Spectral Entry-wise Matrix Estimation for Low-Rank Reinforcement Learning

Stefan Stojanovic
EECS
KTH, Stockholm, Sweden
stesto@kth.se

Yassir Jedra
EECS
KTH, Stockholm, Sweden
jedra@kth.se

Alexandre Proutiere
EECS
KTH, Stockholm, Sweden
alepro@kth.se

Abstract

We study matrix estimation problems arising in reinforcement learning (RL) with low-rank structure. In low-rank bandits, the matrix to be recovered specifies the expected arm rewards, and for low-rank Markov Decision Processes (MDPs), it may for example characterize the transition kernel of the MDP. In both cases, each entry of the matrix carries important information, and we seek estimation methods with low entry-wise error. Importantly, these methods further need to accommodate for inherent correlations in the available data (e.g. for MDPs, the data consists of system trajectories). We investigate the performance of simple spectral-based matrix estimation approaches: we show that they efficiently recover the singular subspaces of the matrix and exhibit nearly-minimal entry-wise error. These new results on low-rank matrix estimation make it possible to devise reinforcement learning algorithms that fully exploit the underlying low-rank structure. We provide two examples of such algorithms: a regret minimization algorithm for low-rank bandit problems, and a best policy identification algorithm for reward-free RL in low-rank MDPs. Both algorithms yield state-of-the-art performance guarantees.

1 Introduction

Learning succinct representations of the reward function or of the system state dynamics in bandit and RL problems is empirically known to significantly accelerate the search for efficient policies [38, 55, 13]. It also comes with interesting theoretical challenges. The design of algorithms learning and leveraging such representations and with provable performance guarantees has attracted considerable attention recently, but remains largely open. In particular, significant efforts have been made towards such design when the representation relies on a low-rank structure. In bandits, assuming such a structure means that the arm-to-reward function can be characterized by a low-rank matrix [37, 32, 7, 27]. In MDPs, it implies that the reward function, the Q-function or the transition kernels are represented by low-rank matrices [56, 4, 46, 60, 49]. In turn, the performance of algorithms exploiting low-rank structures is mainly determined by the accuracy with which we are able to estimate these matrices.

In this paper, we study matrix estimation problems arising in low-rank bandit and RL problems. Two major challenges are associated with these problems. (i) The individual entries of the matrix carry important operational meanings (e.g. in bandits, an entry could correspond to the average reward of an arm), and we seek estimation methods with low entry-wise error. Such requirement calls for a

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

fine-grained analysis, typically much more involved than that needed to only upper bound the spectral or Frobenius norm of the estimation error [22, 21, 12, 2, 54, 15, 53, 47]. (ii) Our estimation methods should further accommodate for inherent correlations in the available data (e.g., in MDPs, we have access to system trajectories, and the data is hence Markovian). We show that, essentially, spectral methods successfully deal with these challenges.

Contributions. 1) We introduce three matrix estimation problems. The first arises in low-rank bandits. The second corresponds to scenarios in RL where the learner wishes to estimate the (low-rank) transition kernel of a Markov chain and to this aim, has access to a generative model. The last problem is similar but assumes that the learner has access to system trajectories only, a setting referred to as the forward model in the RL literature. For all problems, we establish strong performance guarantees for simple spectral-based estimation approaches: these efficiently recover the singular subspaces of the matrix and exhibit nearly-minimal entry-wise error. To prove these results, we develop and combine involved leave-one-out arguments and Poisson approximation techniques (to handle the correlations in the data).

2) We apply the results obtained for our first matrix estimation problem to devise an efficient regret-minimization algorithm for low-rank bandits. We prove that the algorithm enjoys finite-time performance guarantees, with a regret at most roughly scaling as $(m+n)\log^3(T)\bar{\Delta}/\Delta_{\min}^2$ where (m,n) are the reward matrix dimensions, T is the time horizon, $\bar{\Delta}$ is the average of the reward gaps between the best arm and all other arms, and $\Delta_{\min}$ is the minimum of these gaps.

3) Finally, we present an algorithm for best policy identification in low-rank MDPs in the reward-free setting. The results obtained for the second and last matrix estimation problems imply that our algorithm learns an ϵ-optimal policy for any reward function using only a number of samples scaling as $O(nA/\epsilon^2)$ up to logarithmic factors, where n and A denote the number of states and actions, respectively. This sample complexity is mini-max optimal [28], and illustrates the gain achieved by leveraging the low-rank structure (without this structure, the sample complexity would be $\Omega(n^2A/\epsilon^2)$).

Notation. For any matrix $A \in \mathbb{R}^{m\times n}$, $A_{i,:}$ (resp. $A_{:,j}$) denotes its i-th row (resp. its j-th column), $A_{\min} = \min_{(i,j)} A_{i,j}$ and $A_{\max} = \max_{(i,j)} A_{i,j}$. We consider the following norms for matrices: $\|A\|$ denotes the spectral norm, $\|A\|_{1\to\infty} = \max_{i\in[m]} \|A_{i,:}\|_1$, $\|A\|_{2\to\infty} = \max_{i\in[m]} \|A_{i,:}\|_2$, and finally $\|A\|_\infty = \max_{(i,j)\in[m]\times[n]} |A_{i,j}|$. If the SVD of A is $U\Sigma V^\top$, we denote by $\mathrm{sgn}(A) = UV^\top$ the matrix sign function of A (see Definition 4.1 in [14]). $\mathcal{O}^{r\times r}$ denotes the set of $(r\times r)$ real orthogonal matrices. For any finite set $\mathcal{S}$, let $\mathcal{P}(\mathcal{S})$ be the set of distributions over $\mathcal{S}$. The notation $a(n,m,T) \lesssim b(n,m,T)$ (resp. $a(n,m,T) = \Theta(b(n,m,T))$) means that there exists a universal constant $C > 0$ (resp. $c, C > 0$) such that $a(n,m,T) \le Cb(n,m,T)$ (resp. $cb(n,m,T) \le a(n,m,T) \le Cb(n,m,T)$) for all n,m,T. Finally, we use $a\wedge b = \min(a,b)$ and $a\vee b = \max(a,b)$.

2 Models and Objectives

Let $M \in \mathbb{R}^{m\times n}$ be an unknown rank r matrix that we wish to estimate from T noisy observations of its entries. We consider matrices arising in two types of learning problems with low-rank structure, namely low-rank bandits and RL. The SVD of M is $U\Sigma V^\top$ where the matrices $U \in \mathbb{R}^{m\times r}$ and $V \in \mathbb{R}^{n\times r}$ contain the left and right singular vectors of M, respectively, and $\Sigma = \mathrm{diag}(\sigma_1,\dots,\sigma_r)$. We assume without loss of generality that the singular values have been ordered, i.e., $\sigma_1 \ge \dots \ge \sigma_r$. The accuracy of our estimate $\widehat{M}$ of M will be assessed using the following criteria:

(i) Singular subspace recovery. Let the SVD of $\widehat{M}$ be $\widehat{U}\widehat{\Sigma}\widehat{V}^\top$. To understand how well the singular subspaces of M are recovered, we will upper bound $\min_{O\in\mathcal{O}^{r\times r}} \|U - \widehat{U}O\|_{2\to\infty}$ and $\min_{O\in\mathcal{O}^{r\times r}} \|V - \widehat{V}O\|_{2\to\infty}$ (the $\min_{O\in\mathcal{O}^{r\times r}}$ problem corresponds to the orthogonal Procrustes problem and its solution aligns $\widehat{U}$ and U as closely as possible, see Remark 4.1 in [14]).

(ii) Matrix estimation. To assess the accuracy of $\widehat{M}$, we will upper bound the row-wise error $\|\widehat{M} - M\|_{1\to\infty}$ or $\|\widehat{M} - M\|_{2\to\infty}$, as well as the entry-wise error $\|\widehat{M} - M\|_\infty$ (the spectral error $\|\widehat{M} - M\|$ is easier to deal with and is presented in appendix only).

We introduce two classical quantities characterizing the heterogeneity and incoherence of the matrix M [11, 48]. Let $\kappa = \sigma_1/\sigma_r$, and let $\mu(U) = \sqrt{m/r}\|U\|_{2\to\infty}$ (resp. $\mu(V) = \sqrt{n/r}\|V\|_{2\to\infty}$) denote the row-incoherence (resp. column-incoherence) parameter of M. Let $\mu = \max\{\mu(U), \mu(V)\}$. Next, we specify the matrices M of interest in low-rank bandits and RL, and the way the data used for their estimation is generated.

Model I: Reward matrices in low-rank bandits. For bandit problems, M corresponds to the average rewards of various arms. To estimate M, the learner has access to data sequentially generated as follows. In each round $t = 1, \dots, T$, an arm $(i_t, j_t) \in [m] \times [n]$ is randomly selected (say uniformly at random for simplicity) and the learner observes $M_{i_t,j_t} + \xi_t$, an unbiased sample of the corresponding entry of M. $(\xi_t)_{t\geq 1}$ is a sequence of zero-mean and bounded random variables. Specifically, we assume that for all $t \geq 1$, $|\xi_t| \leq c_1\|M\|_\infty$ a.s., for some constant $c_1 > 0$.

Model II: Transition matrices in low-rank MDPs. In low-rank MDPs, we encounter Markov chains whose transition matrices have low rank r (refer to Section 5 for details). Let $P \in \mathbb{R}^{n\times n}$ be such a transition matrix. We assume that the corresponding Markov chain is irreducible with stationary distribution ν. The objective is to estimate P from the data consisting of samples of transitions of the chain. More precisely, from the data, we will estimate the *long-term frequency matrix* $M = \mathrm{diag}(\nu)P$ (M_{ij} is the limiting proportion of transitions from state i to state j as the trajectory grows large). Observe that M is of rank r, and that $P_{i,:} = M_{i,:}/\|M_{i,:}\|_1$. To estimate M, the learner has access to the data $(x_1, \dots, x_T) \in [n]^T$ generated according to one of the following two models.

(a) In the *generative* model, for any $t \in [T]$, if t is odd, x_t is selected at random according to some distribution ν_0, and x_{t+1} is sampled from $P_{x_t,:}$.

(b) In the *forward* model, the learner has access to a trajectory $(x_1, \dots, x_T)$ of length T of the Markov chain, where $x_1 \sim \nu_0$ and for any $t \geq 1$, $x_{t+1} \sim P_{x_t,:}$.

3 Matrix Estimation via Spectral Decomposition

In the three models (Models I, II(a) and II(b)), we first construct a matrix $\widetilde{M}$ directly from the data, and from there, we build our estimate $\widehat{M}$, typically obtained via spectral decomposition, i.e., by taking the best rank-r approximation of $\widetilde{M}$. In the remaining of this section, we let $\widehat{U}\widehat{\Sigma}\widehat{V}^\top$ denote the SVD of $\widehat{M}$. Next, we describe in more details how $\widetilde{M}$ is constructed in the three models, and analyze the corresponding estimation error.

