. Fortunately, the functions corresponding to block y in the objective of (10) is linear. Thus we can apply the multi-block symmetric Gauss-Sediel based ADMM (sGS-ADMM) (Li et al. 2016b). In literature (Chen et al. 2017, Ferreira et al. 2017, Lam et al. 2018, Li et al. 2016b, Wang and Zou 2018), extensive numerical experiments demonstrate that sGS-ADMM is not only convergent but also faster than the directly extended multi-block ADMM and its many other variants. Specifically, the algorithmic framework of sGS-ADMM for solving (10) is presented in Algorithm 1.

Next, we discuss how the k-th iteration of Algorithm 1 is performed:

Algorithm 1 An sGS-ADMM for solving (10)

Input: initial point $(\mathbf{\Xi}^0, y^0, \mathbf{S}^0, \mathbf{X}^0)$, penalty parameter $\sigma > 0$, maximum iteration number K, and the step-length $\gamma \in (0, (1+\sqrt{5})/2)$

for $k = 0$ **to** K **do**

$y^{k+\frac{1}{2}} = \arg\min_y \mathcal{L}_\sigma(\mathbf{\Xi}^k, y, \mathbf{S}^k; \mathbf{X}^k)$

$\mathbf{\Xi}^{k+1} = \arg\min_{\mathbf{\Xi}} \mathcal{L}_\sigma(\mathbf{\Xi}, y^{k+\frac{1}{2}}, \mathbf{S}^k; \mathbf{X}^k)$

$y^{k+1} = \arg\min_y \mathcal{L}_\sigma(\mathbf{\Xi}^{k+1}, y, \mathbf{S}^k; \mathbf{X}^k)$

$\mathbf{S}^{k+1} = \arg\min_{\mathbf{S}} \mathcal{L}_\sigma(\mathbf{\Xi}^{k+1}, y^{k+1}, \mathbf{S}, ; \mathbf{X}^k)$

$\mathbf{X}^{k+1} = \mathbf{X}^k + \gamma\sigma(\mathbf{\Xi}^{k+1} + \mathcal{A}^*(y^{k+1}) + \mathbf{S}^{k+1})$

end for

Computation of $y^{k+\frac{1}{2}}$ and y^{k+1}. Simple calculations show that $y^{k+\frac{1}{2}}$ and y^{k+1} can be obtained by solving the following linear systems:

$$\begin{cases} y^{k+\frac{1}{2}} = \dfrac{1}{\sigma}(\mathcal{A}\mathcal{A}^*)^{-1}\big(b - X^k - \sigma(\mathbf{\Xi}^k + \mathbf{S}^k)\big), \\ y^{k+1} = \dfrac{1}{\sigma}(\mathcal{A}\mathcal{A}^*)^{-1}\big(b - X^k - \sigma(\mathbf{\Xi}^{k+1} + \mathbf{S}^k)\big). \end{cases}$$

In our estimation problem, it is not difficult to verify that $\mathcal{A}\mathcal{A}^* y = py$ for any $y \in \Re^p$. Thanks to this special structure, the above formulas can be further reduced to

$$y^{k+\frac{1}{2}} = \frac{1}{\sigma p}\big(b - \mathbf{X}^k - \sigma(\mathbf{\Xi}^k + \mathbf{S}^k)\big) \text{ and } y^{k+1} = \frac{1}{\sigma p}\big(b - \mathbf{X}^k - \sigma(\mathbf{\Xi}^{k+1} + \mathbf{S}^k)\big).$$

Computation of $\mathbf{\Xi}^{k+1}$. To compute $\mathbf{\Xi}^{k+1}$, we need to solve the following optimization problem:

$$\min_{\mathbf{\Xi}} \left\{ g^*(-\mathbf{\Xi}) + \frac{\sigma}{2}\|\mathbf{\Xi} + \mathbf{R}^k\|^2 \right\},$$

where $\mathbf{R}^k \in \Re^{p\times p}$ is given. Careful calculations, together with the Moreau identity (Rockafellar 2015, Theorem 31.5), show that

$$\mathbf{\Xi}^{k+1} = \frac{1}{\sigma}[\mathbf{Z}^k - \sigma\mathbf{R}^k] \text{ and } \mathbf{Z}^k = \arg\min_{\mathbf{Z}} \left\{ \sigma g(\mathbf{Z}) + \frac{1}{2}\|\mathbf{Z} - \sigma\mathbf{R}^k\|^2 \right\}.$$

For our estimation problem, i.e., $g(\mathbf{X}) = \ell_n(\mathbf{X}) + \delta(\mathbf{X} \geq 0)$, it is easy to see that Z^k admits the following form:

$$Z^k_{ij} = \frac{\sigma R^k_{ij} + \sigma\sqrt{(R^k_{ij})^2 + 4n_{ij}/(n\sigma)}}{2} \quad \text{if } (i,j) \in \Omega \text{ and } Z^k_{ij} = \sigma \max(R^k_{ij}, 0) \quad \text{if } (i,j) \in \overline{\Omega}.$$

Computation of $\mathbf{S}^{k+1}$. The computation of $\mathbf{S}^{k+1}$ can be simplified as:

$$\mathbf{S}^{k+1} = \arg\min_{\mathbf{S}} \left\{ \frac{\sigma}{2} \|\mathbf{S} + \mathbf{\Xi}^{k+1} + \mathcal{A}^* y^{k+1} + \mathbf{X}^k/\sigma\|^2 \mid \|\mathbf{S}\|_2 \le c \right\}.$$

Let $\mathbf{W}_k := -(\mathbf{\Xi}^{k+1} + \mathcal{A}^* y^{k+1} + X^k/\sigma)$ admit the following singular value decomposition (SVD) $\mathbf{W}_k = \mathbf{U}_k \mathbf{\Sigma}_k \mathbf{V}_k^\top$, where $\mathbf{U}_k$ and $\mathbf{V}_k$ are orthogonal matrices, $\mathbf{\Sigma}_k = \mathrm{Diag}(\alpha_1^k, \ldots, \alpha_p^k)$ is the diagonal matrix of singular values of W_k, with $\alpha_1^k \ge \ldots \ge \alpha_p^k \ge 0$. Then, by Lemma 2.1 in (Jiang et al. 2014), we know that

$$\mathbf{S}^{k+1} = \mathbf{U}_k \min(\mathbf{\Sigma}_k, c) \mathbf{V}_k^\top,$$

where $\min(\mathbf{\Sigma}_k, c) = \mathrm{Diag}\big(\min(\alpha_1^k, c), \ldots, \min(\alpha_p^k, c)\big)$. We also note that in the implementation, only partial SVD, which is much cheaper than full SVD, is needed as $r \ll p$.

The nontrivial convergence results and the sublinear non-ergodic iteration complexity of Algorithm 1 can be obtained from Li et al. (2016b) and Chen et al. (2017). We put the convergence theorem and a sketch of the proof in the supplementary material.

4.2. Optimization methods for the rank-constrained likelihood problem

Next we develop the optimization method for computing the rank-constrained likelihood maximizer from (5). In Subsection 4.2.1, a penalty approach is applied to transform the original intractable rank-constrained problem into a DC programming problem. Then we solve this problem by a proximal DC (PDC) algorithm in Subsection 4.2.2. We also discuss the solver for the subproblems involved in the proximal DC algorithm. Lastly, a unified convergence analysis of a class of majorized indefinite-proximal DC (Majorized iPDC) algorithms is provided in Subsection 4.2.3.

4.2.1. A penalty approach for problem (5)**.** Recall (5) is intractable due to the non-convex rank constraint, we introduce a penalty approach to relax. We particularly study the following optimization problem:

$$\min \left\{ f(\mathbf{X}) \mid \mathcal{A}(\mathbf{X}) = b, \, \mathrm{rank}(\mathbf{X}) \le r \right\}, \tag{11}$$

where $f : \Re^{p\times p} \to (-\infty, +\infty]$ is a closed proper convex, but possibly non-smooth, function. The original rank-constraint maximum likelihood problem (5) can be viewed as a special case of the general model (11).

Given $\mathbf{X} \in \Re^{p\times p}$, let $\sigma_1(\mathbf{X}) \geq \cdots \geq \sigma_p(\mathbf{X}) \geq 0$ be the singular values of $\mathbf{X}$. Since $\operatorname{rank}(\mathbf{X}) \leq r$ if and only $\sigma_{r+1}(\mathbf{X}) + \ldots + \sigma_p(\mathbf{X}) = \|\mathbf{X}\|_* - \|\mathbf{X}\|_{(r)} = 0$ ($\|\mathbf{X}\|_{(r)} = \sum_{i=1}^r \sigma_i(\mathbf{X})$ is the Ky Fan r-norm of $\mathbf{X}$), (11) can be equivalently formulated as

$$\min\left\{ f(\mathbf{X}) \mid \|\mathbf{X}\|_* - \|\mathbf{X}\|_{(r)} = 0,\, \mathcal{A}(\mathbf{X}) = b \right\}.$$

See also (Gao and Sun 2010, Equation (29)). The penalized formulation of problem (11) is

$$\min\left\{ f(\mathbf{X}) + c(\|\mathbf{X}\|_* - \|\mathbf{X}\|_{(r)}) \mid \mathcal{A}(\mathbf{X}) = b \right\}, \tag{12}$$

where $c > 0$ is a penalty parameter. Since $\|\cdot\|_{(r)}$ is convex, the objective in problem (12) is a difference of two convex functions: $f(\mathbf{X}) + c\|\mathbf{X}\|_*$ and $c\|\mathbf{X}\|_{(r)}$, i.e., (12) is a DC program.

Let $\mathbf{X}_c^*$ be an optimal solution to the penalized problem (12). The following proposition shows that $\mathbf{X}_c^*$ is also the optimizer to (11) when it is low-rank.

PROPOSITION 1. *If* $\operatorname{rank}(\mathbf{X}_c^*) \leq r$*, then* $\mathbf{X}_c^*$ *is also an optimal solution to the original problem* (11).

In practice, one can gradually increase the penalty parameter c to obtain a sufficient low rank solution $\mathbf{X}_c^*$. In our numerical experiments, we can obtain solutions with the desired rank with a properly chosen parameter c.

4.2.2. A PDC algorithm for the penalized problem (12)**.** The central idea of the DC algorithm (Pham Dinh and Le Thi 1997) is as follows: at each iteration, one approximates the concave part of the objective function by its affine majorant, then solves the resulting convex optimization problem. In this subsection, we present a variant of the classic DC algorithm for solving (12). For the execution of the algorithm, we recall that the sub-gradient of Ky Fan r-norm at a point $\mathbf{X} \in \Re^{p\times p}$ (Watson 1993) is

$$\partial\|\mathbf{X}\|_{(r)} = \left\{ \mathbf{U}\,\mathrm{Diag}(q^*)\mathbf{V}^\top \mid q^* \in \Delta \right\},$$

where $\mathbf{U}$ and $\mathbf{V}$ are the singular vectors of $\mathbf{X}$, and Δ is the optimal solution set of the following problem

$$\max_{q\in\Re^p}\left\{\sum_{i=1}^{p}\sigma_i(\mathbf{X})q_i \mid \langle \mathbf{1}_p,\, q\rangle \le r,\, 0\le q\le 1\right\}.$$

Note that one can efficiently obtain a component of $\partial\|\mathbf{X}\|_{(r)}$ by computing the SVD of X and picking up the SVD vectors corresponding to the r largest singular values. After these preparations, we are ready to state the PDC algorithm for problem (12) in Algorithm 2. Different from the classic DC algorithm, an additional proximal term is added to ensure that solutions of subproblems (13) exist and the difference of two consecutive iterations converges. See Theorem 5 and Remark 8 for more details.

Algorithm 2 A PDC algorithm for solving (12)

Given $c>0$, $\alpha\ge 0$, and the stopping tolerance η, choose initial point $\mathbf{X}^0\in\Re^{p\times p}$. Iterate the following steps for $k=0,1,\ldots:$

1. Choose $\mathbf{W}_k\in\partial\|\mathbf{X}^k\|_{(r)}$. Compute

$$\begin{aligned}\mathbf{X}^{k+1} \quad = \quad & \arg\min f(\mathbf{X})+c\left(\|\mathbf{X}\|_*-\langle \mathbf{W}_k,\,\mathbf{X}-\mathbf{X}^k\rangle-\|\mathbf{X}^k\|_{(r)}\right)+\frac{\alpha}{2}\|\mathbf{X}-\mathbf{X}^k\|_F^2\\ & \text{subject to } \mathcal{A}(\mathbf{X})=b.\end{aligned} \tag{13}$$

2. If $\|\mathbf{X}^{k+1}-\mathbf{X}^k\|_F\le\eta$, stop.

We say that $\mathbf{X}$ is a critical point of problem (12) if

$$\partial(f(\mathbf{X})+c\|\mathbf{X}\|_*+\delta(\mathcal{A}(X)=b))\cap(c\partial\|\mathbf{X}\|_{(r)})\neq\emptyset.$$

We have the following convergence results for Algorithm 2.

THEOREM 5 **(Convergence of Algorithm 2)**. *Let $\{\mathbf{X}^k\}$ be the sequence generated by Algorithm 2 and $\alpha\ge0$. Then $\{f(\mathbf{X}^k)+c(\|\mathbf{X}^k\|_*-\|\mathbf{X}^k\|_{(r)})\}$ is a non-increasing sequence. If $\mathbf{X}^{k+1}=\mathbf{X}^k$ for some integer $k\ge0$, then $\mathbf{X}^k$ is a critical point of (12). Otherwise, it holds that*

$$\left(f(\mathbf{X}^{k+1})+c(\|\mathbf{X}^{k+1}\|_*-\|\mathbf{X}^{k+1}\|_{(r)})\right)-\left(f(\mathbf{X}^k)+c(\|\mathbf{X}^k\|_*-\|\mathbf{X}^k\|_{(r)})\right)\le-\frac{\alpha}{2}\|\mathbf{X}^{k+1}-\mathbf{X}^k\|_F^2.$$

Moreover, any accumulation point of the bounded sequence $\{\mathbf{X}^k\}$ *is a critical point of problem* (12)*. In addition, if* $\alpha > 0$*, it holds that* $\lim_{k\to\infty} \|\mathbf{X}^{k+1} - \mathbf{X}^k\|_F = 0$.

Remark 8 (Adjusting Parameters). In practice, a small $\alpha > 0$ is suggested to ensure strict decrease of the objective value and convergence of $\{\|\mathbf{X}^{k+1} - \mathbf{X}^k\|_F\}$; if f is strongly convex, one achieves these nice properties even if $\alpha = 0$ based on the results of Theorem 6. The penalty parameter c can be adaptively adjusted according to the rank of the sequence generated by Algorithm 2.

Remark 9 (Number of iterations of Algorithm 2). Let $\eta > 0$ be the stopping tolerance and F^* be the optimal value of problem (12). By using the inequality in Theorem 5, it can be shown that if $\alpha > 0$, then Algorithm 2 terminates in no more than K iterations, where

$$K = \left\lceil \frac{2\big(f(X^0) + c(\|X^0\|_* - \|X^0\|_{(r)}) - F^*\big)}{\alpha\eta^2} \right\rceil + 1.$$

Remark 10 (Statistical properties). The statistical rate we derived in Theorem 2 does not carry over to the iterates of the DC algorithm here. Though we show in Theorem 5 that the DC algorithm can converge to a critical point, it remains unclear whethere this point is close to the global optimum and provably enjoys the statistical guarantee. Recently there have been many works conveying positive messages on the statistical properties of the non-convex optimization algorithms. For example, Loh and Wainwright (2015) showed that any stationary point of the composite objective function they considered lies within statistical precision of the true parameter. We hope to establish similar theory for the proposed DC approach in future research.

Next, we discuss how to solve subproblems (13). (13) is still a nuclear norm penalized convex optimization problem and is a special case of model (9) with $g(\mathbf{X}) = f(\mathbf{X}) + \langle \mathbf{W}, \mathbf{X} \rangle + \frac{\alpha}{2}\|\mathbf{X}\|_F^2$. Hence, Algorithm 1 can directly solve these subproblems efficiently. When Algorithm 1 is executed on this new function g, all computations, except for the update of $\mathbf{\Xi}$, have already been discussed in Section 4.1. To update $\mathbf{\Xi}$ in the process of executing Algorithm 1 for solving (13) with $g(\mathbf{X}) =$

$\ell_n(\mathbf{X}) + \delta(\mathbf{X} \geq 0) + \langle \mathbf{W}, \mathbf{X} \rangle + \frac{\alpha}{2}\|\mathbf{X}\|_F^2$, we need to solve the following minimization problem for given $\mathbf{R} \in \Re^{p\times p}$ and $\sigma > 0$,

$$\mathbf{Z}^* = \arg\min_{\mathbf{Z}} \left\{ \sigma g(\mathbf{Z}) + \frac{1}{2}\|\mathbf{Z} - \sigma \mathbf{R}\|^2 \right\}.$$

$\mathbf{Z}^*$ here can be calculated by

$$Z_{ij}^* = \begin{cases} \dfrac{(\sigma R_{ij} - W_{ij}) + \sigma\sqrt{(R_{ij} - W_{ij}/\sigma)^2 + 4(\alpha+1)n_{ij}/(n\sigma)}}{2(\alpha+1)} & \text{if } (i,j) \in \Omega; \\ \sigma \max(R_{ij} - W_{ij}/\sigma, 0) \quad \text{if } (i,j) \in \overline{\Omega}. \end{cases}$$

4.2.3. A unified analysis for the majorized iPDC algorithm. Due to the presence of the proximal term $\frac{\alpha}{2}\|\mathbf{X} - \mathbf{X}^k\|^2$ in Algorithm 2, the classical DC analyses cannot be applied directly. In this subsection, we provide a unified convergence analysis for the majorized indefinite-proximal DC (majorized iPDC) algorithm which includes Algorithm 2 as a special instance. Let $\mathbb{X}$ be a finite-dimensional real Euclidean space endowed with inner product $\langle \cdot, \cdot \rangle$ and induced norm $\|\cdot\|$. Consider the following optimization problem

$$\min_{x \in \mathbb{X}} \theta(x) \triangleq g(x) + p(x) - q(x), \tag{14}$$

where $g : \mathbb{X} \to \Re$ is a continuously differentiable function (not necessarily convex) with a Lipschitz continuous gradient and Lipschitz modulus $L_g > 0$, i.e.,

$$\|\nabla f(x) - \nabla f(x')\| \leq L_g\|x - x'\| \quad \forall\, x, x' \in \mathbb{X},$$

$p : \mathbb{X} \to (-\infty, +\infty]$ and $q : \mathbb{X} \to (-\infty, +\infty]$ are two proper closed convex functions. It is not difficult to observe that penalized problem (12) is a special instance of problem (14). For general model (14), one can only expect the DC algorithm converges to a critical point $\bar{x} \in \mathbb{X}$ of (14) satisfying

$$(\nabla g(\bar{x}) + \partial p(\bar{x})) \cap \partial q(\bar{x}) \neq \emptyset.$$

Since g is continuously differentiable with Lipschitz continuous gradient, there exists a self-adjoint positive semidefinite linear operator $\mathcal{G} : \mathbb{X} \to \mathbb{X}$ such that for any $x, x' \in \mathbb{X}$,

$$g(x) \leq \widehat{g}(x; x') \triangleq g(x') + \langle \nabla g(x'), x - x' \rangle + \frac{1}{2}\|x - x'\|_{\mathcal{G}}^2.$$

Algorithm 3 A majorized indefinite-proximal DC algorithm for solving problem (14)

Given initial point $x^0 \in \mathbb{X}$ and stopping tolerance η, choose a self-adjoint, possibly indefinite, linear operator $\mathcal{T} : \mathbb{X} \to \mathbb{X}$. Iterate the following steps for $k = 0, 1, \dots$:

1. Choose $\xi^k \in \partial q(x^k)$. Compute

$$x^{k+1} \in \arg\min_{x \in \mathbb{X}} \left\{ \widehat{\theta}(x; x^k) + \frac{1}{2} \|x - x^k\|_{\mathcal{T}}^2 \right\}, \tag{15}$$

where $\widehat{\theta}(x; x^k) \triangleq \widehat{g}(x; x^k) + p(x) - \big(q(x^k) + \langle x - x^k, \xi^k \rangle\big)$.

2. If $\|x^{k+1} - x^k\| \leq \eta$, stop.

We present the majorized iPDC algorithm for solving (14) in Algorithm 3 and provide the following convergence results.

Theorem 6 **(Convergence of iPDC)**. *Assume that* $\inf_{x \in \mathbb{X}} \theta(x) > -\infty$. *Let* $\{x^k\}$ *be the sequence generated by Algorithm 3. If* $x^{k+1} = x^k$ *for some* $k \geq 0$, *then* x^k *is a critical point of* (14). *If* $\mathcal{G} + 2\mathcal{T} \succeq 0$, *then any accumulation point of* $\{x^k\}$, *if exists, is a critical point of* (14). *In addition, if* $\mathcal{G} + 2\mathcal{T} \succ 0$, *it holds that* $\lim_{k \to \infty} \|x^{k+1} - x^k\| = 0$.

The proof of Theorem 6 is provided in the supplementary material.

Remark 11. Here, we discuss the roles of linear operators $\mathcal{G}$ and $\mathcal{T}$. First, $\mathcal{G}$ makes the subproblems (15) in Algorithm 3 more amenable to efficient computations. Theorem 6 shows the algorithm is convergent if $\mathcal{G} + 2\mathcal{T} \succeq 0$. This indicates that instead of adding the commonly used positive semidefinte or positive definite proximal terms, we allow $\mathcal{T}$ to be indefinite for better practical performance. The computational benefit of using indefinite proximal terms has also been observed in (Gao and Sun 2010, Li et al. 2016a). As far as we know, Theorem 6 provides the first rigorous convergence proof of the DC algorithms with indefinite proximal terms. Second, $\mathcal{G}$ and $\mathcal{T}$ also help to guarantee that the solutions of the subproblems (15) exist. Since $\mathcal{G} + 2\mathcal{T} \succeq 0$ and $\mathcal{G} \succeq 0$, we have that $2\mathcal{G} + 2\mathcal{T} \succeq 0$, i.e., $\mathcal{G} + \mathcal{T} \succeq 0$. Hence, $\mathcal{G} + 2\mathcal{T} \succeq 0$ ($\mathcal{G} + 2\mathcal{T} \succ 0$) implies that subproblems (15) are (strongly) convex. Third, the choices of $\mathcal{G}$ and $\mathcal{T}$ are very much problem dependent. The general principle is that $\mathcal{G} + \mathcal{T}$ should be as small as possible while ensuring x^{k+1} is relatively easy to compute.

5. Simulation results

In this section, we conduct numerical experiments to validate our theoretical results. We first compare the proposed nuclear-norm regularized estimator and the rank-constrained estimator with previous methods in literature using synthetic data. We then use the rank-constrained method to analyze a dataset of Manhattan taxi trips to reveal citywide traffic patterns. All of our computational results are obtained by running MATLAB (version 9.5) on a windows workstation (8-core, Intel Xeon W-2145 at 3.70GHz, 64 G RAM).

5.1. Experiments with simulated data

We randomly draw the transition matrix $\mathbf{P}$ as follows. Let $\mathbf{U}_0, \mathbf{V}_0 \in \Re^{p\times r}$ be random matrices with i.i.d. standard normal entries and let

$$\widetilde{\mathbf{U}}_{[i,:]} = (\mathbf{U}_0 \circ \mathbf{U}_0)_{[i,:]}/\|(\mathbf{U}_0)_{[i,:]}\|_2^2 \text{ and } \widetilde{\mathbf{V}}_{[:,j]} = (\mathbf{V}_0 \circ \mathbf{V}_0)_{[:,j]}/\|(\mathbf{V}_0)_{[:,j]}\|_2^2, \quad i = 1,\ldots,p, j = 1,\ldots,r,$$

where $\circ$ is the Hadamard product and $\widetilde{\mathbf{U}}_{[i,:]}$ denotes the i-th row of $\widetilde{\mathbf{U}}$. The transition matrix $\mathbf{P}$ is obtained via $\mathbf{P} = \widetilde{\mathbf{U}}\widetilde{\mathbf{V}}^\top$. Then we simulate a Markov chain trajectory of length $n = \mathrm{round}(krp\log(p))$ on p states, $\{X_0,\ldots,X_n\}$, with varying values of k.

We compare the performance of four procedures: the nuclear norm penalized MLE, rank-constrained MLE, empirical estimator and spectral estimator. Here, the empirical estimator is the empirical count distribution matrix defined as follows:

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

The empirical estimator is in fact the unconstrained maximum likelihood estimator without taking into account the low-rank structure. The spectral estimator (Zhang and Wang 2017, Algorithm 1) is based on a truncated SVD. In the implementation of the nuclear norm penalized estimator, the regularization parameter λ in (3) is set to be $C\sqrt{p\log p/n}$ with constant C selected by cross-validation. For each method, let $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{V}}$ be the leading r left and right singular vectors of the resulting estimator $\widehat{\mathbf{P}}$. We measure the statistical performance of $\widehat{\mathbf{P}}$ through three quantities:

$$\eta_F := \|\mathbf{P} - \widehat{\mathbf{P}}\|_F^2, \quad \eta_{KL} := D_{\mathrm{KL}}(\mathbf{P}, \widehat{\mathbf{P}}), \text{ and } \eta_{UV} := \max\left\{\|\sin\Theta(\widehat{\mathbf{U}}, \mathbf{U})\|_F^2, \|\sin\Theta(\widehat{\mathbf{V}}, \mathbf{V})\|_F^2\right\}.$$

We consider the following setting with $p = 1000$, $r = 10$, and $k \in [10, 100]$. The results are plotted in Figure 1. One can observe from these results that for rank-constrained, nuclear norm penalized and spectral methods, η_F, η_{KL} and η_{UV} converge to zero quickly as the number of the state transitions n increases, while the statistical error of the empirical estimator decreases in a much slower rate. Among the three estimators in the zoomed plots (second rows of Figure 1), the rank constrained estimator slightly outperforms the nuclear norm penalized estimator and the spectral estimator. This observation is consistent with our algorithmic design: the nuclear norm minimization procedure is actually the initial step of Algorithm 2; thus the rank-constrained estimator can be seen as a refined version of the nuclear norm regularized estimator.

We also consider the case where the invariant distribution π is "imbalanced", i.e., we construct $\mathbf{P}$ such that $\min_{i=1,\dots,p} \pi_i$ is quite small and the appearance of some states is significantly less than the others. Specifically, given $\gamma_1, \gamma_2 > 0$, we generate a diagonal matrix $\mathbf{D}$ with i.i.d. beta-distributed ($\mathrm{Beta}(\gamma_1, \gamma_2)$) diagonal elements. After obtaining $\widetilde{\mathbf{U}}$ and $\widetilde{\mathbf{V}}$ in the same way as in the beginning of this subsection, we compute $\widetilde{\mathbf{P}} = \widetilde{\mathbf{U}}\widetilde{\mathbf{V}}^\top \mathbf{D}$. The ground truth transition matrix $\mathbf{P}$ is obtained after a normalization of $\widetilde{\mathbf{P}}$. Then, we simulate a Markov chain trajectory of length $n = \mathrm{round}(krp\log(p))$ on p states. In our experiment, we set $p = 1000$, $r = 10$, $k \in [10, 100]$, and $\gamma_1 = \gamma_2 = 0.5$. The detailed results are plotted in Figure 2. As can be seen from the figure, under the imbalanced setting, the rank-constrained, nuclear norm penalized and spectral methods perform much better than the empirical approach in terms of all the three statistical performance measures (η_F, η_{KL} and η_{UV}). In addition, the rank-constrained estimator exhibits a clear advantage over two other approaches.

5.2. Experiments with Manhattan Taxi data

In this experiment, we analyze a real dataset of 1.1×10^7 trip records of NYC Yellow cabs (Link: `https://s3.amazonaws.com/nyc-tlc/trip+data/yellow_tripdata_2016-01.csv`) in January 2016. Our goal is to partition the Manhattan island into several areas, in each of which the taxi customers share similar destination preference. This can provide guidance for balancing the supply and demand of taxi service and optimizing the allocation of traffic resources.

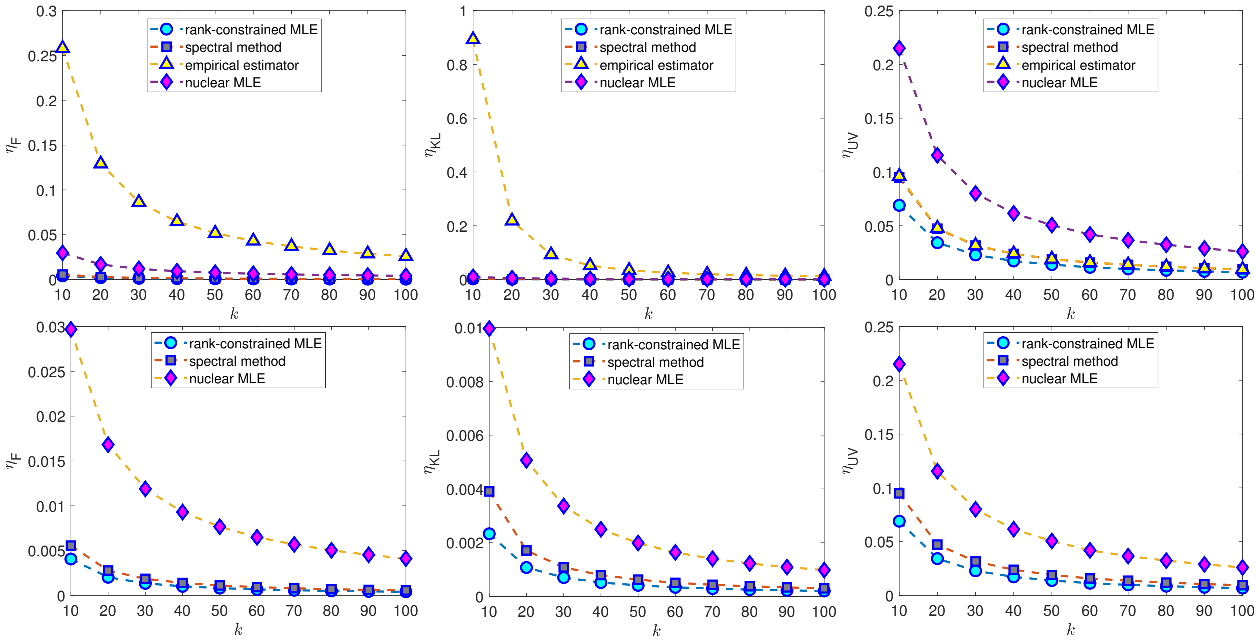

Figure 1 The first row compares the rank-constrained estimator, nuclear norm penalized estimator, spectral method, and empirical estimator in terms of $\eta_F = \|\mathbf{P} - \widehat{\mathbf{P}}\|_F^2, \eta_{KL} = D_{\mathrm{KL}}(\mathbf{P}, \widehat{\mathbf{P}})$, and $\eta_{UV} = \max\{\|\sin\Theta(\widehat{\mathbf{U}}, \mathbf{U},)\|_F^2, \|\sin\Theta(\widehat{\mathbf{V}}, \mathbf{V})\|_F^2\}$. The second row provides the zoomed plots of the first row without the empirical estimator. Here, $n = \mathrm{round}(krp\log p)$ with $p = 1,000$, $r = 10$ and k ranging from 10 to 100.

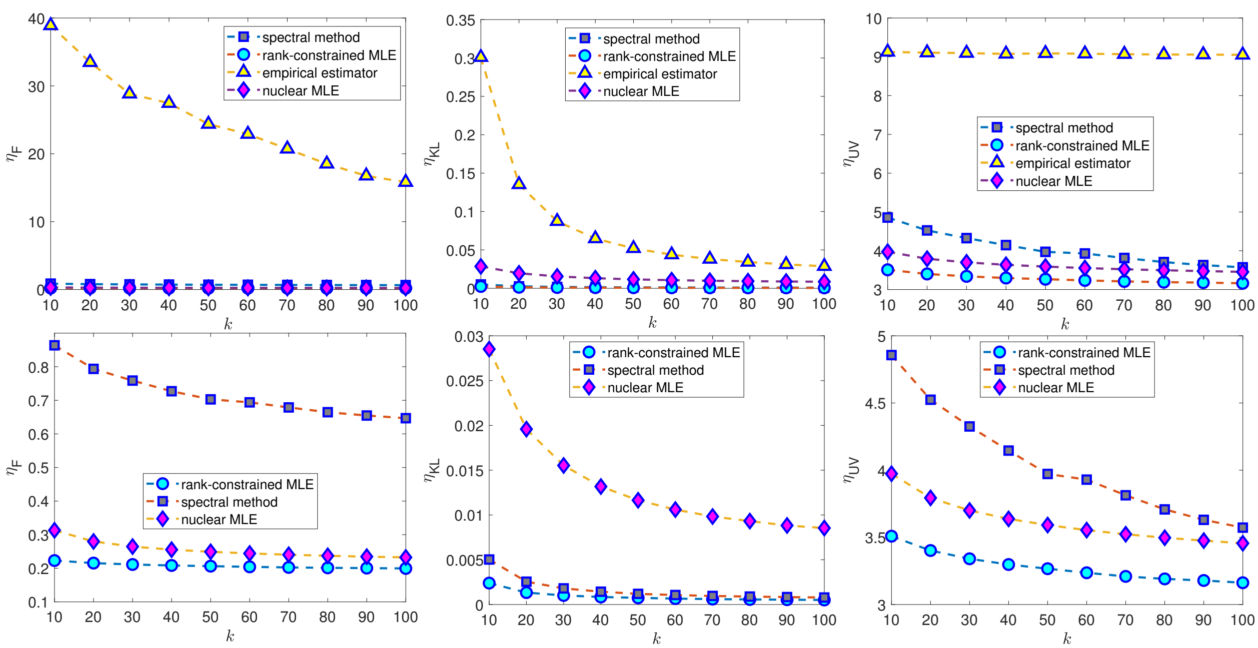

Figure 2 The first row compares the rank-constrained estimator, nuclear norm penalized estimator, spectral method, and empirical estimator in terms of $\eta_F = \|\mathbf{P} - \widehat{\mathbf{P}}\|_F^2, \eta_{KL} = D_{\mathrm{KL}}(\mathbf{P}, \widehat{\mathbf{P}})$, and $\eta_{UV} = \max\{\|\sin\Theta(\widehat{\mathbf{U}}, \mathbf{U},)\|_F^2, \|\sin\Theta(\widehat{\mathbf{V}}, \mathbf{V})\|_F^2\}$ with imbalanced invariant distribution. The second row provides the zoomed plots of the first row without the empirical estimator. Here, $n = \mathrm{round}(krp\log p)$ with $p = 1,000$, $r = 10$ and k ranging from 10 to 100.