3.1 Reward matrices

For Model I, for $t = 1, \dots, T$, we define $\widetilde{M}_t = \big((M_{i_t,j_t} + \xi_t)\mathbb{1}_{\{(i,j)=(i_t,j_t)\}}\big)_{i,j\in[m]\times[n]}$ and $\widetilde{M} = \frac{nm}{T}\sum_{t=1}^T \widetilde{M}_t$. Let $\widehat{M}$ denote the best rank-r approximation of $\widetilde{M}$.

Theorem 1. *Let $\delta > 0$. We introduce:*

$$\mathcal{B} = \sqrt{\frac{nm}{T}}\left(\sqrt{(n+m)\log\left(\frac{e(n+m)T}{\delta}\right)} + \log^{3/2}\left(\frac{e(n+m)T}{\delta}\right)\right).$$

Assume that $T \geq c\mu^4\kappa^2 r^2(n+m)\log^3\left(e(m+n)T/\delta\right)$ for some universal constant $c > 0$. Then there exists a universal constant $C > 0$ such that the following inequalities hold with probability at least $1-\delta$:

$$(i) \qquad \max\left(\|U - \widehat{U}(\widehat{U}^\top U)\|_{2\to\infty}, \|V - \widehat{V}(\widehat{V}^\top V)\|_{2\to\infty}\right) \leq C\frac{(\mu^3\kappa^2 r^{3/2})}{\sqrt{mn(n\wedge m)}}\mathcal{B},$$

$$(ii) \qquad \|\widehat{M} - M\|_{2\to\infty} \leq C\frac{(\mu^3\,\kappa^2 r^{3/2})}{\sqrt{m\wedge n}}\|M\|_\infty\mathcal{B},$$

$$(iii) \qquad \|\widehat{M} - M\|_\infty \leq C\left(\mu^{11/2}\,\kappa^2 r^{1/2} + \mu^3\kappa r^{3/2}\frac{m+n}{\sqrt{mn}}\right)\frac{1}{(n\wedge m)}\|M\|_\infty\mathcal{B}.$$

Corollary 2. *(Homogeneous reward matrix) When $m = \Theta(n)$, $\kappa = \Theta(1)$, $\mu = \Theta(1)$, $\|M\|_\infty = \Theta(1)$, $r = \Theta(1)$, we say that the reward matrix M is homogeneous. In this case, for any $\delta > 0$, when $T \geq c(n+m)\log^3\left(e(m+n)T/\delta\right)$ for some universal constant $c > 0$, we have with probability at least $1-\delta$:*

$$\max\left(\|U - \widehat{U}(\widehat{U}^\top U)\|_{2\to\infty}, \|V - \widehat{V}(\widehat{V}^\top V)\|_{2\to\infty}\right) \lesssim \frac{1}{\sqrt{T}}\log^{3/2}\left(\frac{(n+m)T}{\delta}\right),$$
$$\|\widehat{M} - M\|_{2\to\infty} \lesssim \frac{(n+m)}{\sqrt{T}}\log^{3/2}\left(\frac{(n+m)T}{\delta}\right),$$
$$\|\widehat{M} - M\|_\infty \lesssim \sqrt{\frac{(n+m)}{T}}\log^{3/2}\left(\frac{(n+m)T}{\delta}\right).$$

For a homogeneous reward matrix, $\|U\|_{2\to\infty} = \Theta(1/\sqrt{m})$ and $\|M\|_\infty = \Theta(1)$, and hence, from the above corollary, we obtain estimates whose relative errors (e.g., $\|\widehat{M} - M\|_\infty / \|M\|_\infty$) scale at most as $\sqrt{m/T}$ up to the logarithmic factor.

We may also compare the results of the above corollary to those of Theorem 4.4 presented in [14]. There, the data consists for each pair (i,j) of a noisy observation $M_{i,j} + E_{i,j}$. The $E_{i,j}$'s are independent across (i,j). This model is simpler than ours and does not include any correlation in the data. But it roughly corresponds to the case where $T = nm$ in our Model I. Despite having to deal with correlations, we obtain similar results as those of Theorem 4.4: for example, $\|\widehat{M} - M\|_\infty \lesssim \sqrt{1/(n+m)}$ (up to logarithmic terms) with high probability.

3.2 Transition matrices under the generative model

For Model II(a), the matrix $\widetilde{M}$ records the empirical frequencies of the transitions: for any pair of states (i,j), $\widetilde{M}_{i,j} = \frac{1}{\lfloor T/2 \rfloor}\sum_{k=1}^{\lfloor T/2 \rfloor} \mathbb{1}_{\{(x_{2k-1}, x_{2k}) = (i,j)\}}$. $\widehat{M}$ is the best rank-r approximation of $\widetilde{M}$ and the estimate $\widehat{P}$ of the transition matrix P is obtained normalizing the rows of $\widehat{M}$: for all $i \in [n]$,

$$\widehat{P}_{i,:} = \begin{cases} (\widehat{M}_{i,:})_+ / \|(\widehat{M}_{i,:})_+\|_1, & \text{if } \|(\widehat{M}_{i,:})_+\|_1 > 0, \\ \frac{1}{n}\mathbf{1}_n, & \text{if } \|(\widehat{M}_{i,:})_+\|_1 = 0. \end{cases} \tag{1}$$

where $(\cdot)_+$ is the function applying $\max(0,\cdot)$ component-wise and $\mathbf{1}_n$ is the n-dimensional vector of ones. The next theorem is a simplified version and a consequence of a more general and tighter theorem presented in App. B.2. To simplify the presentation of our results, we define $g(M,T,\delta) = n\log(\frac{n\sqrt{T}}{\delta})\max\left\{\mu^6\kappa^6 r^3, \frac{\log(\frac{n\sqrt{T}}{\delta})\mathbb{1}_{\{\exists \ell : T\|M_{\ell,:}\|_\infty \leq 1\}}}{\log(1+\frac{1}{T\|M\|_\infty})}\right\}$.

Theorem 3. *Let $\delta > 0$. Introduce $\mathcal{B} = \mu\kappa\sqrt{(r\|M\|_\infty/T)\log(n\sqrt{T}/\delta)}$. Assume that we have $(\nu_0)_{\min} = \min_{i\in[n]}(\nu_0)_i > 0$. If (a) $n \geq c\log^2(nT^{3/2}/\delta)$ and (b) $T \geq cg(M,T,\delta)$ for some universal constant $c > 0$, then there exists a universal constant $C > 0$ such that the following inequalities hold with probability at least $1-\delta$:*

$$(i) \qquad \max\left\{\|U - \widehat{U}(\widehat{U}^\top U)\|_{2\to\infty}, \|V - \widehat{V}(\widehat{V}^\top V)\|_{2\to\infty}\right\} \leq C\frac{\kappa\mu^2 r}{n\|M\|_\infty}\mathcal{B},$$
$$(ii) \qquad \|\widehat{M} - M\|_{2\to\infty} \leq C\kappa\mathcal{B}, \quad \|\widehat{P} - P\|_{1\to\infty} \leq C\frac{\kappa\sqrt{n}}{(\nu_0)_{\min}}\mathcal{B},$$
$$(iii) \qquad \|\widehat{M} - M\|_\infty \leq C\frac{\kappa\mu^2 r}{\sqrt{n}}\mathcal{B},$$
$$(iv) \qquad \|\widehat{P} - P\|_\infty \leq C\frac{\mathcal{B}}{(\nu_0)_{\min}}\left[\sqrt{n}\kappa\frac{\|M\|_\infty}{(\nu_0)_{\min}} + \left(1 + \frac{\kappa\mathcal{B}}{\sqrt{n}\,\|M\|_\infty}\right)\frac{\kappa\mu^2 r}{\sqrt{n}}\right],$$

where (iv) holds if in addition $T \geq cn\|M\|_\infty(\nu_0)_{\min}^{-2} r\mu^2\kappa^4\log(n\sqrt{T}/\delta)$

Note that in theorem, the condition (a) on n has been introduced just to simplify the expression of $\mathcal{B}$ (refer to App. B.2 for a full statement of the theorem without this condition).

Corollary 4. *(Homogeneous transition matrix) When $\kappa = \Theta(1)$, $\mu = \Theta(1)$, $r = \Theta(1)$, $M_{\max} = \Theta(M_{\min})$, we say that the frequency matrix M is homogeneous. If $T \geq cn\log(nT)$ for some universal constant $c > 0$, then we have with probability at least $1 - \min\{n^{-2}, T^{-1}\}$:*

$$\max\left\{\|U - \widehat{U}(\widehat{U}^\top U)\|_{2\to\infty}, \|V - \widehat{V}(\widehat{V}^\top V)\|_{2\to\infty}\right\} \lesssim \sqrt{\frac{\log(nT)}{T}},$$

$$\|\widehat{M} - M\|_{2\to\infty} \lesssim \frac{1}{n}\sqrt{\frac{\log(nT)}{T}},\ \|\widehat{M} - M\|_{\infty} \lesssim \frac{1}{n}\sqrt{\frac{\log(nT)}{nT}},$$

$$\|\widehat{P} - P\|_{1\to\infty} \lesssim \sqrt{\frac{n\log(nT)}{T}},\ \|\widehat{P} - P\|_{\infty} \lesssim \sqrt{\frac{\log(nT)}{nT}}.$$

For a homogeneous frequency matrix, $\|U\|_{2\to\infty} = \Theta(1/\sqrt{n})$, $\|M\|_{2\to\infty} = \Theta(1/n\sqrt{n})$, $\|M\|_\infty = \Theta(1/n^2)$, $\|P\|_{1\to\infty} = 1$, $\|P\|_\infty = \Theta(1/n)$. Thus for all these metrics, our estimates achieve a relative error scaling at most as $\sqrt{n/T}$ up to the logarithmic factor.