We discretize the Manhattan island into a fine grid and model each cell of the grid as a state of the Markov chain; each taxi trip can thus be viewed as a state transition of this Markov chain (Yang et al. 2017, Benson et al. 2017, Liu et al. 2012). For stability concerns, our model ignores the cells that have fewer than $1,000$ taxi visits. Given that the traffic dynamics typically vary over time, we fit the MC under three periods of a day, i.e., $06:00 \sim 11:59$ (morning), $12:00 \sim 17:59$ (afternoon) and $18:00 \sim 23:59$ (evening), where the number of the active states $p = 803$, 999 and $1,079$ respectively. We apply the rank-constrained likelihood approach to obtain the estimator $\widehat{\mathbf{P}}^r$ of the transition matrix, and then apply k-means to the left singular subspaces of $\widehat{\mathbf{P}}^r$ to classify all the states into several clusters. Figure 3 presents the clustering result with $r = 4$ and $k = 4$ for the three periods of a day.

First of all, we notice that the locations within the same cluster are close with each other in geographical distance. This is non-trivial: we do not have exposure to GPS location in the clustering analysis. This implies that taxi customers in neighboring locations have similar destination preference, which is consistent with common sense. Furthermore, to track the variation of the traffic dynamics over time, Figure 4 visualizes the distribution of the destination choice that is correspondent to the center of the green cluster in the morning, afternoon and evening respectively. We identify the varying popular destinations in different periods of the day and provide corresponding explanations in the following table:

Time	Popular Destinations	Explanation
Morning	New York–Presbyterian Medical Center, 42–59 St. Park Ave, Penn Station	hospitals, workplaces, the train station
Afternoon	66 St. Broadway	lunch, afternoon break, short trips
Evening	Penn Station	go home

Finally, it might be tempting to model the taxi trips by an HMM, where regions of Manhattan correspond to hidden states. However, such a region is always part of the current observation (i.e.,

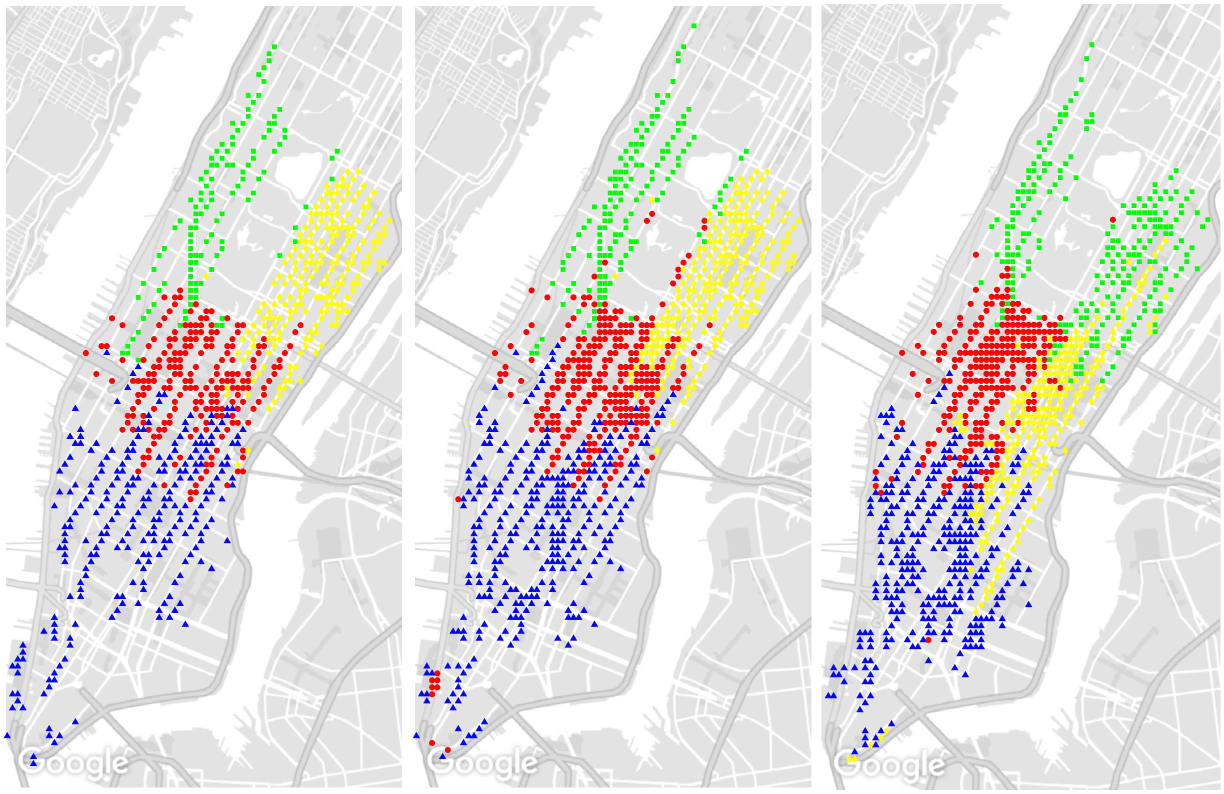

Figure 3 The meta-states compression of Manhattan traffic network via rank-constrained approach with $r = 4$: mornings (left), afternoons (middle) and evenings (right). Each color or symbol represents a meta-state. One can see the day-time state aggregation results differ significantly from that of the evening time.

location of taxi): It is observable and is not a hidden state that has to be inferred from all past observations. As a result, although both HMM and the low-rank Markov model could apply to taxi trips, the low-rank Markov model is simpler and more accurate.

6. Conclusion

This paper studies the recovery and state compression of low-rank Markov chains from empirical trajectories via a rank-constrained likelihood approach. We provide statistical upper bounds for the ℓ_2 risk and Kullback-Leiber divergence between the estimator and the true probability transition matrix for the proposed estimator. Then, a novel DC programming algorithm is developed to solve the associated rank-constrained optimization problem. The proposed algorithm non-trivially combines several recent optimization techniques, such as the penalty approach, the proximal DC algorithm, and the multi-block sGS-ADMM. We further study a new class of majorized indefinite-proximal DC algorithms for solving general non-convex non-smooth DC programming problems

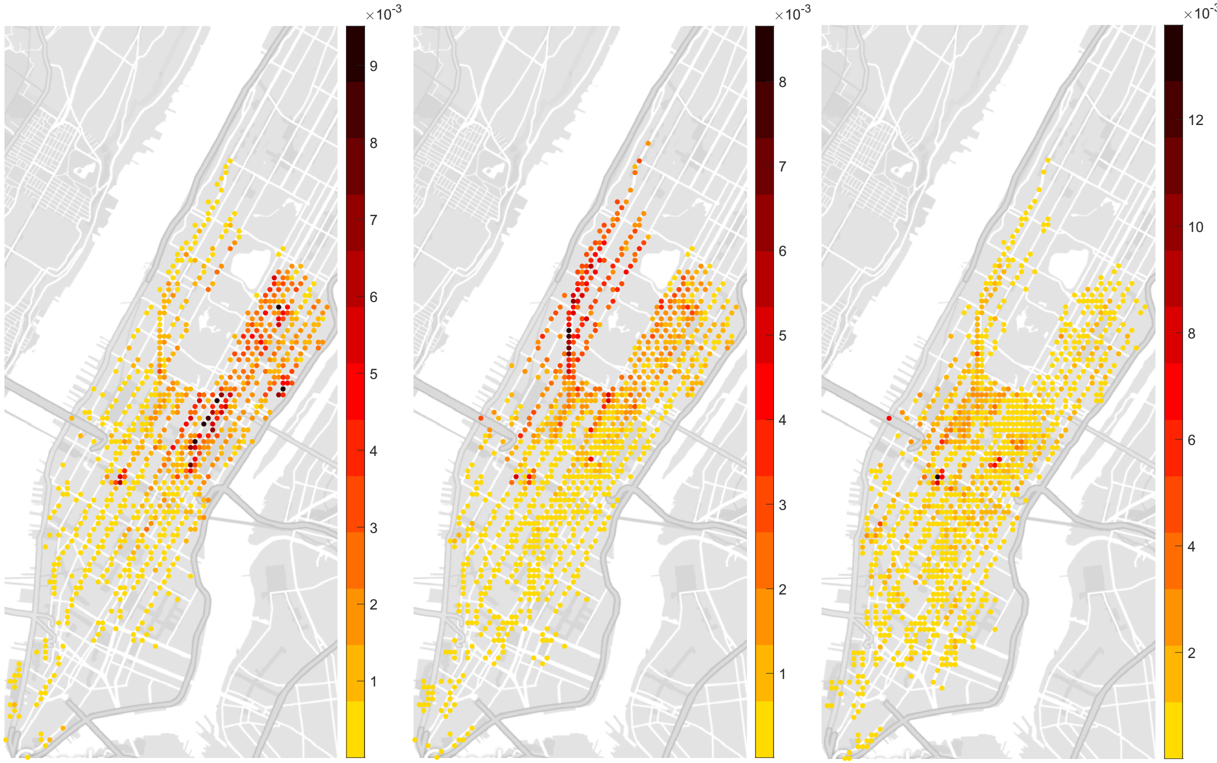

Figure 4 Visualization of the destination distributions corresponding to the pick-up locations in the green clusters in Figure 3: mornings (left), afternoons (middle) and evenings (right).

and provide a unified convergence analysis. Experiments on simulated data illustrate the merits of our approach.

References

Adamczak R (2008) A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electronic Journal of Probability* 13(34):1000–1034.

Anandkumar A, Ge R, Hsu D, Kakade SM, Telgarsky M (2014) Tensor decompositions for learning latent variable models. *The Journal of Machine Learning Research* 15(1):2773–2832.

Azizzadenesheli K, Lazaric A, Anandkumar A (2016) Reinforcement learning in rich-observation MDPs using spectral methods. *arXiv preprint arXiv:1611.03907* .

Benson AR, Gleich DF, Lim LH (2017) The spacey random walk: A stochastic process for higher-order data. *SIAM Review* 59(2):321–345.

Bertsekas DP (1995) *Dynamic Programming and Optimal Control*, volume 1 (Athena Scientific, Belmont, MA).

Bertsekas DP, Tsitsiklis JN (1995) Neuro-dynamic programming: an overview. *Proceedings of the 34th IEEE Conference on Decision and Control*, volume 1, 560–564 (IEEE).

Boucheron S, Lugosi G, Massart P (2013) *Concentration Inequalities: A Nonasymptotic Theory of Independence* (Oxford university press).

Brémaud P (1999) *Markov chains: Gibbs fields, Monte Carlo simulation, and queues*, volume 31 (Springer Science & Business Media).

Buchholz P (1994) Exact and ordinary lumpability in finite Markov chains. *Journal of Applied Probability* 31(1):59–75.

Cao Y, Xie Y (2016) Poisson matrix recovery and completion. *IEEE Transactions on Signal Processing* 64(6):1609–1620.

Chen C, He B, Ye Y, Yuan X (2016) The direct extension of admm for multi-block convex minimization problems is not necessarily convergent. *Mathematical Programming* 155(1-2):57–79.

Chen L, Sun D, Toh KC (2017) An efficient inexact symmetric Gauss–Seidel based majorized ADMM for high-dimensional convex composite conic programming. *Mathematical Programming* 161(1-2):237–270.

Chodera JD, Singhal N, Pande VS, Dill KA, Swope WC (2007) Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *The Journal of Chemical Physics* 126(15).

Coifman RR, Kevrekidis IG, Lafon S, Maggioni M, Nadler B (2008) Diffusion maps, reduction coordinates, and low dimensional representation of stochastic systems. *Multiscale Modeling & Simulation* 7(2):842–864.

Deng K, Huang D (2012) Model reduction of Markov chains via low-rank approximation. *American Control Conference (ACC), 2012*, 2651–2656 (IEEE).

Deng K, Mehta PG, Meyn SP (2011) Optimal Kullback-Leibler aggregation via spectral theory of Markov chains. *IEEE Transactions on Automatic Control* 56(12):2793–2808.

E W, Li T, Vanden-Eijnden E (2008) Optimal partition and effective dynamics of complex networks. *Proceedings of the National Academy of Sciences* 105(23):7907–7912.

Fan J, Liu H, Sun Q, Zhang T (2018) I-LAMM for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *The Annals of Statsitics* 46(2):814–841.

Ferreira JFB, Khoo Y, Singer A (2017) Semidefinite programming approach for the quadratic assignment problem with a sparse graph. *Computational Optimization and Applications* 1–36.

Fill JA (1991) Eigenvalue bounds on convergence to stationarity for nonreversible Markov chains, with an application to the exclusion process. *The Annals of Applied Probability* 1(1):62–87.

Gao Y, Sun D (2010) A majorized penalty approach for calibrating rank constrained correlation matrix problems. *technical reprot* .

Hall EC, Raskutti G, Willett R (2016) Inference of high-dimensional autoregressive generalized linear models. *arXiv preprint arXiv:1605.02693* .

Han Y, Jiao J, Weissman T (2015) Minimax estimation of discrete distributions under ℓ_1 loss. *IEEE Transactions on Information Theory* 61(11):6343–6354.

Hao Y, Orlitsky A, Pichapati V (2018) On learning Markov chains. *Advances in Neural Information Processing Systems*, 648–657.

Hsu D, Kakade SM, Zhang T (2012) A spectral algorithm for learning hidden Markov models. *Journal of Computer and System Sciences* 78(5):1460–1480.

Huang Q, Kakade SM, Kong W, Valiant G (2016) Recovering structured probability matrices. *arXiv preprint arXiv:1602.06586* .

Jiang B, Fan J, Sun Q (2018) Bernstein's inequality for general markov chains. *arXiv:1805.10721* .

Jiang K, Sun D, Toh KC (2014) A partial proximal point algorithm for nuclear norm regularized matrix least squares problems. *Mathematical Programming Computation* 6(3):281–325.

Jiang X, Raskutti G, Willett R (2015) Minimax optimal rates for poisson inverse problems with physical constraints. *IEEE Transactions on Information Theory* 61(8):4458–4474.

Keshavan RH, Montanari A, Oh S (2010) Matrix completion from a few entries. *IEEE Transactions on Information Theory* 56(6):2980–2998.

Koltchinskii V, Lounici K, Tsybakov AB (2011) Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics* 39(5):2302–2329.

Kontorovich L (2007) *Measure concentration of strongly mixing processes with applications*. Ph.D. thesis, Carnegie Mellon University, School of Computer Science, Machine Learning.

Kontorovich LA, Ramanan K (2008) Concentration inequalities for dependent random variables via the martingale method. *The Annals of Probability* 36(6):2126–2158.

Lam XY, Marron JS, Sun D, Toh KC (2018) Fast algorithms for large-scale generalized distance weighted discrimination. *Journal of Computational and Graphical Statistics* 27(2):368–379, URL `http://dx.doi.org/10.1080/10618600.2017.1366915`.

Le Thi HA, Le HM, Phan DN, Tran B (2017) Stochastic DCA for the large-sum of non-convex functions problem and its application to group variable selection in classification. *International Conference on Machine Learning*, 3394–3403.

Le Thi HA, Pham Dinh T (2018) DC programming and DCA: thirty years of developments. *Mathematical Programming* 1–64, URL `http://dx.doi.org/10.1007/s10107-018-1235-y`.

Le Thi HA, Pham Dinh T, Van Ngai H (2012) Exact penalty and error bounds in DC programming. *Journal of Global Optimization* 52(3):509–535.

Ledoux M, Talagrand M (2013) *Probability in Banach Spaces: Isoperimetry and Processes* (Springer Science & Business Media).

Lehmann EL, Casella G (2006) *Theory of Point Estimation* (Springer Science & Business Media).

Levin DA, Peres Y (2017) *Markov Chains and Mixing Times*, volume 107 (American Mathematical Soc.).

Li M, Sun D, Toh KC (2016a) A majorized ADMM with indefinite proximal terms for linearly constrained convex composite optimization. *SIAM Journal on Optimization* 26(2):922–950.

Li X, Sun D, Toh KC (2016b) A Schur complement based semi-proximal ADMM for convex quadratic conic programming and extensions. *Mathematical Programming* 155(1-2):333–373.

Liu Y, Kang C, Gao S, Xiao Y, Tian Y (2012) Understanding intra-urban trip patterns from taxi trajectory data. *Journal of Geographical Systems* 14(4):463–483.

Loh PL, Wainwright MJ (2015) Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *The Journal of Machine Learning Research* 16(1):559–616.

Marton K (1996) Bounding $\bar{d}$-distance by informational divergence: A method to prove measure concentration. *The Annals of Probability* 24(2):857–866.

Moore AW (1991) Variable resolution dynamic programming: Efficiently learning action maps in multivariate real-valued state-spaces. *Machine Learning Proceedings 1991*, 333–337 (Elsevier).

Negahban S, Oh S, Shah D (2016) Rank centrality: Ranking from pairwise comparisons. *Operations Research* 65(1):266–287.

Negahban S, Wainwright MJ (2011) Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics* 39(2):1069–1097.

Negahban S, Wainwright MJ (2012) Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research* 13(May):1665–1697.

Negahban SN, Ravikumar P, Wainwright MJ, Yu B (2012) A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science* 27(4):538–557.

Newman ME (2013) Spectral methods for community detection and graph partitioning. *Physical Review E* 88(4):042822.

Paulin D (2015) Concentration inequalities for markov chains by Marton couplings and spectral methods. *Electronic Journal of Probability* 20.

Pham Dinh T, Le Thi Ha (1997) Convex analysis approach to DC programming: Theory, algorithms and applications. *Acta Mathematica Vietnamica* 22(1):289–355.

Pham Dinh T, Le Thi Ha (2005) The DC (difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems. *Annals of Operations Research* 133(1-4):23–46.

Rockafellar RT (2015) *Convex Analysis* (Princeton University Press).

Rohrdanz MA, Zheng W, Maggioni M, Clementi C (2011) Determination of reaction coordinates via locally scaled diffusion map. *The Journal of Chemical Physics* 134(12):03B624.

Singh SP, Jaakkola T, Jordan MI (1995) Reinforcement learning with soft state aggregation. *Advances in Neural Information Processing Systems*, 361–368.

Steinhaus H (1957) The problem of estimation. *The Annals of Mathematical Statistics* 28(3):633–648.

Sutton RS, Barto AG (1998) *Reinforcement Learning: An Introduction*, volume 1 (MIT press Cambridge).

Tropp JA (2011) Freedman's inequality for matrix martingales. *Electronic Communications in Probability* 16:262–270.

Van Dinh B, Kim DS, Jiao L (2015) Convergence analysis of algorithms for DC programming. *arXiv preprint arXiv:1508.03899* .

Wang B, Zou H (2018) Another look at distance-weighted discrimination. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80(1):177–198.

Watson G (1993) On matrix approximation problems with Ky Fan$k$ norms. *Numerical Algorithms* 5(5):263–272.

Wen B, Chen X, Pong TK (2017) A proximal difference-of-convex algorithm with extrapolation. *Computational Optimization and Applications* 1–28.

Wolfer G, Kontorovich A (2019a) Estimating the mixing time of ergodic Markov chains. *COLT* .

Wolfer G, Kontorovich A (2019b) Minimax learning of ergodic Markov chains. *International Conference on Algorithmic Learning Theory* .

Yang LF, Braverman V, Zhao T, Wang M (2017) Dynamic partition of complex networks. *arXiv preprint arXiv:1705.07881* .

Yu B (1997) Assouad, fano, and le cam. *Festschrift for Lucien Le Cam*, 423–435 (Springer).

Zhang A, Wang M (2017) Spectral state compression of Markov processes. *arXiv preprint arXiv:1802.02920* .

Zhang A, Wang M (2019) Spectral state compression of markov processes. *IEEE Transactions on Information Theory* .

Technical lemmas and proofs

EC.1. Technical lemmas

LEMMA EC.1. *Given two discrete distributions $u, v \in \mathbb{R}^p$, if there exist $\alpha, \beta > 0$ such that $u_j \in \{0\} \cup [\alpha/p, \beta/p]$ and $v_j \in [\alpha/p, \beta/p]$ for any $j \in [p]$, then we have*

$$D_{\mathrm{KL}}(u, v) \geq \{p\alpha/(2\beta^2)\}\|u - v\|_2^2.$$

This implies that under Assumption 1, for any $\mathbf{Q} \in \mathcal{C}$,

$$\|\mathbf{P} - \mathbf{Q}\|_{\mathrm{F}}^2 \leq \frac{2\beta^2}{\alpha \pi_{\min} p} D_{\mathrm{KL}}(\mathbf{P}, \mathbf{Q}).$$

Proof of Lemma EC.1 By the mean value theorem, for any $j \in [p]$ such that $u_j \neq 0$, there exists $\xi_j \in [\alpha/p, \beta/p]$ such that

$$\log(v_j) - \log(u_j) = \frac{v_j - u_j}{u_j} - \frac{(v_j - u_j)^2}{2\xi_j^2}.$$

Therefore,

$$\begin{aligned} D_{\mathrm{KL}}(u, v) &= \sum_{j: u_j \neq 0} u_j \log(u_j / v_j) = \sum_{j: u_j \neq 0} (u_j - v_j) + \sum_{j: u_j \neq 0} \frac{(u_j - v_j)^2}{2\xi_j^2} \\ &\geq 1 - \sum_{j: u_j \neq 0} v_j + \sum_{j: u_j \neq 0} \frac{p\alpha(u_j - v_j)^2}{2\beta^2} = \sum_{j: u_j = 0} v_j - u_j + \sum_{j: u_j \neq 0} \frac{p\alpha(u_j - v_j)^2}{2\beta^2} \\ &\geq \sum_{j: u_j = 0} \frac{p(v_j - u_j)^2}{\beta} + \sum_{j: u_j \neq 0} \frac{p\alpha(u_j - v_j)^2}{2\beta^2} \geq \frac{p\alpha}{2\beta^2}\|u - v\|_2^2. \end{aligned}$$

Then we have

$$\|\mathbf{P} - \mathbf{Q}\|_{\mathrm{F}}^2 = \sum_{i \in [p]} \|P_{i\cdot} - Q_{i\cdot}\|_2^2 \leq \sum_{i \in [p]} \frac{2\beta^2 \pi_i}{p\alpha\pi_{\min}} D_{\mathrm{KL}}(P_{i\cdot}, Q_{i\cdot}) = \frac{2\beta^2}{p\alpha\pi_{\min}} D_{\mathrm{KL}}(\mathbf{P}, \mathbf{Q}).$$

EC.2. Proof of Theorem 1

Given the definition of $\widehat{\mathbf{P}}$,

$$\widetilde{D}_{\mathrm{KL}}(\mathbf{P}, \widehat{\mathbf{P}}) = \frac{1}{n}\sum_{i=1}^{n} \langle \log(\mathbf{P}) - \log(\widehat{\mathbf{P}}), \mathbf{X}_i \rangle = \ell_n(\widehat{\mathbf{P}}) - \ell_n(\mathbf{P}) \leq \lambda(\|\widehat{\mathbf{P}}\|_* - \|\mathbf{P}\|_*) \leq \lambda\|\mathbf{P} - \widehat{\mathbf{P}}\|_*. \quad \text{(EC.1)}$$

Then we have

$$
\begin{aligned}
D_{\mathrm{KL}}(\mathbf{P},\widehat{\mathbf{P}}) &= \mathcal{L}(\widehat{\mathbf{P}})-\mathcal{L}(\mathbf{P}) = \mathcal{L}(\widehat{\mathbf{P}})-\ell_n(\widehat{\mathbf{P}})+\ell_n(\widehat{\mathbf{P}})-\ell_n(\mathbf{P})+\ell_n(\mathbf{P})-\mathcal{L}(\mathbf{P}) \\
&\le \mathcal{L}(\widehat{\mathbf{P}})-\ell_n(\widehat{\mathbf{P}})+\ell_n(\mathbf{P})-\mathcal{L}(\mathbf{P})+\lambda\|\mathbf{P}-\widehat{\mathbf{P}}\|_* \qquad \text{(EC.2)} \\
&= D_{\mathrm{KL}}(\mathbf{P},\widehat{\mathbf{P}})-\widetilde{D}_{\mathrm{KL}}(\mathbf{P},\widehat{\mathbf{P}})+\lambda\|\mathbf{P}-\widehat{\mathbf{P}}\|_*.
\end{aligned}
$$

Define $\mathcal{E} := \{\lambda \ge 2\|\Pi_{\mathcal{N}}(\nabla \ell_n(\mathbf{P}))\|_2\}$. If $\mathcal{E}$ holds, then by Lemma 1 and then Lemma EC.1, we obtain that

$$
\begin{aligned}
D_{\mathrm{KL}}(\mathbf{P},\widehat{\mathbf{P}}) &\le D_{\mathrm{KL}}(\mathbf{P},\widehat{\mathbf{P}})-\widetilde{D}_{\mathrm{KL}}(\mathbf{P},\widehat{\mathbf{P}})+4(2r)^{1/2}\lambda\|\mathbf{P}-\widehat{\mathbf{P}}\|_{\mathrm{F}} \\
&\le D_{\mathrm{KL}}(\mathbf{P},\widehat{\mathbf{P}})-\widetilde{D}_{\mathrm{KL}}(\mathbf{P},\widehat{\mathbf{P}})+8\lambda\beta\left(\frac{rD_{\mathrm{KL}}(\mathbf{P},\widehat{\mathbf{P}})}{p\pi_{\min}\alpha}\right)^{1/2}.
\end{aligned}
$$

For any $\xi > 1$, an application of Lemma 3 with $\eta = \xi\pi_{\min}/(rp\pi_{\max}\log p)$ yields

$$
\mathbb{P}\left[\left\{D_{\mathrm{KL}}(\mathbf{P},\widehat{\mathbf{P}}) \le 16\lambda\beta\left(\frac{rD_{\mathrm{KL}}(\mathbf{P},\widehat{\mathbf{P}})}{p\pi_{\min}\alpha}\right)^{1/2}+\frac{2C_1 r\pi_{\max}\beta^2 p\log p}{\pi_{\min}\alpha^3 n}+\eta\right\}\cap\mathcal{E}\right] \ge 1-C_2e^{-\xi}-\mathbb{P}(\mathcal{E}^c),
$$

where C_1 and C_2 are exactly the same constants as in Lemma 3. Some algebra yields that

$$
\mathbb{P}\left[\left\{D_{\mathrm{KL}}(\mathbf{P},\widehat{\mathbf{P}}) \le \frac{256\lambda^2\beta^2 r}{p\pi_{\min}\alpha}+\frac{2C_1 r\pi_{\max}\beta^2 p\log p}{\pi_{\min}\alpha^3 n}+\eta\right\}\cap\mathcal{E}\right] \ge 1-C_2e^{-\xi}-\mathbb{P}(\mathcal{E}^c).
$$

By Lemma 2, there exists a universal constant $C_3 > 0$ such that if we choose

$$
\lambda = C_3\left\{\left(\frac{\xi p^2\pi_{\max}\log p}{n\alpha}\right)^{1/2}+\frac{\xi p\log p}{n\alpha}\right\},
$$

then for any $\xi > 1$, whenever $n\pi_{\max}(1-\rho_+) \ge \max(20,\xi^2)\log p$, we have that

$$
\mathbb{P}\left(D_{\mathrm{KL}}(\mathbf{P},\widehat{\mathbf{P}}^r) \gtrsim \frac{\xi r\pi_{\max}\beta^2 p\log p}{\pi_{\min}\alpha^3 n}+\frac{\xi\pi_{\min}}{rp\pi_{\max}\log p}\right) \lesssim e^{-\xi}+p^{-(\xi-1)}+p^{-10},
$$

as desired. The Frobenius-norm error bound follows immediately by applying Lemma EC.1.

EC.3. Proof of Theorem 2

Given the definition of $\widehat{\mathbf{P}}^r$,

$$
\widetilde{D}_{\mathrm{KL}}(\mathbf{P},\widehat{\mathbf{P}}^r) = \frac{1}{n}\sum_{i=1}^{n}\langle \log(\mathbf{P})-\log(\widehat{\mathbf{P}}^r), \mathbf{X}_i\rangle = \ell_n(\widehat{\mathbf{P}}^r)-\ell_n(\mathbf{P}) \le 0. \qquad \text{(EC.3)}
$$

Then we have

$$
\begin{aligned}
D_{\mathrm{KL}}(\mathbf{P},\widehat{\mathbf{P}}^r) &= \mathcal{L}(\widehat{\mathbf{P}}^r)-\mathcal{L}(\mathbf{P}) = \mathcal{L}(\widehat{\mathbf{P}}^r)-\ell_n(\widehat{\mathbf{P}}^r)+\ell_n(\widehat{\mathbf{P}}^r)-\ell_n(\mathbf{P})+\ell_n(\mathbf{P})-\mathcal{L}(\mathbf{P}) \\
&\le \mathcal{L}(\widehat{\mathbf{P}}^r)-\ell_n(\widehat{\mathbf{P}}^r)+\ell_n(\mathbf{P})-\mathcal{L}(\mathbf{P}) = D_{\mathrm{KL}}(\mathbf{P},\widehat{\mathbf{P}}^r)-\widetilde{D}_{\mathrm{KL}}(\mathbf{P},\widehat{\mathbf{P}}^r). \qquad \text{(EC.4)}
\end{aligned}
$$

For any $\xi > 1$, an application of Lemma 3 with $\eta = \pi_{\min}\xi/(rp\pi_{\max}\log p)$ yields

$$\mathbb{P}\left\{D_{\mathrm{KL}}(\mathbf{P},\widehat{\mathbf{P}}^r) \geq \max\left(\frac{2C_1 r\pi_{\max}\beta^2 p\log p}{\pi_{\min}\alpha^3 n}, \frac{\xi\alpha^2}{rp^2\pi_{\max}\log p}\right)\right\} \leq C_2 e^{-\xi},$$

as desired. The Frobenius-norm error bound immediately follows by Lemma EC.1.

EC.4. Proof of Theorem 3

To simplify the notation, assume without loss of generality that p is a multiple of $4(r-1)$. For any $1 \leq k \leq m$, consider

$$\begin{aligned}\mathbf{P}^{(k)} = &\begin{bmatrix} \frac{2-\alpha}{p}\mathbf{1}_{p\times(p/2)} & \frac{\alpha}{p}\mathbf{1}_{p\times(p/2)}\end{bmatrix}\\ &+ \frac{\eta(2-\alpha)}{2p}\begin{bmatrix} \mathbf{0}_{(p/2)\times(p/4)} & \mathbf{0}_{(p/2)\times(p/4)} & \mathbf{0}_{(p/2)\times(p/2)} \\ \mathbf{R}^{(k)} \;\cdots\; \mathbf{R}^{(k)} & -\mathbf{R}^{(k)} \;\cdots\; -\mathbf{R}^{(k)} & \mathbf{0}_{(p/4)\times(p/2)} \\ \underbrace{-\mathbf{R}^{(k)} \;\cdots\; -\mathbf{R}^{(k)}}_{l_0} & \underbrace{\mathbf{R}^{(k)} \;\cdots\; \mathbf{R}^{(k)}}_{l_0} & \mathbf{0}_{(p/4)\times(p/2)}\end{bmatrix},\end{aligned} \tag{EC.5}$$

where $l_0 = \frac{p}{4(r-1)}$, $\mathbf{R}^{(k)} \in \{0,1\}^{(p/4)\times(r-1)}$, and η is some positive value to be determined later. Let

$$\mu := \left(\frac{2-\alpha}{p}1_{p/2}^\top \quad \frac{\alpha}{p}1_{p/2}^\top\right)^\top. \tag{EC.6}$$

First of all, regardless of the value of $\mathbf{R}^{(k)}$, one can see that for any $k \in [m]$,

1. $\mathrm{rank}(\mathbf{P}^{(k)}) \leq r$;
2. $\mu^\top\mathbf{P}^{(k)} = \mu^\top$, and hence μ is the invariant distribution of $\mathbf{P}^{(k)}$;
3. $\mathbf{P}^{(k)} \in \Theta$.