3.3 Transition matrices under the forward model

For Model II(b), we first split the data into τ subsets of transitions: for $k = 1, \ldots, \tau$, the k-th subset is $((x_k, x_{k+1}), (x_{k+\tau}, x_{k+1+\tau}), \ldots, (x_{k+(T_\tau-1)\tau}, x_{k+1+(T_\tau-1)\tau}))$ where $T_\tau = \lfloor T/\tau \rfloor$. By separating two transitions in the same subset, we break the inherent correlations in the data if τ is large enough. Now we let $\widetilde{M}^{(k)}$ be the matrix recording the empirical frequencies of the transitions in the k-th subset: $\widetilde{M}^{(k)}_{i,j} = \frac{1}{T_\tau}\sum_{l=0}^{T_\tau-1} \mathbb{1}_{\{(x_{k+l\tau}, x_{k+1+l\tau}) = (i,j)\}}$ for any pair of states (i, j). Let $\widehat{M}^{(k)}$ be the best r-rank approximation of $\widetilde{M}^{(k)}$. As in (1), we define the corresponding $\widehat{P}^{(k)}$. Finally we may aggregate these estimates $\widehat{M} = \frac{1}{\tau}\sum_{k=1}^{\tau}\widehat{M}^{(k)}$ and $\widehat{P} = \frac{1}{\tau}\sum_{k=1}^{\tau}\widehat{P}^{(k)}$. We present below the performance analysis for the estimates coming from a single subset; the analysis of the aggregate estimates easily follows.

For any $\varepsilon > 0$, we define the ε-mixing time of the Markov chain with transition matrix P as $\tau(\varepsilon) = \min\{t \geq 1 :\ \max_{1\leq i\leq n} \frac{1}{2}\|P^t_{i,:} - \nu^\top\|_1 \leq \varepsilon\}$, and its mixing time as $\tau^\star = \tau(1/4)$. The next theorem is a simplified version and a consequence of a more general and tighter theorem presented in App. B.3. To simplify the presentation, we define:

$$h(M, T, \delta) = n\tau^\star \log(\tfrac{n\sqrt{T}}{\delta}) \log(T\nu_{\min}^{-1}) \max\left\{\mu^6\kappa^6 r^3, \frac{\log^2(\frac{n\sqrt{T_\tau}}{\delta})\mathbb{1}_{\{\exists \ell : T_\tau\|M_{\ell,:}\|_\infty \leq 1\}}}{\log^2(1 + \frac{1}{T_\tau\|M\|_\infty})}\right\}.$$

Theorem 5. *Let $\delta > 0$. Assume that $\nu_{\min} = \min_{i\in[n]}\nu_i > 0$ and that $\tau/(\tau^\star\log(T\nu_{\min}^{-1})) \in [c_1, c_2]$ for some universal constants $c_2 > c_1 \geq 2$. Introduce:*

$$\mathcal{B} = \mu\kappa\sqrt{\frac{r\tau^\star\|M\|_\infty}{T}\log\left(\frac{n\sqrt{T_\tau}}{\delta}\right)\log\left(\frac{T}{\nu_{\min}}\right)}.$$

If (a) $n \geq c\tau^\star \log^{3/2}(nT^{3/2}/\delta)\log^{1/2}(T\nu_{\min}^{-1})$ and (b) $T \geq ch(M, T, \delta)$ for some universal constant $c > 0$, then there exists a universal constant $C > 0$ such that the following inequalities hold with probability at least $1 - \delta$:

$$(i) \qquad \max\left\{\|U - \widehat{U}(\widehat{U}^\top U)\|_{2\to\infty}, \|V - \widehat{V}(\widehat{V}^\top V)\|_{2\to\infty}\right\} \leq C\frac{\kappa\mu^2 r}{n\|M\|_\infty}\mathcal{B},$$

$$(ii) \qquad \|\widehat{M} - M\|_{2\to\infty} \leq C\kappa\mathcal{B}, \quad \|\widehat{P} - P\|_{1\to\infty} \leq C\frac{\kappa\sqrt{n}}{\nu_{\min}}\mathcal{B},$$

$$(iii) \qquad \|\widehat{M} - M\|_\infty \leq C\frac{\kappa\mu^2 r}{\sqrt{n}}\mathcal{B},$$

$$(iv) \qquad \|\widehat{P} - P\|_\infty \leq C\frac{\mathcal{B}}{\nu_{\min}}\left[\sqrt{n}\kappa\frac{\|M\|_\infty}{\nu_{\min}} + \left(1 + \frac{\kappa\mathcal{B}}{\sqrt{n}\,\|M\|_\infty}\right)\frac{\kappa\mu^2 r}{\sqrt{n}}\right],$$

where (iv) holds if in addition $T \geq cn\|M\|_\infty\nu_{\min}^{-2}\tau^\star r\mu^2\kappa^4\log(n\sqrt{T}/\delta)\log(T\nu_{\min}^{-1})$.

Note that our guarantees hold when τ roughly scales as $\tau^\star\log(T\nu_{\min}^{-1})$. Hence to select τ, one would need an idea of the latter quantity. It can be estimated typically using $\tau^\star\nu_{\min}^{-1}$ samples [61] (which is small when compared to the constraint $T \geq ch(M, T, \delta)$ as soon as $\nu_{\min} = \Omega(1/n)$). Further observe that in the theorem, the condition (a) can be removed (refer to App. B.3 for a full statement of the theorem without this condition).

Corollary 6. *(Homogeneous transition matrices) Assume that M is homogeneous (as defined in Corollary 4). Let $\tau = \log(Tn)$. If $T \geq cn\log^2(nT)$ for some universal constant $c > 0$, then we have with probability at least $1 - \min\{n^{-2}, T^{-1}\}$:*

$$\max\left\{\|U - \widehat{U}(\widehat{U}^\top U)\|_{2\to\infty}, \|V - \widehat{V}(\widehat{V}^\top V)\|_{2\to\infty}\right\} \lesssim \frac{1}{\sqrt{T}}\log(nT),$$

$$\|\widehat{M} - M\|_{2\to\infty} \lesssim \frac{1}{n\sqrt{T}}\log(nT),\ \|\widehat{M} - M\|_{\infty} \lesssim \frac{1}{n\sqrt{nT}}\log(nT),$$

$$\|\widehat{P} - P\|_{1\to\infty} \lesssim \sqrt{\frac{n}{T}}\log(nT),\ \|\widehat{P} - P\|_{\infty} \lesssim \frac{1}{\sqrt{nT}}\log(nT).$$

As for the generative model, for a homogeneous frequency matrix, our estimates achieve a relative error scaling at most as $\sqrt{n/T}$ up to the logarithmic factor for all metrics. Note that up to a logarithmic factor, the upper bound for $\|\widehat{P} - P\|_{1\to\infty}$ (and similarly for $\widehat{M}$) matches the minimax lower bound derived in [63].

3.4 Elements of the proofs

The proofs of the three above theorems share similar arguments. We only describe elements of the proof of Theorem 5, corresponding to the most challenging model. The most difficult result concerns the singular subspace recovery (the upper bounds (i) in our theorems), and it can be decomposed into the following three steps. The first two steps are meant to deal with the Markovian nature of the data. The third step consists in applying a leave-one-out analysis to recover the singular subspaces.

Step 1: Multinomial approximation of Markovian data. We treat the matrix $\widetilde{M}^{(k)}$ arising from one subset of data, and for simplicity, we remove the superscript (k), i.e., $\widetilde{M} = \widetilde{M}^{(k)}$. Note that $T_\tau\widetilde{M}$ is a matrix recording the numbers of transitions observed in the data for any pair of states: denote by $N_{i,j}$ this number for (i,j). We approximate the joint distribution of $N = (N_{i,j})_{(i,j)}$ by a multinomial distribution with n^2 components and parameter $T_\tau M_{i,j}$ for component (i,j). Denote by $Z = (Z_{i,j})_{(i,j)}$ the corresponding multinomial random variable. Using the mixing property of the Markov chain and the choice of τ, we establish (see Lemma 21 in App. C) that for any subset $\mathcal{Z}$ of $\{z \in \mathbb{N}^{n^2} : \sum_{(i,j)} z_{i,j} = T_\tau\}$, we have $\mathbb{P}[N \in \mathcal{Z}] \leq 3\mathbb{P}[Z \in \mathcal{Z}]$.

Step 2: Towards Poisson random matrices with independent entries. The random matrix Z does not have independent entries. Independence is however a requirement if we wish to apply the leave-one-out argument. Consider the random matrix Y whose entries are independent Poisson random variables with mean $T_\tau M_{i,j}$ for the (i,j)-th entry. We establish the following connection between the distribution of Z and that of Y: for any $\mathcal{Z} \subset \mathbb{N}^{n^2}$, we have $\mathbb{P}[Z \in \mathcal{Z}] \leq e\sqrt{T_\tau}\mathbb{P}[Y \in \mathcal{Z}]$. Refer to Lemma 22 in App. C for details.

Step 3: The leave-one-out argument for Poisson matrices. Combining the two first steps provides a connection between the observation matrix $\widetilde{M}$ and a Poisson matrix Y with independent entries. This allows us to apply a leave-one-out analysis to $\widetilde{M}$ as if it had independent entries (replacing $\widetilde{M}$ by Y). The analysis starts by applying the standard dilation trick (see Section 4.10 in [14]) so as to make $\widetilde{M}$ symmetric. Then, we can decompose the error $\|U - \widehat{U}(\widehat{U}^\top U)\|_{2\to\infty}$ (see Lemma 32 in App. E) into several terms. The most challenging of these terms is $\|(M - \widetilde{M})(U - \widehat{U}(\widehat{U}^\top U))\|_{2\to\infty} = \max_{l\in[n]} \|(M_{l,:} - \widetilde{M}_{l,:})(U - \widehat{U}(\widehat{U}^\top U))\|_2$ because of inherent dependence between $M - \widetilde{M}$ and $U - \widehat{U}(\widehat{U}^\top U)$. The leave-one-out analysis allows us to decouple this statistical dependency. It consists in exploiting the row and column independence of matrix $\widetilde{M}$ to approximate $\|(M_{l,:} - \widetilde{M}_{l,:})(U - \widehat{U}(\widehat{U}^\top U))\|_2$ by $\|(M_{l,:} - \widetilde{M}_{l,:})(U - \widehat{U}^{(l)}((\widehat{U}^{(l)})^\top U)\|_2$ where $\widehat{U}^{(l)}$ is the matrix of eigenvectors of matrix $\widetilde{M}^{(l)}$ obtained by zeroing the l-th row and column of $\widetilde{M}$. By construction, $(M_{l,:} - \widetilde{M}_{l,:})$ and $U - \widehat{U}^{(l)}((\widehat{U}^{(l)})^\top U)$ are independent, which simplifies the analysis. The proof is completed by a further appropriate decomposition of this term, combined with concentration inequalities for random Poisson matrices (see App. D).