Let $\{\mathbf{R}^{(k)}\}_{k=1}^m$ be i.i.d. matrices of independent Rademacher entries, i.e., for any $k \in [m]$, $\{R_{ij}^{(k)}\}_{i\in[n],j\in[d]}$ are independent Rademacher variables, and $\{\mathbf{R}^{(k)}\}_{k\in[m]}$ are independent. For any $k \neq l$, one can see that $\left\{\left|\mathbf{R}_{ij}^{(k)} - \mathbf{R}_{ij}^{(l)}\right|\right\}$ are i.i.d. uniformly distributed on $\{0,2\}$, and that

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

By Bernstein's inequality (Boucheron et al. 2013, Theorem 2.10), for any $t > 0$,

$$\mathbb{P}\left\{\left|\left\|\mathbf{R}^{(k)} - \mathbf{R}^{(l)}\right\|_1 - \frac{p(r-1)}{4}\right| \geq \left(\frac{p(r-1)t}{2}\right)^{1/2} + t\right\} \leq 2e^{-t}.$$

Let $t = p(r-1)/64$ and $m = \lfloor \exp\{p(r-1)/128\}/2^{1/2} \rfloor$. Since $p(r-1) \geq 192\log 2$, we have that $m \geq 2$. Then a union bound yields that

$$\mathbb{P}\left(\forall 1 \leq k < l \leq m,\ \frac{p(r-1)}{8} \leq \left\| \mathbf{R}^{(k)} - \mathbf{R}^{(l)} \right\|_1 \leq \frac{3p(r-1)}{8} \right) \geq 1 - 2m^2 \exp\left(\frac{-p(r-1)}{64} \right) > 0.$$

Hence, there exist $\mathbf{R}^{(1)}, \ldots, \mathbf{R}^{(m)} \subseteq \{-1, 1\}^{(p/4)\times(r-1)}$ such that

$$\forall 1 \leq k < l \leq m,\ \frac{p(r-1)}{8} \leq \left\| \mathbf{R}^{(k)} - \mathbf{R}^{(l)} \right\|_1 \leq \frac{3p(r-1)}{8}, \tag{EC.7}$$

which, given that $\|\mathbf{R}^{(j)} - \mathbf{R}^{(k)}\|_{\mathrm{F}}^2 = 2\|\mathbf{R}^{(j)} - \mathbf{R}^{(k)}\|_1$, further implies that

$$\forall 1 \leq k < l \leq m,\ \frac{p(r-1)}{4} \leq \|\mathbf{R}^{(k)} - \mathbf{R}^{(l)}\|_{\mathrm{F}}^2 \leq \frac{3p(r-1)}{4}. \tag{EC.8}$$

Now we have that

$$\begin{aligned} \left\| \mathbf{P}^{(k)} - \mathbf{P}^{(l)} \right\|_1 &= \frac{2l_0\eta(2-\alpha)}{p} \|\mathbf{R}^{(k)} - \mathbf{R}^{(l)}\|_1 \geq \frac{\eta p(2-\alpha)}{16}, \left\| \mathbf{P}^{(k)} - \mathbf{P}^{(l)} \right\|_1 \leq \frac{3\eta p(2-\alpha)}{16}, \\ \|\mathbf{P}^{(k)} - \mathbf{P}^{(l)}\|_{\mathrm{F}}^2 &= \frac{l_0\eta^2(2-\alpha)^2}{p^2} \|\mathbf{R}^{(k)} - \mathbf{R}^{(l)}\|_{\mathrm{F}}^2 \geq \frac{\eta^2(2-\alpha)^2}{16}, \left\| \mathbf{P}^{(k)} - \mathbf{P}^{(l)} \right\|_1 \leq \frac{3\eta^2(2-\alpha)^2}{16}. \end{aligned}$$

Besides,

$$\begin{aligned} D_{\mathrm{KL}}(\mathcal{X}^{(k)} || \mathcal{X}^{(l)}) &= n \sum_{i \in [p]} \pi_i D_{\mathrm{KL}}\big(\mathbf{P}_{[i,:]}^{(k)}, \mathbf{P}_{[i,:]}^{(l)}\big) = n \sum_{i=(p/2)+1}^{p} \sum_{j=1}^{p/2} \frac{\alpha}{p} P_{ij}^{(k)} \log\big(P_{ij}^{(k)}/P_{ij}^{(l)}\big) \\ &= \frac{2n\alpha}{p} \sum_{i=1}^{p/4} \frac{2-\alpha}{2} D_{\mathrm{KL}}\big(u_i^{(k)}, u_i^{(l)}\big), \end{aligned}$$

where $u_i^{(k)} = \frac{2}{p} 1_{p/2} + \frac{\eta}{p} \begin{bmatrix} \mathbf{R}_{[i,:]}^{(k)} & \cdots & \mathbf{R}_{[i,:]}^{(k)} & -\mathbf{R}_{[i,:]}^{(k)} & \cdots & -\mathbf{R}_{[i,:]}^{(k)} \end{bmatrix}$ corresponds to a $(p/2)$-dimensional distribution. By Zhang and Wang (2019, Lemma 4), we have that

$$D_{\mathrm{KL}}\big(u_i^{(k)}, u_i^{(l)}\big) \leq \frac{3l_0\eta^2}{p} \left\| \mathbf{R}_{[i,:]}^{(k)} - \mathbf{R}_{[i,:]}^{(l)} \right\|_2^2 = \frac{6l_0\eta^2}{p} \left\| \mathbf{R}_{[i,:]}^{(k)} - \mathbf{R}_{[i,:]}^{(l)} \right\|_1.$$

Therefore,

$$\begin{aligned} D_{\mathrm{KL}}(\mathcal{X}^{(k)}, \mathcal{X}^{(l)}) &= \frac{6n\alpha(2-\alpha)l_0\eta^2}{p^2} \sum_{i=1}^{p/4} \left\| \mathbf{R}_{[i,:]}^{(k)} - \mathbf{R}_{[i,:]}^{(l)} \right\|_1 \leq \frac{12n\alpha l_0\eta^2}{p^2} \|\mathbf{R}^{(k)} - \mathbf{R}^{(l)}\|_1 \\ &\leq \frac{12n\alpha l_0\eta^2}{p^2} \frac{3p(r-1)}{8} = \frac{9n\eta^2\alpha}{8}. \end{aligned}$$

By Fano's inequality (Yu 1997, Lemma 3), we have that

$$\inf_{\widehat{\mathbf{P}}} \sup_{\mathbf{P}\in\Theta} \|\widehat{\mathbf{P}}-\mathbf{P}\|_{\mathrm{F}}^2 \geq \inf_{\widehat{\mathbf{P}}} \sup_{\mathbf{P}\in\{\mathbf{P}^{(1)},\ldots,\mathbf{P}^{(m)}\}} \|\hat{\mathbf{P}}-\mathbf{P}\|_{\mathrm{F}}^2 \geq \frac{\eta^2(2-\alpha)^2}{16}\left(1-\frac{9n\eta^2\alpha-\log 2}{\log m}\right).$$

There exist universal constants $c_1, c_2 > 0$ such that when $p(r-1) \geq 192\log 2$, choosing $\eta = c_1\{p(r-1)/(n\alpha)\}^{1/2}$ yields that

$$\inf_{\widehat{\mathbf{P}}} \sup_{\mathbf{P}\in\Theta} \|\widehat{\mathbf{P}}-\mathbf{P}\|_{\mathrm{F}}^2 \geq c_2\frac{p(r-1)}{n\alpha}.$$

□

EC.5. Proof of Theorem 4

Let $\widehat{\mathbf{U}}_\perp, \widehat{\mathbf{V}}_\perp \in \Re^{p\times(p-r)}$ be the orthogonal complement of $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{V}}$. Since $\mathbf{U}, \mathbf{V}, \widehat{\mathbf{U}}$, and $\widehat{\mathbf{V}}$ are the leading left and right singular vectors of $\mathbf{P}$ and $\widehat{\mathbf{P}}$, we have

$$\|\widehat{\mathbf{P}}-\mathbf{P}\|_F \geq \|\widehat{\mathbf{U}}_\perp^\top(\widehat{\mathbf{P}}-\mathbf{U}\mathbf{U}^\top\mathbf{P})\|_F = \|\widehat{\mathbf{U}}_\perp^\top\mathbf{U}\mathbf{U}^\top\mathbf{P}\|_F \geq \|\widehat{\mathbf{U}}_\perp^\top\mathbf{U}\|_F\sigma_r(\mathbf{U}^\top\mathbf{P}) = \|\sin\Theta(\widehat{\mathbf{U}},\mathbf{U})\|_F\sigma_r(\mathbf{P}).$$

Similar argument also applies to $\|\sin\Theta(\widehat{\mathbf{V}},\mathbf{V})\|$. Thus,

$$\max\big(\|\sin\Theta(\widehat{\mathbf{U}},\mathbf{U})\|_F, \|\sin\Theta(\widehat{\mathbf{V}},\mathbf{V})\|_F\big) \leq \min\left(\frac{\|\widehat{\mathbf{P}}-\mathbf{P}\|_F}{\sigma_r(\mathbf{P})}, r^{1/2}\right).$$

The rest of the proof immediately follows from Theorem 1.

EC.6. Proof of Lemma 1

By the inequality (52) in Lemma 3 in the Appendix of Negahban and Wainwright (2012), we have for any $\mathbf{\Delta} \in \Re^{p\times p}$,

$$\|\mathbf{P}+\mathbf{\Delta}\|_* - \|\mathbf{P}\|_* \geq \|\mathbf{\Delta}_{\overline{\mathcal{M}}^\perp}\|_* - \|\mathbf{\Delta}_{\overline{\mathcal{M}}}\|_* - 2\|\mathbf{P}_{\mathcal{M}^\perp}\|_*.$$

Besides,

$$\begin{aligned}\ell_n(\mathbf{P}+\mathbf{\Delta}) - \ell_n(\mathbf{P}) &\geq \langle\nabla\ell_n(\mathbf{P}),\mathbf{\Delta}\rangle = \langle\Pi_{\mathcal{N}}(\nabla\ell_n(\mathbf{P})),\mathbf{\Delta}\rangle \geq -|\langle\Pi_{\mathcal{N}}(\nabla\ell_n(\mathbf{P})),\mathbf{\Delta}\rangle| \\ &\geq -\|\Pi_{\mathcal{N}}(\nabla\ell_n(\mathbf{P}))\|_2\|\mathbf{\Delta}\|_* \geq -\frac{\lambda}{2}\big(\|\mathbf{\Delta}_{\overline{\mathcal{M}}}\|_* + \|\mathbf{\Delta}_{\overline{\mathcal{M}}^\perp}\|_*\big).\end{aligned}$$

By the optimality of $\widehat{\mathbf{P}}$, $\ell_n(\widehat{\mathbf{P}}) + \lambda\|\widehat{\mathbf{P}}\|_* \leq \ell_n(\mathbf{P}) + \lambda\|\mathbf{P}\|_*$. Therefore,

$$\lambda\big(\|\mathbf{\Delta}_{\overline{\mathcal{M}}}\|_* + 2\|\mathbf{P}_{\mathcal{M}^\perp}\|_* - \|\mathbf{\Delta}_{\overline{\mathcal{M}}^\perp}\|_*\big) \geq \lambda(\|\mathbf{P}\|_* - \|\widehat{\mathbf{P}}\|_*) \geq -\frac{\lambda}{2}\big(\|\widehat{\mathbf{\Delta}}_{\overline{\mathcal{M}}}\|_* + \|\widehat{\mathbf{\Delta}}_{\overline{\mathcal{M}}^\perp}\|_*\big),$$

from which we deduce that

$$\|\widehat{\mathbf{\Delta}}_{\overline{\mathcal{M}}^\perp}\|_* \leq 3\|\widehat{\mathbf{\Delta}}_{\overline{\mathcal{M}}}\|_* + 4\|\mathbf{P}_{\mathcal{M}^\perp}\|_*.$$

EC.7. Proof of Lemma 2

Some algebra yields that

$$\nabla \ell_n(\mathbf{Q}) = \frac{1}{n}\sum_{i=1}^{n} -\frac{\mathbf{X}_i}{\langle \mathbf{Q}, \mathbf{X}_i\rangle}. \tag{EC.9}$$

For ease of notation, write $\mathbf{Z}_i := -\mathbf{X}_i/\langle \mathbf{P}, \mathbf{X}_i\rangle$. Note that $\mathbf{Z}_i$ is well-defined almost surely. Besides,

$$\mathbb{E}(\mathbf{Z}_i|\mathbf{Z}_{i-1}) = \mathbb{E}(\mathbf{Z}_i|\mathbf{X}_{i-1}) = \sum_{j=1}^{p} -\frac{e_{X_{i-1}} e_j^\top}{P_{X_{i-1},j}} P_{X_{i-1},j} = -e_{X_{i-1}} 1^\top.$$

Thus $\|\mathbf{Z}_i - \mathbb{E}(\mathbf{Z}_i|\mathbf{Z}_{i-1})\|_2 \le p/\alpha + \sqrt{p} =: R < \infty$. Define $\mathbf{S}_k := \sum_{i=1}^{k} \mathbf{Z}_i - \mathbb{E}(\mathbf{Z}_i|\mathbf{Z}_{i-1})$, then $\{\mathbf{S}_k\}_{k=1}^n$ is a matrix martingale. In addition,

$$\begin{aligned}
&\mathbb{E}\big\{(\mathbf{Z}_i - \mathbb{E}(\mathbf{Z}_i|\mathbf{Z}_{i-1}))^\top (\mathbf{Z}_i - \mathbb{E}(\mathbf{Z}_i|\mathbf{Z}_{i-1}))|\{\mathbf{S}_k\}_{k=1}^{i-1}\big\} = \mathbb{E}\big\{(\mathbf{Z}_i - \mathbb{E}(\mathbf{Z}_i|\mathbf{Z}_{i-1}))^\top (\mathbf{Z}_i - \mathbb{E}(\mathbf{Z}_i|\mathbf{Z}_{i-1}))|\mathbf{Z}_{i-1}\big\} \\
&\quad = \mathbb{E}(\mathbf{Z}_i^\top \mathbf{Z}_i|\mathbf{Z}_{i-1}) - \mathbb{E}(\mathbf{Z}_i|\mathbf{Z}_{i-1})^\top \mathbb{E}(\mathbf{Z}_i|\mathbf{Z}_{i-1}) = \left(\sum_{j=1}^{p} \frac{e_j e_j^\top}{P_{X_{i-1},j}}\right) - 11^\top =: \mathbf{W}_i^{(1)},
\end{aligned}$$

and similarly,

$$\mathbb{E}\big\{(\mathbf{Z}_i - \mathbb{E}(\mathbf{Z}_i|\mathbf{Z}_{i-1}))(\mathbf{Z}_i - \mathbb{E}(\mathbf{Z}_i|\mathbf{Z}_{i-1}))^\top|\{\mathbf{S}_k\}_{k=1}^{i-1}\big\} = \left(\sum_{j=1}^{p} \frac{e_{X_{i-1}} e_{X_{i-1}}^\top}{P_{X_{i-1},j}}\right) - p e_{X_{i-1}} e_{X_{i-1}}^\top =: \mathbf{W}_i^{(2)}.$$

Write $\|\sum_{i=1}^n \mathbf{W}_i^{(1)}\|_2$ as $W_n^{(1)}$, $\|\sum_{i=1}^n \mathbf{W}_i^{(2)}\|_2$ as $W_n^{(2)}$ and $\max(W_n^{(1)}, W_n^{(2)})$ as W_n. By the matrix Freedman inequality (Tropp 2011, Corollary 1.3), for any $t \ge 0$ and $\sigma^2 > 0$,

$$\mathbb{P}(\|\mathbf{S}_n\|_2 \ge t, W_n \le \sigma^2) \le 2p \exp\left(-\frac{t^2/2}{\sigma^2 + Rt/3}\right). \tag{EC.10}$$

Now we need to choose an appropriate σ^2 so that $W_n \le \sigma^2$ holds with high probability. Note that $W_n^{(1)} \le np(\alpha^{-1}+1)$ and $W_n^{(2)} \le (p^2\alpha^{-1} - p)\sup_{j\in[p]} \sum_{i=1}^n 1_{\{X_i = s_j\}}$. In the following we derive a bound for $\sup_{j\in[p]} \sum_{i=1}^n 1_{\{X_i = s_j\}}$. For any $j \in [p]$, by Jiang et al. (2018, Theorem 1.2), which is a variant of Bernstein's inequality for Markov chains, we have that

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n} (1_{\{X_i = s_j\}} - \pi_j) > \epsilon\right\} \le \exp\left(-\frac{n\epsilon^2}{2(A_1\beta/p + A_2\epsilon)}\right), \tag{EC.11}$$

where

$$A_1 = \frac{1+\max(\rho_+,0)}{1-\max(\rho_+,0)} \quad \text{and} \quad A_2 = \frac{1}{3}1_{\{\rho_+ \le 0\}} + \frac{5}{1-\rho_+}1_{\{\rho_+ > 0\}}.$$

Some algebra yields that for any $\xi > 0$,

$$\mathbb{P}\bigg\{\frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i=s_j\}} - \pi_j > \bigg(\frac{4A_1\xi}{np}\bigg)^{1/2} + \frac{4A_2\xi}{n}\bigg\} \le \exp(-\xi).$$

A union bound over $j \in [p]$ yields that

$$\mathbb{P}\bigg\{\sup_{j\in[p]} \frac{1}{n}\sum_{i=1}^{n}(1_{\{X_i=s_j\}} - \pi_j) > \bigg(\frac{4A_1\xi \log p}{np}\bigg)^{1/2} + \frac{4A_2\xi\log p}{n}\bigg\} \le p^{-(\xi-1)},$$

which implies that

$$\mathbb{P}\bigg\{\sup_{j\in[p]} \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i=s_j\}} > \pi_{\max} + \bigg(\frac{4A_1\xi \log p}{np}\bigg)^{1/2} + \frac{4A_2\xi\log p}{n}\bigg\} \le p^{-(\xi-1)}.$$

Therefore, whenever $n\pi_{\max}(1-\rho_+) \ge 2\log p$, we have that

$$\mathbb{P}\bigg(\sup_{j\in[p]} \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i=s_j\}} \gtrsim \pi_{\max}\bigg) \le \exp\bigg(-\frac{n\pi_{\max}(1-\rho_+)}{2}\bigg).$$

Combining this with the bounds of $W_n^{(1)}$ and $W_n^{(2)}$, we have that

$$\mathbb{P}\bigg(W_n \ge \frac{C_1 np^2\pi_{\max}}{\alpha}\bigg) \le \exp\bigg(-\frac{n\pi_{\max}(1-\rho_+)}{2}\bigg),$$

where C_1 is a universal constant. Now choosing $\sigma^2 = C_1 np^2\pi_{\max}/\alpha$, we deduce that for any $t \ge 0$,

$$\begin{aligned}\mathbb{P}(\|\mathbf{S}_n\|_2 \ge t) &= \mathbb{P}(\|\mathbf{S}_n\|_2 \ge t, W_n \le \sigma^2) + \mathbb{P}(\|\mathbf{S}_n\|_2 \ge t, W_n > \sigma^2)\\ &\le \mathbb{P}(\|\mathbf{S}_n\|_2 \ge t, W_n \le \sigma^2) + \mathbb{P}(W_n > \sigma^2)\\ &\le 2p\exp\bigg(-\frac{t^2/2}{\sigma^2 + Rt/3}\bigg) + \exp\bigg(-\frac{n\pi_{\max}(1-\rho_+)}{2}\bigg).\end{aligned}$$

Equivalently, for any $\xi > 1$,

$$\mathbb{P}\bigg\{\bigg\|\frac{1}{n}\mathbf{S}_n\bigg\|_2 \gtrsim \bigg(\frac{\xi p^2\pi_{\max}\log p}{n\alpha}\bigg)^{1/2} + \frac{\xi p\log p}{n\alpha}\bigg\} \le 4p^{-(\xi-1)} + \exp\bigg(-\frac{n\pi_{\max}(1-\rho_+)}{2}\bigg).$$

Finally, observe that for any $i \in [n]$, $\Pi_{\mathcal{N}}(\mathbb{E}(\mathbf{Z}_i|\mathbf{Z}_{i-1})) = \Pi_{\mathcal{N}}(-e_{X_{i-1}}1^\top) = \mathbf{0}$. Therefore, $\Pi_{\mathcal{N}}(\nabla\ell_n(\mathbf{P})) = n^{-1}\mathbf{S}_n$ and the final conclusion then follows.