4 Regret Minimization in Low-Rank Bandits

Consider a low-rank bandit problem with a homogeneous rank-r reward matrix M. We wish to devise an algorithm π with low regret. π selects in round t an entry (i_t^π, j_t^π) based on previous observations, and receives as a feedback the noisy reward $M_{i_t^\pi, j_t^\pi} + \xi_t$. The regret up to round T is defined by $R^\pi(T) = TM_{i^\star,j^\star} - \mathbb{E}[\sum_{t=1}^T M_{i_t^\pi, j_t^\pi}]$, where $(i^\star, j^\star)$ is an optimal entry. One could think of a simple Explore-Then-Commit (ETC) algorithm, where in the first phase entries are sampled uniformly at random, and where in a second phase, the algorithm always selects the highest entry of $\widehat{M}$ built using the samples gathered in the first phase and obtained by spectral decomposition. When the length of the first phase is $T^{2/3}(n+m)^{1/3}$, the ETC algorithm would yield a regret upper bounded by $O(T^{2/3}(n+m)^{1/3})$ for $T = \Omega((n+m)\log^3(n+m))$.

To get better regret guarantees, we present SME-AE (Successive Matrix Estimation and Arm Elimination), an algorithm meant to identify the best entry as quickly as possible with a prescribed level of certainty. After the SME-AE has returned the estimated best entry, we commit and play this entry for the remaining rounds. The pseudo-code of SME-AE is presented in Algorithm 1. The algorithm runs in epochs: in epoch ℓ, it samples T_ℓ entries uniformly at random among all entries (in T_ℓ, the constant C just depends on upper bounds of the parameters μ, κ, and $\|M\|_\infty$, refer to App. G); from these samples, a matrix $\widehat{M}^{(\ell)}$ is estimated and $\mathcal{A}_\ell$, the set of candidate arms, is pruned. The pruning procedure is based on the estimated gaps: $\widehat{\Delta}_{i,j}^{(\ell)} = \widehat{M}_\star^{(\ell)} - \widehat{M}_{i,j}^{(\ell)}$ where $\widehat{M}_\star^{(\ell)} = \max_{i,j} \widehat{M}_{i,j}^{(\ell)}$.

Algorithm 1: **S**uccesive **M**atrix **E**stimation and **A**rm **E**limination (**SME-AE**)

```
Input: Arms $[m] \times [n]$, confidence level $\delta$
$\ell = 1$ ;
$\mathcal{A}_1 = [m] \times [n]$;
while $|\mathcal{A}_\ell| > 1$ do
    $\delta_\ell = \delta/\ell^2$;
    $T_\ell = \left\lceil C\left(2^{\ell+2}\right)^2 (m+n)\log^3\left(2^{2\ell+4}(m+n)/\delta_\ell\right)\right\rceil$ ;
    Sample uniformly at random $T_\ell$ entries from $\mathcal{A}_1$: $(M_{i_t,j_t} + \xi_t)_{t=1,\dots,T_\ell}$ ;
    Estimate $\widehat{M}^{(\ell)}$ via spectral decomposition as described in Section 3.1 ;
    $\mathcal{A}_{\ell+1} = \left\{(i,j) \in \mathcal{A}_\ell : \widehat{\Delta}_{i,j}^{(\ell)} \le 2^{-(\ell+2)}\right\}$; $\ell = \ell + 1$;
end
Output: Recommend the remaining pair $(\hat{\imath}_\tau, \hat{\jmath}_\tau)$ in $\mathcal{A}_\ell$.
```

The following theorem characterizes the performance of SME-AE and the resulting regret. To simplify the notation, we introduce the gaps: for any entry (i,j), $\Delta_{i,j} = (M_{i^\star,j^\star} - M_{i,j})$, $\Delta_{\min} = \min_{(i,j):\Delta_{i,j}>0} \Delta_{i,j}$, $\Delta_{\max} = \max_{(i,j)} \Delta_{i,j}$, and $\bar{\Delta} = \sum_{(i,j)} \Delta_{i,j}/(mn)$. We define the function $\psi(n,m,\delta) = \frac{c(m+n)\log(e/\Delta_{\min})}{\Delta_{\min}^2} \log^3\left(\frac{e(m+n)\log(e/\Delta_{\min})}{\Delta_{\min}\delta}\right)$ for some universal constant $c > 0$.

Theorem 7. *(Best entry identification) For any $\delta \in (0,1)$, SME-AE(δ) stops at time τ and recommends arm $(\hat{\imath}_\tau, \hat{\jmath}_\tau)$ with the guarantee $\mathbb{P}\big((\hat{\imath}_\tau, \hat{\jmath}_\tau) = (i^\star, j^\star), \tau \le \psi(n,m,\delta)\big) \ge 1 - \delta$. Moreover, for any $T \ge 1$ and $\alpha > 0$, the sample complexity τ of SME-AE($1/T^\alpha$) satisfies $\mathbb{E}[\tau \wedge T] \le \psi(n,m,T^{-\alpha}) + T^{1-\alpha}$.*
(Regret) Let $T \ge 1$. Consider the algorithm π that first runs SME-AE($1/T^2$) and then commits to its output $(\hat{\imath}_\tau, \hat{\jmath}_\tau)$ after τ. We have: $R^\pi(T) \le \bar{\Delta}\left(\psi(n,m,T^{-2}) + 1\right) + \frac{\Delta_{\max}}{T}$.

The proof of Theorem 7 is given in App. G. Note that the regret upper bounds hold for any time horizon $T \ge 1$, and that it scales as $O((m+n)\log^3(T)\bar{\Delta}/\Delta_{\min}^2)$ (up to logarithmic factors in m, n and $1/\Delta_{\min}$). The cubic dependence in $\log^3(T)$ is an artifact of our proof techniques. More precisely, it is due to the Poisson approximation used to obtain entry-wise guarantees. Importantly, for any time horizon, the regret upper bound only depends on $(m+n)$ rather than mn (the number of arms / entries), and hence, the low-rank structure is efficiently exploited. If we further restrict our attention to problems with gap ratio $\Delta_{\max}/\Delta_{\min}$ upper bounded by ζ, our regret upper bound becomes $O(\zeta(m+n)\log^3(T)/\Delta_{\min})$, and can be transformed into the minimax gap-independent upper bound $O(\zeta((m+n)T)^{1/2}\log^2(T))$, see App. G. Finally note that $\Omega(((m+n)T)^{1/2})$ is an obvious minimax regret lower bound for our low-rank bandit problem.

A very similar low-rank bandit problem has been investigated in [7]. There, under similar assumptions (see Assumption 1 and Definition 1), the authors devise an algorithm with both gap-dependent and gap-independent regret guarantees. The latter are difficult to compare with ours. Their guarantees exhibit a better dependence in T and $\Delta_{\min}$, but worse in the matrix dimensions n and m. Indeed in our model, $b^\star$ in [7] corresponds to $\|M\|_{2\to\infty}$ and scales as $\sqrt{n}$. As a consequence, the upper bounds in [7] have a dependence in n and m scaling as $\sqrt{n}(n+m)$ in the worst case for gap-dependent guarantees and even nm (through the constant C_2 in [7]) for gap-independent guarantees.

5 Representation Learning in Low-Rank MDPs

The results derived for Models II(a) and II(b) are instrumental towards representation learning and hence towards model-based or reward-free RL in low-rank MDPs. In this section, we provide an example of application of these results, and mention other examples in Section 7. A low-rank MDP is defined by $(\mathcal{S},\mathcal{A},\{P^a\}_{a\in\mathcal{A}},R,\gamma)$ where $\mathcal{S}$, $\mathcal{A}$ denote state and action spaces of cardinalities n and A, respectively, P^a denotes the rank-r transition matrix when taking action a, R is the reward function, and γ is the discount factor. We assume that all rewards are in $[0,1]$. The value function of a policy $\pi:\mathcal{S}\to\mathcal{A}$ is defined as $V_R^\pi(x)=\mathbb{E}[\sum_{t=1}^\infty \gamma^{t-1}R(x_t^\pi,\pi_t(x_t^\pi))|x_1^\pi=x]$ where x_t^π is the state visited under π in round t. We denote by $\pi^\star(R)$ an optimal policy (i.e., with the highest value function).

Reward-free RL. In the reward-free RL setting (see e.g. [36, 30, 66]), the learner does not receive any reward signal during the exploration process. The latter is only used to construct estimates $\{\widehat{P}^a\}_{a\in\mathcal{A}}$ of $\{P^a\}_{a\in\mathcal{A}}$. The reward function R is revealed at the end, and the learner may compute $\hat{\pi}(R)$ an optimal policy for the MDP $(\mathcal{S},\mathcal{A},\{\widehat{P}^a\}_{a\in\mathcal{A}},R,\gamma)$. The performance of this model-based approach is often assessed through $\Gamma=\sup_R\|V_R^{\pi^\star(R)}-V_R^{\hat{\pi}(R)}\|_\infty$. In tabular MDP, to identify an ϵ-optimal policy for all reward functions, i.e., to ensure that $\Gamma\le\epsilon$, we believe that the number of samples that have to be collected should be $\Omega(\mathrm{poly}(\frac{1}{1-\gamma})\frac{n^2A}{\epsilon^2})$ (the exact degree of the polynomial in $1/(1-\gamma)$ has to be determined). This conjecture is based on the sample complexity lower bounds derived for reward-free RL in episodic tabular MDP [30, 43]. Now for low-rank MDPs, the equivalent lower bound would be $\Omega(\mathrm{poly}(\frac{1}{1-\gamma})\frac{nA}{\epsilon^2})$ [28] (this minimax lower bound is valid for Block MDPs, a particular case of low-rank MDPs).

Leveraging our low-rank matrix estimation guarantees, we propose an algorithm matching the aforementioned sample complexity lower bound (up to logarithmic factors) at least when the frequency matrices $\{M^a\}_{a\in\mathcal{A}}$ are homogeneous. The algorithm consists of two phases: (1) in the model estimation phase, it collects A trajectories, each of length T/A, corresponding to the Markov chains with transition matrices $\{P^a\}_{a\in\mathcal{A}}$. From this data, it uses the spectral decomposition method described in §3 to build estimates $\{\widehat{P}^a\}_{a\in\mathcal{A}}$. (2) In the planning phase, based on the reward function R, it computes the best policy $\hat{\pi}(R)$ for the MDP $(\mathcal{S},\mathcal{A},\{\widehat{P}^a\}_{a\in\mathcal{A}},R,\gamma)$. The following theorem summarizes the performance of this algorithm. To simplify the presentation, we only provide the performance guarantees of the algorithm for homogeneous transition matrices (guarantees for more general matrices can be derived plugging in the results from Theorem 5).