EC.8. Proof of Lemma 3

We first split $\mathcal{C}(\eta)$ as the union of the sets

$$\mathcal{C}_l := \left\{ \mathbf{Q} \in \mathcal{C}(\eta) : 2^{l-1}\eta \le D_{\mathrm{KL}}(\mathbf{P},\mathbf{Q}) \le 2^l \eta,\ \operatorname{rank}(\mathbf{Q}) \le r \right\}, \quad l = 1,2,3,\ldots.$$

Define

$$\begin{aligned}\gamma_l &= \sup_{\mathbf{Q}\in\mathcal{C}_l} \left| D_{\mathrm{KL}}(\mathbf{P},\mathbf{Q}) - \widetilde{D}_{\mathrm{KL}}(\mathbf{P},\mathbf{Q}) \right| \\ &= \sup_{\mathbf{Q}\in\mathcal{C}_l} \left| \frac{1}{n}\sum_{i=1}^{n} \langle \log(\mathbf{P}) - \log(\mathbf{Q}), \mathbf{X}_i\rangle - \mathbb{E}\langle \log(\mathbf{P}) - \log(\mathbf{Q}), \mathbf{X}_i\rangle \right|.\end{aligned}$$

First, we wish to apply Adamczak (2008, Theorem 7) to bound $|\gamma_l - \mathbb{E}\gamma_l|$. Adamczak's bound entails the following asymptotic weak variance

$$\sigma^2 := \sup_{\mathbf{Q}\in\mathcal{C}_l} \operatorname{Var}\left\{ \sum_{i=S_1+1}^{S_2} \langle \log(\mathbf{P}) - \log(\mathbf{Q}), \mathbf{X}_i\rangle - \mathbb{E}\langle \log(\mathbf{P}) - \log(\mathbf{Q}), \mathbf{X}_i\rangle \right\} / \mathbb{E}T_2.$$

We have that

$$\begin{aligned}\sigma^2 &\le \sup_{\mathbf{Q}\in\mathcal{C}_l} \mathbb{E}\left[\left\{ \sum_{i=S_1+1}^{S_2} \langle \log(\mathbf{P}) - \log(\mathbf{Q}), \mathbf{X}_i\rangle - \mathbb{E}\langle \log(\mathbf{P}) - \log(\mathbf{Q}), \mathbf{X}_i\rangle \right\}^2\right] / \mathbb{E}T_2 \\ &= \frac{1}{2}\sup_{\mathbf{Q}\in\mathcal{C}_l} \sum_{j=1}^{\infty} \mathbb{E}\left[\left\{ \sum_{i=S_1+1}^{S_2} \langle \log(\mathbf{P}) - \log(\mathbf{Q}), \mathbf{X}_i\rangle - \mathbb{E}\langle \log(\mathbf{P}) - \log(\mathbf{Q}), \mathbf{X}_i\rangle \right\}^2 1_{\{T_2=j\}}\right] \\ &\le \frac{1}{2}\sum_{j=1}^{\infty} 4j^2 \log^2(\beta/\alpha)\mathbb{P}(T_2=j) = 2\log^2(\beta/\alpha)\mathbb{E}(T_2^2) = 8\log^2(\beta/\alpha).\end{aligned}$$

By Adamczak (2008, Theorem 7), there exists a universal constant $K > 1$ such that for any $\xi > 0$,

$$\mathbb{P}\left\{ |\gamma_l - \mathbb{E}\gamma_l| \ge K\mathbb{E}\gamma_l + 2\log(\beta/\alpha)\left(\frac{2K\xi}{n}\right)^{1/2} + \frac{16K\log(\beta/\alpha)\xi\log n}{n} \right\} \le Ke^{-\xi}.$$

Since $n^{-1/2} \ge 2n^{-1}\log n$ for any positive integer n, we have that

$$\mathbb{P}\left\{ |\gamma_l - \mathbb{E}\gamma_l| \ge K\mathbb{E}\gamma_l + 11K\log(\beta/\alpha)\left(\frac{\xi}{n}\right)^{1/2} \right\} \le Ke^{-\xi}. \tag{EC.12}$$

Next, we bound $\mathbb{E}\gamma_l$. Let $\{\varepsilon_i\}_{i=1}^n$ be n independent Rademacher random variables. By a symmetrization argument,

$$\begin{aligned}\mathbb{E}\gamma_l =&\mathbb{E}\left(\sup_{\mathbf{Q}\in\mathcal{C}_l} \left| \frac{1}{n}\sum_{i=1}^{n} \langle \log(\mathbf{P}) - \log(\mathbf{Q}), \mathbf{X}_i\rangle - \mathbb{E}\langle \log(\mathbf{P}) - \log(\mathbf{Q}), \mathbf{X}_i\rangle \right| \right) \\ \le& 2\mathbb{E}\left(\sup_{\mathbf{Q}\in\mathcal{C}_l} \left| \frac{1}{n}\sum_{i=1}^{n} \varepsilon_k \langle \log(\mathbf{P}) - \log(\mathbf{Q}), \mathbf{X}_i\rangle \right| \right).\end{aligned}$$

Let $\phi_i(t)=(\alpha/p)\{\log(\langle \mathbf{P},\mathbf{X}_i\rangle+t)-\log(\langle \mathbf{P},\mathbf{X}_i\rangle)\}$. Then $\phi_i(0)=0$ and $|\phi_i'(t)|\le 1$ for all t such that $t+\langle \mathbf{P},\mathbf{X}_i\rangle \ge \alpha/p$. In other words, ϕ_i is a contraction map for $t\ge \min_{j,k\in[p]}(P_{jk}-\alpha/p)$. By the contraction principle (Theorem 4.12 in Ledoux and Talagrand (2013)),

$$\begin{aligned}\mathbb{E}\gamma_l \le& \frac{2p}{\alpha}\mathbb{E}\left(\sup_{\mathbf{Q}\in\mathcal{C}_l}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\phi_i\left(\langle \mathbf{Q}-\mathbf{P},\mathbf{X}_i\rangle\right)\right|\right)\le \frac{4p}{\alpha}\mathbb{E}\left(\sup_{\mathbf{Q}\in\mathcal{C}_l}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\langle \mathbf{Q}-\mathbf{P},\mathbf{X}_i\rangle\right|\right)\\ \le& \frac{4p}{\alpha}\mathbb{E}\left(\sup_{\mathbf{Q}\in\mathcal{C}_l}\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\mathbf{X}_i\right\|_2\|\mathbf{Q}-\mathbf{P}\|_*\right)\le \frac{4p}{\alpha}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\mathbf{X}_i\right\|_2\sup_{\mathbf{Q}\in\mathcal{C}_l}\|\mathbf{Q}-\mathbf{P}\|_*.\end{aligned} \tag{EC.13}$$

By Lemma EC.1,

$$\sup_{\mathbf{Q}\in\mathcal{C}_l}\|\mathbf{Q}-\mathbf{P}\|_*\le \sup_{\mathbf{Q}\in\mathcal{C}_l}(2r)^{1/2}\|\mathbf{Q}-\mathbf{P}\|_{\mathrm{F}}\le 2\beta\left(\frac{2^l\eta r}{p\alpha\pi_{\min}}\right)^{1/2}. \tag{EC.14}$$

Hence, the remaining task is to bound $\mathbb{E}\|n^{-1}\sum_{i=1}^n\varepsilon_i\mathbf{X}_i\|$. From now on, we denote $\varepsilon_i\mathbf{X}_i$ by $\mathbf{Z}_i$. One can see that $(\mathbf{Z}_i)_{i=1}^n$ is a martingale difference sequence. We wish to apply the matrix Freedman inequality (Tropp 2011, Corollary 1.3) to bound the average of $(\mathbf{Z}_i)_{i=1}^n$. We have that

$$\begin{aligned}\left\|\sum_{i=1}^n\mathbb{E}\big(\mathbf{Z}_i^\top\mathbf{Z}_i|X_{i-1}\big)\right\|_2 =& \left\|\sum_{i=1}^n\sum_{j=1}^p P_{X_{i-1},j}(e_{X_{i-1}}e_j^\top)^\top(e_{X_{i-1}}e_j^\top)\right\|_2=\left\|\sum_{j=1}^p\sum_{i=1}^n P_{X_{i-1},j}e_je_j^\top\right\|_2\\ =&\max_{j\in[p]}\sum_{i=1}^n P_{X_{i-1},j}=:W_n^{(1)}\end{aligned}$$

and that

$$\begin{aligned}\left\|\sum_{i=1}^n\mathbb{E}\big(\mathbf{Z}_i\mathbf{Z}_i^\top|X_{i-1}\big)\right\|_2 =& \left\|\sum_{i=1}^n\sum_{j=1}^p P_{X_{i-1},j}e_{X_{i-1}}e_{X_{i-1}}^\top\right\|_2=\left\|\sum_{i=1}^n e_{X_{i-1}}e_{X_{i-1}}^\top\right\|_2\\ =&\max_{j\in[p]}\sum_{i=1}^n 1_{\{X_{i-1}=j\}}=:W_n^{(2)}.\end{aligned}$$

We first bound $W_n^{(1)}$. Note that for any $j\in[p]$, $\mathbb{E}(P_{X_{i-1},j})=\pi_j$, and that

$$\mathrm{Var}_\pi(P_{X_{i-1},j})=\sum_{k=1}^p\pi_k(P_{kj}-\pi_j)^2=\sum_{k=1}^p\pi_kP_{kj}^2-\pi_j^2\le\pi_j(1-\pi_j).$$

By a variant of Bernstein's inequality for Markov chains (Jiang et al. 2018, Theorem 1.2), we have that for any $j\in[p]$,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n P_{X_{i-1},j}-\pi_j>\epsilon\right)\le\exp\left\{-\frac{n\epsilon^2}{2(A_1\pi_j+A_2\epsilon)}\right\},$$

where

$$A_1 := \frac{1+\max(\rho_+,0)}{1-\max(\rho_+,0)} \quad \text{and} \quad A_2 := \frac{1}{3}1_{\{\rho_+\le 0\}} + \frac{5}{1-\rho_+}1_{\{\rho_+>0\}}.$$

A union bound yields that

$$\mathbb{P}\big\{W_n^{(1)} \ge n\pi_{\max} + (4nA_1\pi_{\max}\xi\log p)^{1/2} + 4A_2\xi\log p\big\} \le p^{-(\xi-1)}. \tag{EC.15}$$

Next we bound $W_n^{(2)}$. Note that $W_n^{(2)} \le \max_{j\in[p]}\sum_{i=1}^n 1_{\{X_{i-1}=s_j\}}$. Similarly, by Jiang et al. (2018, Theorem 1.2), for any $j\in[p]$,

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^n 1_{\{X_{i-1}=s_j\}} - \pi_j > \epsilon\right\} \le \exp\left\{-\frac{n\epsilon^2}{2(A_1\pi_j + A_2\epsilon)}\right\}, \tag{EC.16}$$

Some algebra yields that for any $\xi>0$,

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^n 1_{\{X_{i-1}=s_j\}} - \pi_j > \left(\frac{4A_1\pi_j\xi}{n}\right)^{1/2} + \frac{4A_2\xi}{n}\right\} \le \exp(-\xi).$$

By a union bound over $j\in[p]$,

$$\mathbb{P}\left\{\max_{j\in[p]}\frac{1}{n}\sum_{i=1}^n 1_{\{X_{i-1}=s_j\}} > \pi_{\max} + \left(\frac{4A_1\pi_{\max}\xi\log p}{n}\right)^{1/2} + \frac{4A_2\xi\log p}{n}\right\} \le p^{-(\xi-1)},$$

which further implies that

$$\mathbb{P}\big\{W_n^{(2)} \ge n\pi_{\max} + (4nA_1\pi_{\max}\xi\log p)^{1/2} + 4A_2\xi\log p\big\} \le p^{-(\xi-1)}. \tag{EC.17}$$

Define $W_n := \max(W_n^{(1)}, W_n^{(2)})$. Let $\mathbf{S}_n := \sum_{i=1}^n \varepsilon_i\mathbf{X}_i$. By matrix Freedman's inequality (Tropp 2011, Corollary 1.3), for any $t\ge 0$ and $\sigma^2>0$,

$$\mathbb{P}(\|\mathbf{S}_n\|_2 \ge t, W_n\le\sigma^2) \le 2p\exp\left(-\frac{t^2/2}{\sigma^2+t/3}\right). \tag{EC.18}$$

Now we need to choose an appropriate σ^2 so that $W_n\le\sigma^2$ holds with high probability. Given that $\rho_+>0$ and $n\pi_{\max}\ge 10\xi\log p/(1-\rho_+)$, combining (EC.15) and (EC.17) yields that

$$\mathbb{P}\big(W_n \ge 4n\pi_{\max}\big) \le 2p^{-(\xi-1)}. \tag{EC.19}$$

Now choosing $\sigma^2 = 4n\pi_{\max}$ in (EC.18), we deduce that

$$\begin{aligned}\mathbb{P}(\|\mathbf{S}_n\|_2 \geq t) &= \mathbb{P}(\|\mathbf{S}_n\|_2 \geq t, W_n \leq \sigma^2) + \mathbb{P}(\|\mathbf{S}_n\|_2 \geq t, W_n > \sigma^2)\\ &\leq \mathbb{P}(\|\mathbf{S}_n\|_2 \geq t, W_n \leq \sigma^2) + \mathbb{P}(W_n > \sigma^2)\\ &\leq 2p\exp\left(-\frac{t^2/2}{\sigma^2 + t/3}\right) + 2p^{-(\xi-1)}.\end{aligned}$$

Choose $\xi = n\pi_{\max}(1-\rho_+)/(10\log p)$. As long as $n\pi_{\max}(1-\rho_+) \geq \max(20\log p, \log n)$, we have that

$$\mathbb{E}\left\|\frac{1}{n}\mathbf{S}_n\right\|_2 \lesssim \left(\frac{\pi_{\max}\log p}{n}\right)^{1/2}. \tag{EC.20}$$

Combining (EC.13), (EC.14) and (EC.20) yields that

$$\mathbb{E}\gamma_l \lesssim \frac{\beta}{\alpha^{3/2}}\left(\frac{2^l\eta\pi_{\max}rp\log p}{\pi_{\min}n}\right)^{1/2}.$$

Then combining this with (EC.12) yields that

$$\mathbb{P}\left\{\gamma_l \gtrsim \frac{\beta}{\alpha^{3/2}}\left(\frac{2^l\eta\pi_{\max}rp\log p}{\pi_{\min}n}\right)^{1/2} + \log(\beta/\alpha)\left(\frac{\xi}{n}\right)^{1/2}\right\} \lesssim e^{-\xi}.$$

Let $\xi = 2^l\eta\pi_{\max}rp\log p/\pi_{\min}$. Then there exist universal constants $C_1, C_2 > 0$ such that

$$\mathbb{P}\left\{\gamma_l \geq \frac{C_1\beta}{\alpha^{3/2}}\left(\frac{2^l\eta\pi_{\max}rp\log p}{\pi_{\min}n}\right)^{1/2}\right\} \leq C_2\exp\left\{-\frac{(2l+1)\eta\pi_{\max}rp\log p}{\pi_{\min}}\right\}.$$

We can thus deduce that there exists a universal constant $C_3 > 0$ such that

$$\begin{aligned}&\mathbb{P}\left(|\widetilde{D}_{\mathrm{KL}}(\mathbf{P},\mathbf{Q}) - D_{\mathrm{KL}}(\mathbf{P},\mathbf{Q})| > \frac{1}{2}D_{\mathrm{KL}}(\mathbf{P},\mathbf{Q}) + \frac{C_3\pi_{\max}\beta^2 rp\log p}{\pi_{\min}\alpha^3 n}\right)\\ \leq &\sum_{l=0}^{\infty} P\left(\exists \mathbf{Q}\in\mathcal{C}_l,\ \left|\widetilde{D}_{\mathrm{KL}}(\mathbf{P},\mathbf{Q}) - D_{\mathrm{KL}}(\mathbf{P},\mathbf{Q})\right| > 2^{l-2}\eta + \frac{C_3\pi_{\max}\beta^2 rp\log p}{\pi_{\min}\alpha^3 n}\right)\\ \leq &\sum_{l=0}^{\infty} P\left\{\gamma_l \geq \frac{C_1\beta}{\alpha^{3/2}}\left(\frac{2^l\eta\pi_{\max}rp\log p}{\pi_{\min}n}\right)^{1/2}\right\}\\ \leq &C_2\sum_{l=0}^{\infty}\exp\left\{-\frac{(2l+1)\eta\pi_{\max}rp\log p}{\pi_{\min}}\right\} \leq 2C_2\exp\left(-\frac{\eta\pi_{\max}rp\log p}{\pi_{\min}}\right).\end{aligned}$$

where we use the Cauchy-Schwarz inequality in the second step.