Theorem 8. *Assume that for any $a\in\mathcal{A}$, M^a is homogeneous (as defined in Corollary 4). If $T\ge cnA\log^2(nAT)$ for some universal constant $c>0$, then we have with probability at least $1-\min\{n^{-2},T^{-1}\}$: $\Gamma=\sup_R\|V_R^{\pi^\star(R)}-V_R^{\hat{\pi}(R)}\|_\infty\lesssim\frac{1}{(1-\gamma)^2}\sqrt{\frac{nA}{T}}\log(nAT)$.*

Theorem 8 is a direct consequence of Corollary 6 and of the fact that for any reward function R: $\|V_R^{\pi^\star(R)}-V_R^{\hat{\pi}(R)}\|_\infty\le\frac{2\gamma}{(1-\gamma)^2}\max_{a\in\mathcal{A}}\|P^a-\widehat{P}^a\|_{1\to\infty}$, see App. A. The theorem implies that if we wish to guarantee $\Gamma\le\epsilon$, we just need to collect $O(\frac{nA}{\epsilon^2(1-\gamma)^4})$ samples up to a logarithmic factor.

This sample complexity is minimax optimal in n, A, and ϵ in view of the lower bound presented in [28].

6 Related Work

Low-rank matrix estimation. Until recently, the main efforts on low-rank matrix recovery were focused on guarantees w.r.t. the spectral or Frobenius norms, see e.g. [19] and references therein. The first matrix estimation and subspace recovery guarantees in $\ell_{2\to\infty}$ and ℓ_∞ were established in [22], [21] via a more involved perturbation analysis than the classical Davis-Kahan bound. An alternative approach based on a leave-one-out analysis was proposed in [2], and further refined in [10, 12, 17], see [14] for a survey. Some work have also adapted the techniques beyond the independent noise assumption [39, 1, 5], but for very specific structural dependence. We deal with a stronger dependence, and in particular with Markovian data (an important scenario in RL).

The estimation of low-rank transition matrices of Markov chains has been studied in [63, 9] using spectral methods and in [40, 67] using maximum-likelihood approaches. [63] does not conduct any fine-grained subspace recovery analysis (such as the leave-one-out), and hence the results pertaining to the $\|\cdot\|_{1\to\infty}$-guarantees are questionable; refer to App. H for a detailed justification. All these papers do not present entry-wise guarantees.

It is worth mentioning that there exist other methods for matrix estimation that do not rely on spectral decompositions like ours, yet enjoy entry-wise matrix estimation guarantees [51, 3, 50]. However, these methods require different assumptions than ours that may be too strong for our purposes, notably having access to the so-called anchor rows and columns. Moreover, we do not know if these methods also lead to guarantees for subspace recovery in the norm $\|\cdot\|_{2\to\infty}$, nor how to extend those results to settings with dependent noise.

Low-rank bandits. Low-rank structure in bandits has received a lot of attention recently [35, 37, 32, 57, 41, 7, 34, 27]. Different set-ups have been proposed (refer to App. H for a detailed exposition, in particular, we discuss how the settings proposed in [32, 7] are equivalent), and regret guarantees in an instance dependent and minimax sense have been both established.

Typically minimax regret guarantees in bandits scale as $\sqrt{T}$, but the scaling in dimension may defer when dealing with a low rank structure [32, 34, 7]. In [32], the authors also leverage spectral methods. They reduce the problem to a linear bandit of dimension nm but where only roughly $n+m$ dimensions are relevant. This entails that a regret lower bound of order $(n+m)\sqrt{T}$ is inevitable. Actually, in their reduction to linear bandits, they only use a subspace recovery in Frobenius norm, which perhaps explains the scaling $(n+m)^{3/2}$ in their regret guarantees. It is worth noting that in [34], the authors manage to improve upon the work [32] and obtain a scaling order $(m+n)$ in the regret. Our algorithm leverages entry-wise guarantees which rely on a stronger subspace recovery guarantee. This allows us to obtain a scaling $\sqrt{n+m}$ in the regret. The work of [27] is yet another closely related work to ours. There, the authors propose an algorithm achieving a regret of order $\text{polylog}(n+m)\sqrt{T}$ for a contextual bandit problem with low rank structure. However, their result only holds for rank 1 and their observation setup is different than ours because in their setting, the learner observes m entries per round while in ours the learner only observes one entry per round. In [7], the authors use matrix estimation with nuclear norm penalization to estimate the matrix M. Their regret guarantees are already discussed in §4.

Some instance-dependent guarantees with logarithmic regret for low rank bandits have been established in [35, 37, 57]. However, these results suffer what may be qualified as serious limitations. Indeed, [35, 57] provide instance dependent regret guarantees but only consider low-rank bandits with rank 1, and the regret bounds of [35] are expressed in terms of the so-called column and row gaps (see their Theorem 1) which are distinct from the standard gap notions. [37] extend the results in [35] to rank r with the limitation that they require stronger assumptions than ours. Moreover, the computational complexity of their algorithm depends exponentially on the rank r; they require a search over spaces of size $\binom{m}{r}$ and $\binom{n}{r}$. Our proposed algorithm does not suffer from such limitations.

We wish to highlight that our entry-wise guarantees for matrix estimation are the key enabling tool that led us to the design and analysis of our proposed algorithm. In fact, the need for such guarantees arises naturally in the analysis of gap-dependent regret bounds (see Appendix G.1). Therefore, we believe that such guarantees can pave the way towards better, faster, and efficient algorithms for bandits with low-rank structure.

Low-rank Reinforcement Learning. RL with low rank structure has been recently extensively studied but always in the function approximation framework [29, 18, 20, 44, 24, 65, 56, 4, 46, 59, 60, 49]. There, the transition probabilities can be written as $\phi(x,a)^\top \mu(x')$ where the unknown feature functions $\phi(x,a), \mu(x') \in \mathbb{R}^r$ belong to some specific class $\mathcal{F}$ of functions. The major issue with algorithms proposed in this literature is that they rely on strong computational oracles (e.g., ERM, MLE), see [33, 25, 64] for detailed discussions. In contrast, we do not assume that the transition matrices are constructed based on a given restricted class of functions, and our algorithms do not rely on any oracle and are computationally efficient. In [51, 50], the authors also depart from the function approximation framework. There, they consider a low rank structure different than ours. Their matrix estimation method enjoys an entry-wise guarantee, but requires to identify a subset of rows and columns spanning the range of the full matrix. Moreover, their results are only limited the generative models, which allows to actually rely on independent data samples.

7 Conclusion and Perspectives

In this paper, we have established that spectral methods efficiently recover low-rank matrices even in correlated noise. We have investigated noise correlations that naturally arise in RL, and have managed to prove that spectral methods yield nearly-minimal entry-wise error. Our results for low-rank matrix estimation have been applied to design efficient algorithms in low-rank RL problems and to analyze their performance. We believe that these results may find many more applications in low-rank RL. They can be applied (i) to reward-free RL in episodic MDPs (this setting is easier than that presented in §5 since successive episodes are independent); (ii) to scenarios corresponding to offline RL [62] where the data consists of a single trajectory generated under a given behavior policy (from this data, we can extract the transitions (x, a, x') where a given action a is involved and apply the spectral method to learn $\widehat{P}^a$); (iii) to traditional RL where the reward function R has to be learnt (learning R is a problem that lies in some sense between the inference problems in our Models I and II); (iv) to model-free RL where we would directly learn the Q function as done in [52] under a generative model; (v) to low-rank RL problems with continuous state spaces (this can be done if the transition probabilities are smooth in the states, and by combining our methods to an appropriate discretization of the state space).

Acknowledgment

This research was supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

[1] Emmanuel Abbe, Jianqing Fan, and Kaizheng Wang. An ℓ_p theory of pca and spectral clustering. *The Annals of Statistics*, 50(4):2359–2385, 2022. (Cited on page 9.)

[2] Emmanuel Abbe, Jianqing Fan, Kaizheng Wang, and Yiqiao Zhong. Entrywise eigenvector analysis of random matrices with low expected rank. *Annals of statistics*, 48(3):1452, 2020. (Cited on pages 2, 9, and 34.)

[3] Alekh Agarwal, Nan Jiang, Sham M Kakade, and Wen Sun. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, pages 10–4, 2019. (Cited on pages 9 and 18.)

[4] Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun. FLAMBE: Structural Complexity and Representation Learning of Low Rank MDPs. In *Advances in Neural Information Processing Systems*, volume 33, pages 20095–20107. Curran Associates, Inc., 2020. (Cited on pages 1 and 10.)

[5] Joshua Agterberg, Zachary Lubberts, and Carey E Priebe. Entrywise estimation of singular vectors of low-rank matrices with heteroskedasticity and dependence. *IEEE Transactions on Information Theory*, 68(7):4618–4650, 2022. (Cited on page 9.)

[6] Afonso S Bandeira and Ramon Van Handel. Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *The Annals of Probability*, 44(4):2479–2506, 2016. (Cited on page 31.)

[7] Mohsen Bayati, Junyu Cao, and Wanning Chen. Speed up the cold-start learning in two-sided bandits with many arms. *arXiv preprint arXiv:2210.00340*, 2022. (Cited on pages 1, 8, 9, and 50.)

[8] George Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45, 1962. (Cited on page 26.)

[9] Shujun Bi, Zhen Yin, and Yihong Weng. A low-rank spectral method for learning markov models. *Optimization Letters*, 17(1):143–162, 2023. (Cited on page 9.)

[10] T Tony Cai and Anru Zhang. Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *The Annals of Statistics*, 46(1):60–89, 2018. (Cited on page 9.)

[11] Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010. (Cited on page 3.)

[12] Joshua Cape, Minh Tang, and Carey E Priebe. The two-to-infinity norm and singular subspace geometry with applications to high-dimensional statistics. *The Annals of Statistics*, 47(5):2405–2439, 2019. (Cited on pages 2, 9, 18, 35, and 43.)

[13] Yash Chandak, Shantanu Thakoor, Zhaohan Daniel Guo, Yunhao Tang, Remi Munos, Will Dabney, and Diana L Borsa. Representations and exploration for deep reinforcement learning using singular value decomposition. In *Proc. of ICML*, 2023. (Cited on page 1.)

[14] Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma, et al. Spectral methods for data science: A statistical perspective. *Foundations and Trends® in Machine Learning*, 14(5):566–806, 2021. (Cited on pages 2, 4, 6, 9, 34, 39, 40, 41, and 43.)

[15] Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma, and Yuling Yan. Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. *SIAM journal on optimization*, 30(4):3098–3121, 2020. (Cited on page 2.)

[16] Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma, and Yuling Yan. Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. *SIAM Journal on Optimization*, 30(4):3098–3121, 2020. (Cited on page 50.)

[17] Yuxin Chen, Jianqing Fan, Cong Ma, and Kaizheng Wang. Spectral method and regularized mle are both optimal for top-k ranking. *Annals of statistics*, 47(4):2204, 2019. (Cited on page 9.)

[18] Christoph Dann, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. On Oracle-Efficient PAC RL with Rich Observations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. (Cited on page 10.)

[19] Mark A. Davenport and Justin K. Romberg. An overview of low-rank matrix recovery from incomplete observations. *IEEE J. Sel. Top. Signal Process.*, 10(4):608–622, 2016. (Cited on page 9.)

[20] Simon Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudik, and John Langford. Provably efficient RL with Rich Observations via Latent State Decoding. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1665–1674. PMLR, 09–15 Jun 2019. (Cited on page 10.)

[21] Justin Eldridge, Mikhail Belkin, and Yusu Wang. Unperturbed: spectral analysis beyond davis-kahan. In *Algorithmic Learning Theory*, pages 321–358. PMLR, 2018. (Cited on pages 2 and 9.)

[22] Jianqing Fan, Weichen Wang, and Yiqiao Zhong. An ℓ_∞ eigenvector perturbation bound and its application to robust covariance estimation. *Journal of Machine Learning Research*, 18(207):1–42, 2018. (Cited on pages 2 and 9.)

[23] Vivek Farias, Andrew A Li, and Tianyi Peng. Near-optimal entrywise anomaly detection for low-rank matrices with sub-exponential noise. In *International Conference on Machine Learning*, pages 3154–3163. PMLR, 2021. (Cited on page 34.)

[24] Dylan Foster, Alexander Rakhlin, David Simchi-Levi, and Yunzong Xu. Instance-Dependent Complexity of Contextual Bandits and Reinforcement Learning: A Disagreement-Based Perspective. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 2059–2059. PMLR, 15–19 Aug 2021. (Cited on page 10.)

[25] Noah Golowich, Ankur Moitra, and Dhruv Rohatgi. Learning in Observable POMDPs, without Computationally Intractable Oracles. In *Advances in Neural Information Processing Systems*, volume 35. Curran Associates, Inc., 2022. (Cited on page 10.)

[26] Samuel B Hopkins, Tselil Schramm, Jonathan Shi, and David Steurer. Fast spectral algorithms from sum-of-squares proofs: tensor decomposition and planted sparse vectors. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 178–191, 2016. (Cited on page 26.)

[27] Prateek Jain and Soumyabrata Pal. Online low rank matrix completion. In *Proc. of ICLR*, 2023. (Cited on pages 1, 9, and 50.)

[28] Yassir Jedra, Junghyun Lee, Alexandre Proutiere, and Se-Young Yun. Nearly optimal latent state decoding in block mdps. In *International Conference on Artificial Intelligence and Statistics*, pages 2805–2904. PMLR, 2023. (Cited on pages 2, 8, and 9.)

[29] Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E. Schapire. Contextual decision processes with low Bellman rank are PAC-learnable. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1704–1713. PMLR, 06–11 Aug 2017. (Cited on page 10.)

[30] Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-Free Exploration for Reinforcement Learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4870–4879. PMLR, 13–18 Jul 2020. (Cited on page 8.)

[31] Kwang-Sung Jun, Rebecca Willett, Stephen Wright, and Robert Nowak. Bilinear bandits with low-rank structure. In *International Conference on Machine Learning*, pages 3163–3172. PMLR, 2019. (Cited on page 50.)

[32] Kwang-Sung Jun, Rebecca Willett, Stephen J. Wright, and Robert D. Nowak. Bilinear bandits with low-rank structure. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 3163–3172. PMLR, 2019. (Cited on pages 1 and 9.)

[33] Daniel Kane, Sihan Liu, Shachar Lovett, and Gaurav Mahajan. Computational-statistical gap in reinforcement learning. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 1282–1302. PMLR, 02–05 Jul 2022. (Cited on page 10.)

[34] Yue Kang, Cho-Jui Hsieh, and Thomas Chun Man Lee. Efficient frameworks for generalized low-rank matrix bandit problems. *Advances in Neural Information Processing Systems*, 35:19971–19983, 2022. (Cited on pages 9 and 50.)

[35] Sumeet Katariya, Branislav Kveton, Csaba Szepesvari, Claire Vernade, and Zheng Wen. Stochastic Rank-1 Bandits. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 392–401. PMLR, 20–22 Apr 2017. (Cited on page 9.)

[36] Emilie Kaufmann, Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Edouard Leurent, and Michal Valko. Adaptive Reward-Free Exploration. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pages 865–891. PMLR, 16–19 Mar 2021. (Cited on page 8.)

[37] Branislav Kveton, Csaba Szepesvari, Anup Rao, Zheng Wen, Yasin Abbasi-Yadkori, and S. Muthukrishnan. Stochastic low-rank bandits, 2017. (Cited on pages 1 and 9.)

[38] Michael Laskin, Aravind Srinivas, and Pieter Abbeel. CURL: Contrastive Unsupervised Representations for Reinforcement Learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5639–5650. PMLR, 13–18 Jul 2020. (Cited on page 1.)

[39] Lihua Lei. Unified $\ell_{2\to\infty}$ eigenspace perturbation theory for symmetric random matrices. *arXiv preprint arXiv:1909.04798*, 2019. (Cited on page 9.)

[40] Xudong Li, Mengdi Wang, and Anru Zhang. Estimation of markov chain via rank-constrained likelihood. In *International Conference on Machine Learning*, pages 3033–3042. PMLR, 2018. (Cited on page 9.)

[41] Yangyi Lu, Amirhossein Meisami, and Ambuj Tewari. Low-rank generalized linear bandit problems. In Arindam Banerjee and Kenji Fukumizu, editors, *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pages 460–468. PMLR, 2021. (Cited on page 9.)

[42] Andrew D McRae and Mark A Davenport. Low-rank matrix completion and denoising under poisson noise. *Information and Inference: A Journal of the IMA*, 10(2):697–720, 2021. (Cited on page 31.)

[43] Pierre Menard, Omar Darwiche Domingues, Anders Jonsson, Emilie Kaufmann, Edouard Leurent, and Michal Valko. Fast active learning for pure exploration in reinforcement learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7599–7608. PMLR, 18–24 Jul 2021. (Cited on page 8.)

[44] Dipendra Misra, Mikael Henaff, Akshay Krishnamurthy, and John Langford. Kinematic State Abstraction and Provably Efficient Rich-Observation Reinforcement Learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6961–6971. PMLR, 13–18 Jul 2020. (Cited on page 10.)

[45] Michael Mitzenmacher and Eli Upfal. *Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis*. Cambridge university press, 2017. (Cited on page 22.)

[46] Aditya Modi, Jinglin Chen, Akshay Krishnamurthy, Nan Jiang, and Alekh Agarwal. Model-free representation learning and exploration in low-rank mdps. *To appear in Journal of Machine Learning Research (JMLR)*, 2023. (Cited on pages 1 and 10.)

[47] Sahand Negahban and Martin J Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research*, 13(1):1665–1697, 2012. (Cited on page 2.)

[48] Benjamin Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12(12), 2011. (Cited on page 3.)

[49] Tongzheng Ren, Tianjun Zhang, Lisa Lee, Joseph E Gonzalez, Dale Schuurmans, and Bo Dai. Spectral decomposition representation for reinforcement learning. In *Proc. of ICLR*, 2023. (Cited on pages 1 and 10.)

[50] Tyler Sam, Yudong Chen, and Christina Lee Yu. Overcoming the long horizon barrier for sample-efficient reinforcement learning with latent low-rank structure. *ACM SIGMETRICS Performance Evaluation Review*, 50(4):41–43, 2023. (Cited on pages 9 and 10.)

[51] Devavrat Shah, Dogyoon Song, Zhi Xu, and Yuzhe Yang. Sample Efficient Reinforcement Learning via Low-Rank Matrix Estimation. In *Advances in Neural Information Processing Systems*, volume 33, pages 12092–12103. Curran Associates, Inc., 2020. (Cited on pages 9 and 10.)

[52] Devavrat Shah, Dogyoon Song, Zhi Xu, and Yuzhe Yang. Sample efficient reinforcement learning via low-rank matrix estimation. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20. Curran Associates Inc., 2020. (Cited on page 10.)

[53] Ohad Shamir and Shai Shalev-Shwartz. Matrix completion with the trace norm: Learning, bounding, and transducing. *The Journal of Machine Learning Research*, 15(1):3401–3423, 2014. (Cited on page 2.)

[54] Nathan Srebro and Adi Shraibman. Rank, trace-norm and max-norm. In *International conference on computational learning theory*, pages 545–560. Springer, 2005. (Cited on page 2.)

[55] Adam Stooke, Kimin Lee, Pieter Abbeel, and Michael Laskin. Decoupling Representation Learning from Reinforcement Learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 9870–9879. PMLR, 18–24 Jul 2021. (Cited on page 1.)

[56] Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-based RL in Contextual Decision Processes: PAC bounds and Exponential Improvements over Model-free Approaches. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2898–2933. PMLR, 25–28 Jun 2019. (Cited on pages 1 and 10.)

[57] Cindy Trinh, Emilie Kaufmann, Claire Vernade, and Richard Combes. Solving bernoulli rank-one bandits with unimodal thompson sampling. In Aryeh Kontorovich and Gergely Neu, editors, *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, volume 117 of *Proceedings of Machine Learning Research*, pages 862–889. PMLR, 08 Feb–11 Feb 2020. (Cited on page 9.)

[58] Joel A Tropp et al. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015. (Cited on page 34.)

[59] Masatoshi Uehara, Xuezhou Zhang, and Wen Sun. Representation learning for online and offline rl in low-rank mdps. *arXiv preprint arXiv:2110.04652*, 2021. (Cited on page 10.)

[60] Masatoshi Uehara, Xuezhou Zhang, and Wen Sun. Representation Learning for Online and Offline RL in Low-rank MDPs. In *International Conference on Learning Representations*, 2022. (Cited on pages 1 and 10.)

[61] Geoffrey Wolfer and Aryeh Kontorovich. Estimating the mixing time of ergodic markov chains. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 3120–3159. PMLR, 25–28 Jun 2019. (Cited on page 5.)

[62] Ming Yin and Yu-Xiang Wang. Optimal Uniform OPE and Model-based Offline Reinforcement Learning in Time-Homogeneous, Reward-Free and Task-Agnostic Settings. In *Advances in Neural Information Processing Systems*, volume 34, pages 12890–12903. Curran Associates, Inc., 2021. (Cited on page 10.)

[63] Anru Zhang and Mengdi Wang. Spectral State Compression of Markov Processes. *IEEE Transactions on Information Theory*, 66(5):3202–3231, 2020. (Cited on pages 6, 9, 18, 21, 44, 49, and 50.)

[64] Tianjun Zhang, Tongzheng Ren, Mengjiao Yang, Joseph Gonzalez, Dale Schuurmans, and Bo Dai. Making Linear MDPs Practical via Contrastive Representation Learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 26447–26466. PMLR, 17–23 Jul 2022. (Cited on page 10.)