EC.9. Alternative statistical error analysis

EC.9.1. Main results

In this section, we provide an alternative proof strategy that follows Negahban et al. (2012) to bound the statistical error of $\widehat{\mathbf{P}}$ and $\widehat{\mathbf{P}}^r$. This strategy resolves the inconsistency issue of Theorems 1 and 2 when $n \gg \{rp\pi_{\max}(\log p)\beta/(\pi_{\min}\alpha^{3/2})\}^2$. For any $R > 0$, define a constraint set $\mathcal{C}(\beta, R, \kappa) := \{\mathbf{\Delta} \in \Re^{p\times p} : \|\mathbf{\Delta}\|_{\max} \le \beta/p, \|\mathbf{\Delta}\|_{\mathrm{F}} \le R, \|\mathbf{\Delta}\|_* \le \kappa r^{1/2}\|\mathbf{\Delta}\|_{\mathrm{F}}\}$. An important ingredient of this statistical analysis is the localized restricted strong convexity (Negahban and Wainwright 2011, Fan et al. 2018) of the loss function $\ell_n(\mathbf{P})$ near $\mathbf{P}$. This property allows us to bound the distance in the parameter space by the difference in the objective function value. Define the first-order Taylor remainder term of the negative log-likelihood function $\ell_n(\mathbf{Q})$ around $\mathbf{P}$ as

$$\delta\ell_n(\mathbf{Q};\mathbf{P}) := \ell_n(\mathbf{Q}) - \ell_n(\mathbf{P}) - \nabla\ell_n(\mathbf{P})^\top(\mathbf{Q}-\mathbf{P}).$$

The following lemma establishes the desired local restricted strong convexity.

Lemma EC.2. *Under Assumption 1, there exists a universal constant K such that for any $\xi > 1$, it holds with probability at least $1 - K\exp(-\xi)$ that for any $\mathbf{\Delta} \in \mathcal{C}(\beta, R, \kappa)$,*

$$\delta\ell_n(\mathbf{P}+\mathbf{\Delta};\mathbf{P}) \ge \frac{\alpha^2}{8\beta^2}\|\mathbf{\Delta}\|_{\mathrm{F}}^2 - 8R\left(\frac{3K\xi}{n}\right)^{1/2} - \frac{8K\xi\alpha^2\log n}{\beta^2 n} - \frac{Kp\kappa R}{\beta}\left(\frac{r\pi_{\max}\log p}{n}\right)^{1/2}. \tag{EC.21}$$

Now we present the statistical rates of $\widehat{\mathbf{P}}$ and $\widehat{\mathbf{P}}^r$.

Theorem EC.1 **(Alternative statistical guarantee for $\widehat{\mathbf{P}}$)**. *Under the same assumptions of Theorem 1, there exists a universal constant $C_1 > 0$, such that for any $\xi > 1$, if we choose*

$$\lambda = C_1\left\{\left(\frac{\xi p^2\pi_{\max}\log p}{n\alpha}\right)^{1/2} + \frac{\xi p\log p}{n\alpha}\right\},$$

then whenever $n\pi_{\max}(1-\rho_+) \ge \max\{\max(20,\xi^2)\log p, \log n\}$, we have with probability at least $1 - K\exp(-\xi) - 4p^{-(\xi-1)} - p^{-1}$ that

$$\|\widehat{\mathbf{P}} - \mathbf{P}\|_{\mathrm{F}} \lesssim \frac{\beta^2}{\alpha^2}\left(\frac{\xi rp^2\pi_{\max}\log p}{n\alpha}\right)^{1/2} \quad \textit{and} \quad D_{\mathrm{KL}}(\mathbf{P},\widehat{\mathbf{P}}) \lesssim \frac{\xi\beta^6\pi_{\max}rp^2\log p}{n\alpha^7},$$

where K is the same constant as in Lemma EC.2.

Theorem EC.2 **(Alternative statistical guarantee for $\widehat{\mathbf{P}}^r$).** *Under the same assumptions of Theorem 1, there exists a universal constant $C_1 > 0$, for any $\xi > 1$, we have with probability at least $1 - K\exp(-\xi) - 4p^{-(\xi-1)} - p^{-1}$ that*

$$\|\widehat{\mathbf{P}}^r - \mathbf{P}\|_{\mathrm{F}} \lesssim \frac{\beta^2}{\alpha^2}\left(\frac{\xi r p^2 \pi_{\max}\log p}{n\alpha}\right)^{1/2} \quad \textit{and} \quad D_{\mathrm{KL}}(\mathbf{P}, \widehat{\mathbf{P}}^r) \lesssim \frac{\xi\beta^6\pi_{\max} r p^2 \log p}{n\alpha^7},$$

where K is the same constant as in Lemma EC.2.

One can see from the theorems above that the derived error bounds converge to zero as n goes to infinity. Nevertheless, their dependence on α and β is worse than those in Theorems 1 and 2 when $n \lesssim \{rp\pi_{\max}(\log p)\beta/(\pi_{\min}\alpha^{3/2})\}^2$. This is why we do not present this result in the main text.

EC.9.2. Proof of Lemma EC.2

Given any $\mathbf{\Delta} \in \mathcal{C}(\beta, R, \kappa)$, it holds that for some $0 \le v \le 1$ that

$$\begin{aligned}\delta\ell_n(\mathbf{P}+\mathbf{\Delta};\mathbf{P}) &= \frac{1}{2}\mathrm{vec}(\mathbf{\Delta})^\top \mathbf{H}_n(\mathbf{P}+v\mathbf{\Delta})\mathrm{vec}(\mathbf{\Delta}) = \frac{1}{2n}\sum_{i=1}^n \frac{\langle \mathbf{X}_i, \mathbf{\Delta}\rangle^2}{\langle \mathbf{P}+v\mathbf{\Delta}, \mathbf{X}_i\rangle^2} \\ &\ge \frac{1}{2n}\sum_{i=1}^n \frac{p^2}{4\beta^2}\langle \mathbf{\Delta}, \mathbf{X}_i\rangle^2.\end{aligned} \tag{EC.22}$$

Define

$$\Gamma_n := \sup_{\mathbf{\Delta}\in\mathcal{C}(\beta,R,\kappa)}\left|\frac{1}{n}\sum_{i=1}^n \langle \mathbf{\Delta}, \mathbf{X}_i\rangle^2 - \mathbb{E}(\langle \mathbf{\Delta}, \mathbf{X}_i\rangle^2)\right|.$$

We first bound the deviation of Γ_r from its expectation $\mathbb{E}\Gamma_r$. Note that $\{\mathbf{X}_i\}_{i=1}^n$ is a Markov chain on $\mathcal{M} := \{e_j e_k^\top\}_{j,k=1}^p$. Here we apply a tail inequality for suprema of unbounded empirical processes due to Adamczak (2008, Theorem 7). To apply this result, we need to verify that $\{\mathbf{X}_i\}_{i=1}^n$ satisfies the "minorization condition" as stated in Section 3.1 of Adamczak (2008). Below we characterize a specialized version of this condition.

Condition 1 (minorized condition). We say that a Markov chain $\mathcal{X}$ on $\mathcal{S}$ satisfies the minorized condition if there exist $\delta > 0$, a set $\mathcal{C} \subset \mathcal{S}$ and a probability measure ν on $\mathcal{S}$ for which $\forall_{x\in\mathcal{C}}\forall_{\mathcal{A}\subset\mathcal{S}}\mathbb{P}(x,\mathcal{A}) \ge \delta\nu(\mathcal{A})$ and $\forall_{x\in\mathcal{S}}\exists_{n\in\mathbb{N}}\mathbb{P}^n(x,\mathcal{C}) > 0$.

One can verify that the Markov chain $\{\mathbf{X}_i\}_{i=1}^n$ satisfies Condition 1 with $\delta = 1/2$, $\mathcal{C} = \{e_1 e_2^\top\}$ and $\nu(e_j e_k^\top) = P_{jk} 1_{\{j=2\}}$ for $j, k \in [p]$.

Now consider a new Markov chain $\{(\widetilde{\mathbf{X}}_i, R_i)\}_{i=1}^n$ constructed as follows. Let $\{R_i\}_{i=1}^n$ be i.i.d. Bernoulli random variables with $\mathbb{E}R_1 = \delta$. For any $i \in \{0, \ldots, n-1\}$, at step i, if $\mathbf{X}_i \notin \mathcal{C}$, we sample $\widetilde{\mathbf{X}}_{i+1}$ according to $\mathbb{P}(\widetilde{\mathbf{X}}_i, \cdot)$; otherwise, the distribution of $\widetilde{\mathbf{X}}_i$ depends on R_i: if $R_i = 1$, the chain regenerates in the sense that we draw $\widetilde{\mathbf{X}}_i$ from ν, and if $R_i = 0$, we draw $\widetilde{\mathbf{X}}_i$ from $(\mathbb{P}(\mathbf{X}_i, \cdot) - \delta\nu(\cdot))/(1-\delta)$. One can verify that the sequence $\{\widetilde{\mathbf{X}}_i\}_{i=1}^n$ has exactly the same distribution as the original Markov chain $\{\mathbf{X}_i\}_{i=1}^n$. Define $T_1 := \inf\{n > 0 : R_n = 1\}$ and $T_{i+1} := \inf\{n > 0 : R_{T_1 + \ldots + T_i + n} = 1\}$ for $i \geq 0$. Note that $\{T_i\}_{i \geq 0}$ are i.i.d. Geometric random variables with $\mathbb{E}T_1 = 2$ and $\|T_1\|_{\psi_1} \leq 4$. Let $S_0 := -1$, $S_j := T_1 + \ldots + T_j$ and $\mathcal{Y}_j := \{\widetilde{\mathbf{X}}_i\}_{i=S_{j-1}+1}^{S_j}$ for $j \geq 1$. Based on our construction, we deduce that $\{\mathcal{Y}_j\}_{j \geq 1}$ are independent. Thus we chop the original Markov chain $\{X_i\}_{i \in [n]}$ into independent sequences. Finally, Adamazak's bound entails the following asymptotic weak variance

$$\sigma^2 := \sup_{\mathbf{\Delta} \in \mathcal{C}(\beta, R, \kappa)} \operatorname{Var}\left\{ \sum_{i=S_1+1}^{S_2} \langle \mathbf{\Delta}, \mathbf{X}_i \rangle^2 - \mathbb{E}(\langle \mathbf{\Delta}, \mathbf{X}_i \rangle^2) \right\} / \mathbb{E}T_2.$$

We have

$$\begin{aligned}
\sigma^2 &\leq \sup_{\mathbf{\Delta} \in \mathcal{C}(\beta, R, \kappa)} \mathbb{E}\left[\left\{ \sum_{i=S_1+1}^{S_2} \langle \mathbf{\Delta}, \mathbf{X}_i \rangle^2 - \mathbb{E}(\langle \mathbf{\Delta}, \mathbf{X}_i \rangle^2) \right\}^2 \right] / \mathbb{E}T_2 \\
&= \frac{1}{2} \sup_{\mathbf{\Delta} \in \mathcal{C}(\beta, R, \kappa)} \sum_{j=1}^{\infty} \mathbb{E}\left[\left\{ \sum_{i=S_1+1}^{S_2} \langle \mathbf{\Delta}, \mathbf{X}_i \rangle^2 - \mathbb{E}(\langle \mathbf{\Delta}, \mathbf{X}_i \rangle^2) \right\}^2 1_{\{T_2 = j\}} \right] \\
&\leq \frac{1}{2} \sum_{j=1}^{\infty} \frac{j^2 R^2 \beta^4}{p^4} \mathbb{P}(T_2 = j) = \frac{R^2 \beta^4 \mathbb{E}(T_2^2)}{2p^4} = \frac{3\beta^4 R^2}{p^4}.
\end{aligned}$$

By Adamczak (2008, Theorem 7), there exists a universal constant K such that for any $\xi > 0$,

$$\mathbb{P}\left\{ |\Gamma_n - \mathbb{E}\Gamma_n| \geq K\mathbb{E}\Gamma_n + \frac{R\beta^2}{p^2} \left(\frac{3K\xi}{n} \right)^{1/2} + \frac{64K\xi\alpha^2 \log n}{np^2} \right\} \leq K \exp(-\xi). \tag{EC.23}$$

Next, by the symmetrization argument and Ledoux-Talagrand contraction inequality (Ledoux and Talagrand 2013), for n independent and identically distributed Rademacher variables $\{\gamma_i\}_{i=1}^n$, when $n\pi_{\max}(1 - \rho_+) \geq \max(20 \log p, \log n)$, we have that

$$\begin{aligned}
\mathbb{E}\Gamma_n &\leq 2\mathbb{E} \sup_{\substack{\|\mathbf{\Delta}\|_{\mathrm{F}} \leq R, \\ \mathbf{\Delta} \in \mathcal{C}(\beta, R, \kappa)}} \left| \frac{1}{n} \sum_{i=1}^n \gamma_i \langle \mathbf{\Delta}, \mathbf{X}_i \rangle^2 \right| \leq \frac{8\beta}{p} \, \mathbb{E} \sup_{\substack{\|\mathbf{\Delta}\|_{\mathrm{F}} \leq R, \\ \mathbf{\Delta} \in \mathcal{C}(\beta, R, \kappa)}} \left| \langle \frac{1}{n} \sum_{i=1}^n \gamma_i \mathbf{X}_i, \mathbf{\Delta} \rangle \right| \\
&\leq \frac{8\beta \|\mathbf{\Delta}\|_*}{p} \, \mathbb{E}\left\| \frac{1}{n} \sum_{i=1}^n \gamma_i \mathbf{X}_i \right\|_2 \leq \frac{8\kappa\beta r^{1/2} R}{p} \mathbb{E}\left\| \frac{1}{n} \sum_{i=1}^n \gamma_i \mathbf{X}_i \right\|_2 \leq \frac{8\kappa\beta R}{p} \left(\frac{r\pi_{\max} \log p}{n} \right)^{1/2},
\end{aligned} \tag{EC.24}$$

where the penultimate inequality is due to the fact that $\mathbf{\Delta} \in \mathcal{C}(\beta, R, \kappa)$, and where the last inequality is due to (EC.20).