[65] Xuezhou Zhang, Yuda Song, Masatoshi Uehara, Mengdi Wang, Alekh Agarwal, and Wen Sun. Efficient Reinforcement Learning in Block MDPs: A Model-free Representation Learning Approach. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 26517–26547. PMLR, 17–23 Jul 2022. (Cited on page 10.)

[66] Zihan Zhang, Simon Du, and Xiangyang Ji. Near Optimal Reward-Free Reinforcement Learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12402–12412. PMLR, 18–24 Jul 2021. (Cited on page 8.)

[67] Ziwei Zhu, Xudong Li, Mengdi Wang, and Anru Zhang. Learning markov models via low-rank optimization. *Operations Research*, 70(4):2384–2398, 2022. (Cited on page 9.)

Contents

A Preliminaries

In this section, we present a few results that are used throughout our analysis.

A.1 Matrix norms

Lemma 9. *Let $A \in \mathbb{R}^{n\times m}, B \in \mathbb{R}^{m\times r}$. Then:*

$$\|AB\|_{2\to\infty} \le \|A\|_{1\to\infty}\|B\|_{2\to\infty}, \tag{2}$$

$$\|AB\|_{2\to\infty} \le \|A\|_{2\to\infty}\|B\|, \tag{3}$$

$$\|AB\|_{\infty} \le \|A\|_{2\to\infty}\|B^\top\|_{2\to\infty}. \tag{4}$$

Proof. The proof of the lemma directly follows from Hölder's inequality (see for example Proposition 6.5 in [12]). □

A.2 Mixing time

Lemma 10. *(Lemma 5 in [63]) Let $\tau(\varepsilon)$ be the ε-mixing time of an irreducible Markov chain. Then if $\varepsilon \le \delta < 1/2$,*

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

A.3 Value difference lemmas

The following lemmas are used in Section 5 to prove Theorem 8. Recall the definition of value function of a policy π: $V_R^\pi(x) = \mathbb{E}[\sum_{t=1}^\infty \gamma^{t-1} R(x_t^\pi, \pi_t(x_t^\pi))|x_1^\pi = x]$. The (state, action) value function of π is also defined as: for any state $x \in \mathcal{S}$ and action $a \in \mathcal{A}$,

$$Q_R^\pi(x,a) = R(x,a) + \gamma \mathbb{E}_{x'\sim P(\cdot|x,a)}[V_R^\pi(x')].$$

We denote by $\widehat{Q}_R^\pi$ the (state, action) value function of π in the MDP where P is replaced by its estimate $\widehat{P}$, and let $\widehat{\pi}(R)$ be the optimal policy for this MDP.

Lemma 11. *We have that*

$$\|V_R^{\pi^\star(R)} - V_R^{\widehat{\pi}(R)}\|_\infty \le 2\sup_\pi \|Q_R^\pi - \widehat{Q}_R^\pi\|_\infty$$

Proof. We remove the subscript R to simplify the notation. For any s, we have

$$\begin{aligned} V^{\pi^\star}(s) - V^{\widehat{\pi}}(s) = Q^{\pi^\star}(s,\pi^\star(s)) - Q^{\widehat{\pi}}(s,\widehat{\pi}(s)) &= [Q^{\pi^\star}(s,\pi^\star(s)) - \widehat{Q}^{\pi^\star}(s,\pi^\star(s))] \\ &\quad + [\widehat{Q}^{\widehat{\pi}}(s,\widehat{\pi}(s)) - Q^{\widehat{\pi}}(s,\widehat{\pi}(s))] + [\widehat{Q}^{\pi^\star}(s,\pi^\star(s)) - \widehat{Q}^{\widehat{\pi}}(s,\widehat{\pi}(s))] \\ &\le 2\sup_\pi \|Q^\pi - \widehat{Q}^\pi\|_\infty, \end{aligned}$$

since $\widehat{Q}^{\pi^\star}(s,\pi^\star(s)) \le \widehat{Q}^{\widehat{\pi}}(s,\widehat{\pi}(s))$ by definition of $\widehat{\pi}$. □

Lemma 12. *(Proposition 2.1 in [3]) For all policies π:*

$$\|Q_R^\pi - \widehat{Q}_R^\pi\|_\infty \le \frac{\gamma}{(1-\gamma)^2}\max_{a\in\mathcal{A}}\|P^a - \widehat{P}^a\|_{1\to\infty}$$

Combining the two lemmas, we get:

$$\|V_R^{\pi^\star(R)} - V_R^{\widehat{\pi}(R)}\|_\infty \le \frac{2\gamma}{(1-\gamma)^2}\max_{a\in\mathcal{A}}\|P^a - \widehat{P}^a\|_{1\to\infty}.$$

This inequality is used in the proof of Theorem 8.

B Statement and proofs of the main results

In this appendix, we present the proofs of the main theorems. In Subsection §B.1, we provide the proof of Theorem 1 and Corollary 2. In Subsection §B.2, we give a complete, non-simplified version of Theorem 3 from which one can deduce Theorem 3 and Corollary 4 given in the main text. Finally, in Subsection §B.3, we present a complete, non-simplified version of Theorem 5 and from the latter, deduce Theorem 5 and Corollary 6.

B.1 Reward matrix estimation – Model I

In this subsection, we present the proofs of Theorem 1. The proof of Corollary 2 is in fact immediate from Theorem 1.

Proof of Theorem 1. **Proof of (i)**. Recall the results from Lemma 30: for all $\delta \in (0,1)$, if

$$\mathcal{B} = \sqrt{\frac{nm}{T}} \left(\sqrt{(n+m)\log\left(\frac{e(n+m)T}{\delta}\right)} + \log^{3/2}\left(\frac{e(n+m)T}{\delta}\right) \right), \tag{5}$$

then for all $T \geq c_1(\mu^4\kappa^2 r^2 + 1)(m+n)\log^3\left(e^2(m+n)T/\delta\right)$, the event

$$\max(\|U - \widehat{U}(\widehat{U}^\top U)\|, \|V - \widehat{V}(\widehat{V}^\top V)\|) \leq C_1 \frac{\|M\|\|M\|_\infty}{\sigma_r(M)^2} \max(\|V\|_{2\to\infty}\|U\|_{2\to\infty})\mathcal{B}$$

holds with probability at least $1-\delta$ for some universal constants $c_1, C_1 > 0$. To obtain the form presented in Theorem 1, we simply recall the definitions $\kappa = \|M\|/\sigma_r(M)$, $\mu = \max(\sqrt{m/r}\|U\|_{2\to\infty}, \sqrt{n/r}\|V\|_{2\to\infty})$ and the bound $\|M\|_\infty/\sigma_r(M) \leq (\mu^2\kappa r)/\sqrt{mn}$ from Lemma 17. We then substitute in the upper bound above. Note that μ, κ and r are larger than 1 by definition.

Proof of (ii). To establish the desired bound, we use the decomposition error established in Lemma 34. Namely, under the event that $\|\widetilde{M} - M\| \leq c_1\sigma_r(M)$ for some universal constant $c_1 > 0$ sufficiently small, there exists a universal constant $c_2 > 0$ such that

$$\|\widehat{M} - M\|_{2\to\infty} \leq c_2\sigma_1(M)\left[\|U - \widehat{U}(\widehat{U}^\top U)\|_{2\to\infty} + \mu\sqrt{\frac{r}{m}}\frac{\|\widetilde{M} - M\|}{\sigma_r(M)}\right]. \tag{6}$$

Hence, we only need high probability bounds on $\|U - \widehat{U}(\widehat{U}^\top U)\|_{2\to\infty}$ which we established in (i), and on $\|\widetilde{M} - M\|$ which we also established in Proposition 26 under the compound Poisson entries model described (15). We can extend the latter result under our observation model using the Poisson approximation Lemma 20, and finally write that for all $\delta \in (0,1)$, using the same definition of $\mathcal{B}$ as above in (5), for all for all $T \geq c_3 \log^3((n+m)/\delta)$, the following statement

$$\frac{\|\widetilde{M} - M\|}{\sigma_r(M)} \leq \frac{C_3\|M\|_\infty}{\sigma_r(M)}\mathcal{B} \tag{7}$$

holds with probability at least $1-\delta$, for some universal constants $c_3, C_3 > 0$ large enough. Note that under the condition $T \geq c_4\mu^4\kappa^2 r^2 \log^3(e(n+m)/\delta)$ for some universal constant c_4 large enough, the high probability statement in (7) holds and in addition we also have $\|\widetilde{M} - M\| \leq c_1\sigma_r(M)$. There, we used the result of Lemma 17. The statement (ii) in Theorem 1 is obtained by first substituting in (6), the upper bound we get in (i) and that we get in (7), and then, using $\sigma_1(M) \leq \sqrt{mn}\|M\|_\infty$ and the bound $\|M\|_\infty/\sigma_r(M) \leq (\mu^2\kappa r)/\sqrt{mn}$ from Lemma 17.

Proof of (iii). To establish the desired bound, we use the decomposition error established in Lemma 36. Namely, under the event that $\|\widetilde{M} - M\| \leq c_1\sigma_r(M)$ for some universal constant $c_1 > 0$ sufficiently small, there exists a universal constant $c_2 > 0$ such that

$$\|\widehat{M} - M\|_\infty \leq c_2\|M\|_{2\to\infty}\left(\frac{\|M - \widetilde{M}\|}{\sigma_r(M)}\|V\|_{2\to\infty} + \|V - \widehat{V}W_{\widehat{V}}\|_{2\to\infty}\right) + c_2\|M - \widehat{M}\|_{2\to\infty}(\|V\|_{2\to\infty} + \|V - \widehat{V}W_{\widehat{V}}\|_{2\to\infty}). \tag{8}$$

To upper bound the above error, we need to control: (a) $\|V - \widehat{V} W_{\widehat{V}}\|_{2\to\infty}$, which we have already done in (i); (b) $\|M - \widetilde{M}\|$, which follows from Lemma 20 as established in the proof of (ii) (see the high probability statement (7)); and (c) $\|M - \widehat{M}\|_{2\to\infty}$, which again we have already done in (ii). The statement (iii) in Theorem 1 follows from first substituting in (8), the upper bounds we get from (a), (b) and (c), and then using $\|M\|_\infty/\sigma_r(M) \le (\mu^2\kappa r)/\sqrt{mn}$, $\mu = \max(\sqrt{m/r}\|U\|_{2\to\infty}, \sqrt{n/r}\|V\|_{2\to\infty})$, and the basic inequality $\|M\|_{2\to\infty} \le \sqrt{m}\|M\|_\infty \le \sqrt{m+n}\|M\|_\infty$. □

B.2 Transition matrix estimation under the generative model – Model II(a)

In this subsection, we present a complete, non-simplified version of Theorem 3, from which one can deduce Theorem 3 and Corollary 4 given in the main text.