Finally,

$$\mathbb{E}\langle \mathbf{\Delta}, \mathbf{X}_i \rangle^2 = \sum_{1 \le j,k \le d} \pi_j P_{jk} \Delta_{jk}^2 \ge \frac{\alpha^2}{p^2} \|\mathbf{\Delta}\|_{\mathrm{F}}^2. \tag{EC.25}$$

Combining all the bounds above, we have for any $\xi > 1$, with probability at least $1 - K\exp(-\xi)$,

$$\delta \ell_n(\mathbf{P} + \mathbf{\Delta}; \mathbf{P}) \ge \frac{\alpha^2}{8\beta^2} \|\mathbf{\Delta}\|_{\mathrm{F}}^2 - 8R\left(\frac{3K\xi}{n}\right)^{1/2} - \frac{8K\xi\alpha^2 \log n}{\beta^2 n} - \frac{Kp\kappa R}{\beta}\left(\frac{r\pi_{\max}\log p}{n}\right)^{1/2}. \tag{EC.26}$$

EC.9.3. Proof of Theorem EC.1

For a specific R whose value will be determined later, we construct an intermediate estimator $\widehat{\mathbf{P}}_\eta$ between $\widehat{\mathbf{P}}$ and $\mathbf{P}$:

$$\widehat{\mathbf{P}}_\eta = \mathbf{P} + \eta(\widehat{\mathbf{P}} - \mathbf{P}),$$

where $\eta = 1$ if $\|\widehat{\mathbf{P}} - \mathbf{P}\|_{\mathrm{F}} \le R$ and $\eta = R/\|\widehat{\mathbf{P}} - \mathbf{P}\|_{\mathrm{F}}$ if $\|\widehat{\mathbf{P}} - \mathbf{P}\|_{\mathrm{F}} > R$. For any $\xi > 1$, there exists a universal constant $C > 0$ such that when

$$\lambda = C\left\{\left(\frac{\xi p^2 \pi_{\max} \log p}{n\alpha}\right)^{1/2} + \frac{\xi p \log p}{n\alpha}\right\},$$

we have by Lemmas EC.2 and 2 that with probability at least $1 - K\exp(-\xi) - 4p^{-(\xi-1)} - p^{-1}$,

$$\begin{aligned}
&\frac{\alpha^2}{8\beta^2}\|\mathbf{\Delta}\|_{\mathrm{F}}^2 - 8R\left(\frac{3K\xi}{n}\right)^{1/2} - \frac{8K\xi\alpha^2\log n}{\beta^2 n} - \frac{Kp\kappa R}{\beta}\left(\frac{r\pi_{\max}\log p}{n}\right)^{1/2} \\
&\quad \le \delta\ell_n(\widehat{\mathbf{P}}_\eta; \mathbf{P}) \le -\langle \Pi_{\mathcal{N}}(\nabla \mathcal{L}_n(\mathbf{P})), \widehat{\mathbf{\Delta}}_\eta \rangle + \lambda(\|\mathbf{P}\|_* - \|\widehat{\mathbf{P}}_\eta\|_*) \\
&\quad \le -\langle \Pi_{\mathcal{N}}(\nabla \mathcal{L}_n(\mathbf{P})), \widehat{\mathbf{\Delta}}_\eta \rangle + \lambda\|\widehat{\mathbf{\Delta}}_\eta\|_* \le (\|\Pi_{\mathcal{N}}(\nabla \mathcal{L}_n(\mathbf{P}))\|_2 + \lambda)\|\widehat{\mathbf{\Delta}}_\eta\|_* \\
&\quad \le 8\lambda\|[\widehat{\mathbf{\Delta}}_\eta]_{\overline{\mathcal{M}}}\|_* \le 8\lambda\sqrt{r}\|\widehat{\mathbf{\Delta}}_\eta\|_{\mathrm{F}},
\end{aligned} \tag{EC.27}$$

where K is the same universal constant as in Theorem EC.2. Some algebra yields that

$$\|\widehat{\mathbf{\Delta}}_\eta\|_{\mathrm{F}}^2 \lesssim \frac{\beta^2}{\alpha^2}\max\left\{\frac{\lambda^2 r\beta^2}{\alpha^2}, R\left(\frac{\xi}{n}\right)^{1/2}, \frac{\xi\alpha^2\log n}{\beta^2 n}, \frac{pR}{\beta}\left(\frac{r\pi_{\max}\log p}{n}\right)^{1/2}\right\}. \tag{EC.28}$$

Letting R^2 be greater than the RHS of the inequality above, we can find a universal constant $C_4 > 0$ such that

$$R \ge \frac{C_4\beta^2}{\alpha^2}\left(\frac{\xi r p^2 \pi_{\max}\log p}{n\alpha}\right)^{1/2} =: R_0.$$

Choose $R = R_0$. Therefore, $\|\widehat{\mathbf{\Delta}}_\eta\|_{\rm F} \le R$ and $\widehat{\mathbf{\Delta}}_\eta = \widehat{\mathbf{\Delta}}$. We can thus reach the conclusion. As to the KL-Divergence, by Zhang and Wang (2017, Lemma 4), we deduce that

$$D_{\rm KL}(\widehat{\mathbf{P}},\mathbf{P}) = \sum_{j=1}^{p} \pi_j D_{\rm KL}(\mathbf{P}_{j\cdot},\widehat{\mathbf{P}}_{j\cdot}) \le \sum_{j=1}^{p} \frac{\beta^2}{2\alpha^2}\|\mathbf{P}_{j\cdot}-\widehat{\mathbf{P}}_{j\cdot}\|_2^2 = \frac{\beta^2}{2\alpha^2}\|\widehat{\mathbf{P}}-\mathbf{P}\|_{\rm F}^2, \tag{EC.29}$$

from which we attain the conclusion.

EC.9.4. Proof of Theorem EC.2

Define $\widehat{\mathbf{\Delta}}(r) := \widehat{\mathbf{P}}^r - \mathbf{P}$. Since $\text{rank}(\mathbf{P}) \le r$ and $\text{rank}(\widehat{\mathbf{P}}^r) \le r$, $\text{rank}(\widehat{\mathbf{\Delta}}(r)) \le 2r$. Thus $\|\widehat{\mathbf{\Delta}}(r)\|_{\rm F} \le (2r)^{1/2}\|\widehat{\mathbf{\Delta}}(r)\|_*$. Now we follow the proof strategy of Theorem 1 to establish the statistical error bound for $\widehat{\mathbf{P}}^r$. Similarly, for a specific $R > 0$ whose value will be determined later, we can construct an intermediate estimator $\widehat{\mathbf{P}}^r_\eta$ between $\widehat{\mathbf{P}}^r$ and $\mathbf{P}$:

$$\widehat{\mathbf{P}}^r_\eta = \mathbf{P} + \eta(\widehat{\mathbf{P}}^r - \mathbf{P}),$$

where $\eta = 1$ if $\|\widehat{\mathbf{P}}^r - \mathbf{P}\|_{\rm F} \le R$ and $\eta = R/\|\widehat{\mathbf{P}}^r - \mathbf{P}\|_{\rm F}$ if $\|\widehat{\mathbf{P}}^r - \mathbf{P}\|_{\rm F} > R$. Let $\widehat{\mathbf{\Delta}}_\eta(r) := \widehat{\mathbf{P}}^r_\eta - \mathbf{P}$. Since $\widehat{\mathbf{\Delta}}_\eta(r) \in \mathcal{C}(\beta, R, \sqrt{2})$, applying Lemma EC.2 yields that

$$\begin{aligned}
&\frac{\alpha^2}{8\beta^2}\|\mathbf{\Delta}\|_{\rm F}^2 - 8R\left(\frac{3K\xi}{n}\right)^{1/2} - \frac{8K\xi\alpha^2\log n}{\beta^2 n} - \frac{Kp\kappa R}{\beta}\left(\frac{r\pi_{\max}\log p}{n}\right)^{1/2} \\
&\le \delta\ell_n(\widehat{\mathbf{P}}^r_\eta;\mathbf{P}) \le -\langle \Pi_{\mathcal{N}}(\nabla\ell_n(\mathbf{P})), \widehat{\mathbf{\Delta}}_\eta(r)\rangle \le \|\Pi_{\mathcal{N}}(\nabla\mathcal{L}_n(\mathbf{P}))\|_2\|\widehat{\mathbf{\Delta}}_\eta(r)\|_* \\
&\le \sqrt{2r}\|\Pi_{\mathcal{N}}(\nabla\mathcal{L}_n(\mathbf{P}))\|_2\|\widehat{\mathbf{\Delta}}_\eta(r)\|_{\rm F},
\end{aligned} \tag{EC.30}$$

which futher implies that there exists C_1 depending only on α and β such that

$$\|\widehat{\mathbf{\Delta}}_\eta(r)\|_{\rm F}^2 \le C_1 \max\left\{ r\|\Pi_{\mathcal{N}}(\nabla\mathcal{L}_n(\mathbf{P}))\|_2^2, R\left(\frac{\xi}{n}\right)^{1/2}, \frac{\xi\alpha^2\log n}{\beta^2 n}, \frac{pR}{\beta}\left(\frac{r\pi_{\max}\log p}{n}\right)^{1/2}\right\}.$$

By a contradiction argument as in the proof of Theorem 1, we can choose an appropriate R large enough such that $\widehat{\mathbf{P}}^r_\eta = \widehat{\mathbf{P}}^r$ and attain the conclusion.

EC.10. Proof of Proposition 1

Since $\text{rank}(\mathbf{X}_c^*) \le r$, we know that $\mathbf{X}_c^*$ is in fact a feasible solution to the original problem (5) and $\|\mathbf{X}_c^*\|_* - \|\mathbf{X}_c^*\|_{(r)} = 0$. Therefore, for any feasible solution X to (5), it holds that

$$\begin{aligned}
f(\mathbf{X}_c^*) &= f(\mathbf{X}_c^*) + c(\|\mathbf{X}_c^*\|_* - \|\mathbf{X}_c^*\|_{(r)}) \\
&\le f(\mathbf{X}) + c(\|\mathbf{X}\|_* - \|\mathbf{X}\|_{(r)}) = f(\mathbf{X}).
\end{aligned}$$

This completes the proof of the proposition.

EC.11. Convergence and $o(1/k)$ non-ergodic iteration complexity of Algorithm 1 (sGS-ADMM)

Before deriving the desired results of Algorithm 1 for solving problem (10), we present some notation and definitions for the subsequent analysis. Assume that the solution sets of (9) and (10) are nonempty. Then, the primal-dual solution pairs associated with problems (9) and (10) satisfy the following Karush-Kuhn-Tucker (KKT) system:

$$0 \in R(\mathbf{X}, \boldsymbol{\Xi}, \mathbf{S}), \quad \mathcal{A}(\mathbf{X}) = b, \quad \boldsymbol{\Xi} + \mathcal{A}^*(y) + \mathbf{S} = 0, \tag{EC.31}$$

with

$$R(\mathbf{X}, \boldsymbol{\Xi}, \mathbf{S}) := \begin{pmatrix} \boldsymbol{\Xi} + \partial g(\mathbf{X}) \\ \\ \mathbf{X} + \partial \delta(\|\mathbf{S}\|_2 \le c) \end{pmatrix}, \quad (\mathbf{X}, \boldsymbol{\Xi}, \mathbf{S}) \in \operatorname{dom} g \times \Re^{p\times p} \times \left\{\mathbf{S} \in \Re^{p\times p} \mid \|\mathbf{S}\|_2 \le c\right\}.$$

Define the KKT residual function $D : \operatorname{dom} g \times \Re^{p\times p} \times \Re^n \times \{\mathbf{S} \in \Re^{p\times p} \mid \|\mathbf{S}\|_2 \le c\} \to [0, +\infty)$ as

$$D(\mathbf{X}, \boldsymbol{\Xi}, y, \mathbf{S}) := \operatorname{dist}^2(0, R(\mathbf{X}, \boldsymbol{\Xi}, \mathbf{S})) + \|\mathcal{A}(\mathbf{X}) - b\|^2 + \|\boldsymbol{\Xi} + \mathcal{A}^*(y) + \mathbf{S}\|^2.$$

We say $(\mathbf{X}, \boldsymbol{\Xi}, y, \mathbf{S}) \in \operatorname{dom} g \times \Re^{p\times p} \times \Re^n \times \{\mathbf{S} \in \Re^{p\times p} \mid \|\mathbf{S}\|_2 \le c\}$ be an ϵ-approximate primal-dual solution pair for problems (9) and (10) if $D(\mathbf{X}, \boldsymbol{\Xi}, y, \mathbf{S}) \le \epsilon$. We show in the following theorem the global convergence and the $o(1/k)$ iteration complexity results of Algorithm sGS-ADMM.

Theorem EC.3. *Suppose that the solution sets of* (9) *and* (10) *are nonempty. Let* $\{(\boldsymbol{\Xi}^k, y^k, \mathbf{S}^k, \mathbf{X}^k)\}$ *be the sequence generated by Algorithm 1. If* $\tau \in (0, (1+\sqrt{5})/2)$, *then the sequence* $\{(\boldsymbol{\Xi}^k, y^k, \mathbf{S}^k)\}$ *converges to an optimal solution of* (10) *and* $\{\mathbf{X}^k\}$ *converges to an optimal solution of* (9). *Moreover, there exist a constant* $\omega > 0$ *such that*

$$\min_{1\le i\le k} \left\{D(\mathbf{X}^k, \boldsymbol{\Xi}^k, y^k, \mathbf{S}^k)\right\} \le \frac{\omega}{k}, \ \forall\, k \ge 1, \quad \text{and} \quad \lim_{k\to\infty} \left\{ k \min_{1\le i\le k} \left\{D(\mathbf{X}^k, \boldsymbol{\Xi}^k, y^k, \mathbf{S}^k)\right\} \right\} = 0.$$

In order to use (Li et al. 2016b, Theorem 3), we need to write problem (10) as following

$$\begin{aligned} &\min\ g^*(-\boldsymbol{\Xi}) - \langle b,\, y\rangle + \delta(\|\mathbf{S}\|_2 \le c) \\ &\text{s.t. } \mathcal{F}(\boldsymbol{\Xi}) + \mathcal{A}_1^*(y) + \mathcal{G}(\mathbf{S}) = 0, \end{aligned}$$

where $\mathcal{F}, \mathcal{A}_1$ and $\mathcal{G}$ are linear operators such that for all $(\mathbf{\Xi}, y, \mathbf{S}) \in \Re^{p\times p} \times \Re^n \times \Re^{p\times p}$, $\mathcal{F}(\mathbf{\Xi}) = \mathbf{\Xi}$, $\mathcal{A}_1^*(y) = \mathcal{A}^*(y)$ and $\mathcal{G}(\mathbf{S}) = \mathbf{S}$. Clearly, $\mathcal{F} = \mathcal{G} = \mathcal{I}$ where $\mathcal{I} : \Re^{p\times p} \to \Re^{p\times p}$ is the identity map. Therefore, we have $\mathcal{A}_1\mathcal{A}_1^* \succ 0$ and $\mathcal{F}\mathcal{F}^* = \mathcal{G}\mathcal{G}^* = \mathcal{I} \succ 0$. Hence, the assumptions and conditions in (Li et al. 2016b, Theorem 3) are satisfied. The convergence results thus follow directly. Meanwhile, the non-ergodic iteration complexity results follows from (Chen et al. 2017, Theorem 6.1).

EC.12. Proof of Theorems 5 and 6

We only need to prove Theorem 6 as Theorem 5 is a special incidence. To prove Theorem 6, we first introduce the following lemma.

LEMMA EC.3. *Suppose that $\{x^k\}$ is the sequence generated by Algorithm 3. Then $\theta(x^{k+1}) \le \theta(x^k) - \frac{1}{2}\|x^{k+1} - x^k\|^2_{\mathcal{G}+2\mathcal{T}}$.*

For any $k \ge 0$, by the optimality condition of problem (10) at x^{k+1}, we know that there exist $\eta^{k+1} \in \partial p(x^{k+1})$ such that

$$0 = \nabla g(x^k) + (\mathcal{G} + \mathcal{T})(x^{k+1} - x^k) + \eta^{k+1} - \xi^k = 0.$$

Then for any $k \ge 0$, we deduce

$$\begin{aligned}
&\theta(x^{k+1}) - \theta(x^k) \le \widehat{\theta}(x^{k+1}; x^k) - \theta(x^k) \\
&= p(x^{k+1}) - p(x^k) + \langle x^{k+1} - x^k, \nabla g(x^k) - \xi^k\rangle + \tfrac{1}{2}\|x^{k+1} - x^k\|^2_{\mathcal{G}} \\
&\le \langle \nabla g(x^k) + \eta^{k+1} - \xi^k, x^{k+1} - x^k\rangle + \tfrac{1}{2}\|x^{k+1} - x^k\|^2_{\mathcal{G}} \\
&= -\tfrac{1}{2}\|x^{k+1} - x^k\|^2_{\mathcal{G}+2\mathcal{T}}.
\end{aligned}$$

This completes the proof of this lemma.

Now we are ready to prove Theorem 6.

From the optimality condition at x^{k+1}, we have that

$$0 \in \nabla g(x^k) + (\mathcal{G} + \mathcal{T})(x^{k+1} - x^k) + \partial p(x^{k+1}) - \xi^k.$$

Since $x^{k+1} = x^k$, this implies that

$$0 \in \nabla g(x^k) + \partial p(x^k) - \partial q(x^k),$$

i.e., x^k is a critical point. Observe that the sequence $\{\theta(x^k)\}$ is non-increasing since

$$\theta(x^{k+1}) \leq \widehat{\theta}(x^{k+1}; x^k) \leq \widehat{\theta}(x^k; x^k) = \theta(x^k), \quad k \geq 0.$$

Suppose that there exists a subsequence $\{x^{k_j}\}$ that converging to $\bar{x}$, i.e., one of the accumulation points of $\{x^k\}$. By Lemma EC.3 and the assumption that $\mathcal{G} + 2\mathcal{T} \succeq 0$, we know that for all $x \in \mathbb{X}$

$$\begin{aligned} &\widehat{\theta}(x^{k_{j+1}}; x^{k_{j+1}}) = \theta(x^{k_{j+1}}) \\ \leq &\theta(x^{k_j+1}) \leq \widehat{\theta}(x^{k_j+1}; x^{k_j}) \leq \widehat{\theta}(x; x^{k_j}). \end{aligned}$$

By letting $j \to \infty$ in the above inequality, we obtain that

$$\widehat{\theta}(\bar{x}; \bar{x}) \leq \widehat{\theta}(x; \bar{x}).$$

By the optimality condition of $\widehat{\theta}(x; \bar{x})$, we have that there exists $\bar{u} \in \partial p(\bar{x})$ and $\bar{v} \in \partial q(\bar{x})$ such that

$$0 \in \nabla g(\bar{x}) + \bar{u} - \bar{v}.$$

This implies that $(\nabla g(\bar{x}) + \partial p(\bar{x})) \cap \partial q(\bar{x}) \neq \emptyset$. To establish the rest of this proposition, we obtain from Lemma 1 that

$$\begin{aligned} &\lim_{t \to +\infty} \frac{1}{2} \sum_{i=0}^{t} \|x^{k+1} - x^k\|^2_{\mathcal{G}+2\mathcal{T}} \\ \leq &\liminf_{t \to +\infty} \left(\theta(x^0) - \theta(x^{k+1}) \right) \leq \theta(x^0) < +\infty\,, \end{aligned}$$

which implies $\lim_{i \to +\infty} \|x^{k+1} - x^i\|_{\mathcal{G}+2\mathcal{T}} = 0$. The proof of this theorem is thus complete by the positive definiteness of the operator $\mathcal{G} + 2\mathcal{T}$.