First, let us define the function $g_\delta : \mathbb{R}^{n\times n} \to \mathbb{R}_+$ as

$$\begin{aligned} g_\delta(M) = \mathbb{1}_{\{\exists \ell : \|M_{\ell,:}\|_\infty \le 1\}} \log\left(\frac{ne}{\delta}\right) \log^{-1}\left(1 + \frac{1}{\|M\|_\infty}\right) \\ + \mathbb{1}_{\{\forall \ell : \|M_{\ell,:}\|_\infty > 1\}} \log\left(\frac{\|M\|_\infty\, ne}{\delta}\right) \sqrt{\|M\|_\infty}. \end{aligned} \tag{9}$$

We also use the following notation:

$$\begin{cases} \mathcal{A} = \frac{1}{\sqrt{T}}\sqrt{\|M\|_{1\to\infty} + \|M^\top\|_{1\to\infty}}, \\ \mathcal{B}' = \mu\kappa\sqrt{\frac{r}{n}}\left(\mathcal{A} + \frac{1}{T} g_{\delta/\sqrt{T}}(TM)\log\left(\frac{n\sqrt{T}}{\delta}\right)\right) + \sqrt{\frac{r\|M\|_\infty}{T}\log\left(\frac{n\sqrt{T}}{\delta}\right)}. \end{cases}$$

We first recall a standard result quantifying how well $\widetilde{M}$ approximates M.

Lemma 13. *$\forall \delta \in (0,1)$, w.p. at least $1-\delta$, $\|\widetilde{M} - M\| \le C\mathcal{A} + \frac{C}{T} g_{\delta/\sqrt{T}}(TM)\sqrt{\log(\frac{n\sqrt{T}}{\delta})}$.*

Proof. The lemma follows directly from Lemma 22 (replacing T_τ by T) and Lemma 28. □

Theorem 14. *Assume that $(\nu_0)_{\min} = \min_{i\in[n]}(\nu_0)_i > 0$. For any $\delta > 0$, if $\|\widetilde{M} - M\| \le c\sigma_r(M)$, $g_{\delta/\sqrt{T}}(TM)\log(n\sqrt{T}/\delta) \le cT\sigma_r(M)$ and $\|M\|_\infty \log(n\sqrt{T}/\delta) \le cT\sigma_r^2(M)$ for some universal constant $c > 0$, then there exists a universal constant $C > 0$ such that with probability at least $1-\delta$ holds:*

$$\begin{array}{ll} (i) & \max\left\{\|U - \widehat{U}(\widehat{U}^\top U)\|_{2\to\infty}, \|V - \widehat{V}(\widehat{V}^\top V)\|_{2\to\infty}\right\} \le C\frac{\mathcal{B}'}{\sigma_r(M)}, \\ (ii) & \|\widehat{M} - M\|_{2\to\infty} \le C\kappa\mathcal{B}', \quad \|\widehat{P} - P\|_{1\to\infty} \le C\frac{\kappa\sqrt{n}}{(\nu_0)_{\min}}\mathcal{B}', \\ (iii) & \|\widehat{M} - M\|_\infty \le C\left(\frac{\|M\|_{2\to\infty} + \kappa\mathcal{B}'}{\sigma_r(M)} + \kappa\mu\sqrt{\frac{r}{n}}\right)\mathcal{B}', \\ (iv) & \|\widehat{P} - P\|_\infty \le C\frac{\mathcal{B}'}{(\nu_0)_{\min}}\left[\sqrt{n}\kappa\frac{\|M\|_\infty}{(\nu_0)_{\min}} + \frac{\|M\|_{2\to\infty} + \kappa\mathcal{B}'}{\sigma_r(M)} + \kappa\mu\sqrt{\frac{r}{n}}\right], \end{array}$$

where (iv) holds if in addition $\|\widehat{M} - M\|_{1\to\infty} \le \frac{1}{2}(\nu_0)_{\min}$.

Proof. The first statement of the theorem follows from Lemma 22 (with T instead of T_τ), Lemmas 28 and 32. The remaining bounds are consequences of (i) and of the results presented in Appendix F. □

Proof of Theorem 3. Theorem 3 follows from Lemma 13 and Theorem 14 by simplifying the term $\mathcal{B}'$ using $\mathcal{B}$ given in Theorem 3. As a result of this simplification, as well as of the assumptions given in statement of Theorem 14, we obtain bounds on n, T required in Theorem 3. Furthermore, we use simple inequalities (check Lemma 17) to rewrite all terms depending on M as functions of $\|M\|_\infty$ and $(\nu_0)_{\min}$. □

Remark 1. *It is worth noting that Corollary 4 is a corollary of Theorem 14, and that the lower bound on n required in Theorem 3 is not required for this corollary. Moreover, results presented in this corollary are valid for almost all $T \ge cn\log(nT)$ - in the case when $T \asymp [n^{2-\epsilon}, n^2]$ for arbitrarily* *small $\epsilon > 0$, bounds in Corollary 4 contain additional* log *term, which is an artifact of our analysis (and splitting concentration into cases $T \lesssim n^2$ and $T \gtrsim n^2$). This discontinuity in the range of T can be resolved, but at the price of a reduced readability.*

B.3 Transition matrix estimation under the forward model – Model II(b)

In this subsection, we present a complete, non-simplified version of Theorem 5, from which one can deduce Theorem 5 and Corollary 6 given in the main text.

Again, we use function the funcion g_δ defined in (9), and we introduce:

$$\mathcal{B}' = \mu\kappa\sqrt{\frac{r}{n}}\left(\sqrt{\frac{\|\nu\|_\infty\tau^\star}{T}\log\left(\frac{ne}{\delta}\right)\log(T\nu_{\min}^{-1})} + \frac{\tau^\star}{T}g_{\delta/\sqrt{T_\tau}}(T_\tau M)\log\left(\frac{n\sqrt{T_\tau}}{\delta}\right)\log(T\nu_{\min}^{-1})\right) \\ + \sqrt{\frac{r\tau^\star\|M\|_\infty}{T}\log\left(\frac{n\sqrt{T_\tau}}{\delta}\right)\log(T\nu_{\min}^{-1})}.$$

Our analysis starts from the following lemma stating how well $\widetilde{M}$ approximates M.

Lemma 15. *(Lemma 7 in [63]) For $\tau \geq 2\tau^\star\log(T\nu_{\min}^{-1})$ and for any $\delta \in (0,1)$, we have with probability at least $1-\delta$: $\|\widetilde{M} - M\| \leq C\sqrt{\frac{\|\nu\|_\infty\tau}{T}\log\left(\frac{ne}{\delta}\right)} + C\frac{\tau}{T}\log\left(\frac{ne}{\delta}\right)$.*

Theorem 16. *Assume that $\nu_{\min} = \min_{i\in[n]}\nu_i > 0$ and that $\tau/(\tau^\star\log(T\nu_{\min}^{-1})) \in [c_1, c_2]$ for some universal constants $c_2 > c_1 \geq 2$. For any $\delta > 0$, if $\|\widetilde{M} - M\| \leq c\sigma_r(M)$, $g_{\delta/\sqrt{T}}(T_\tau M)\log(n\sqrt{T_\tau}/\delta) \leq cT_\tau\sigma_r(M)$ and $\|M\|_\infty\log(n\sqrt{T_\tau}/\delta) \leq cT_\tau\sigma_r^2(M)$ for some universal constant $c > 0$, then there exists a universal constant $C > 0$ such that with probability at least $1-\delta$,*

$$\begin{aligned}
&(i) && \max\left\{\|U - \widehat{U}(\widehat{U}^\top U)\|_{2\to\infty}, \|V - \widehat{V}(\widehat{V}^\top V)\|_{2\to\infty}\right\} \leq C\tfrac{\mathcal{B}'}{\sigma_r(M)},\\
&(ii) && \|\widehat{M} - M\|_{2\to\infty} \leq C\kappa\mathcal{B}', \quad \|\widehat{P} - P\|_{1\to\infty} \leq C\tfrac{\kappa\sqrt{n}}{\nu_{\min}}\mathcal{B}',\\
&(iii) && \|\widehat{M} - M\|_\infty \leq C\left(\tfrac{\|M\|_{2\to\infty}+\kappa\mathcal{B}'}{\sigma_r(M)} + \kappa\mu\sqrt{\tfrac{r}{n}}\right)\mathcal{B}',\\
&(iv) && \|\widehat{P} - P\|_\infty \leq C\tfrac{\mathcal{B}'}{\nu_{\min}}\left[\sqrt{n}\kappa\tfrac{\|M\|_\infty}{\nu_{\min}} + \tfrac{\|M\|_{2\to\infty}+\kappa\mathcal{B}'}{\sigma_r(M)} + \kappa\mu\sqrt{\tfrac{r}{n}}\right],
\end{aligned}$$

where (iv) holds if in addition $\|\widehat{M} - M\|_{1\to\infty} \leq \frac{1}{2}\nu_{\min}$.

Proof. The first statement of the theorem follows from Lemmas 21, 22 and 32, whereas the next four bounds follow from (i) and the bounds presented in Appendix F. □

As for the generative model, Theorem 5 is a direct consequence of Theorem 16, and it is obtained by simplifying the term $\mathcal{B}'$ to $\mathcal{B}$. Corollary 6 is also easily derived from Theorem 16.

B.4 An additional lemma

Lemma 17. *Let M be matrix and $m \times n$ matrix with rank r, incoherence parameter $\mu > 0$, and condition number $\kappa > 0$. Then, we have*

$$\|M\|_\infty \leq \sigma_1(M)\frac{\mu^2 r}{\sqrt{nm}} \leq \sigma_r(M)\frac{\mu^2\kappa r}{\sqrt{nm}}.$$

Proof of Lemma 17. For all $(i,j) \in [m]\times[n]$, we have

$$\begin{aligned}
|M_{i,j}| &= \left|\sum_{\ell=1}^r \sigma_\ell(M)u_{i,\ell}v_{j,\ell}\right| \leq \sigma_1(M)\sum_{\ell=1}^r |u_{i,\ell}v_{j,\ell}|\\
&\leq \sigma_1(M)\|U\|_{2\to\infty}\|V\|_{2\to\infty} \leq \sigma_1(M)\frac{\mu r}{\sqrt{nm}} \leq \sigma_r(M)\frac{\mu\kappa r}{\sqrt{nm}}.
\end{aligned}$$

The first inequality follows from the triangular inequality and the fact that $\sigma_1(M) \geq \sigma_2(M) \geq \cdots \geq \sigma_r(M)$. The second inequality follows from Cauchy-Schwarz inequality. The last inequalities follow by definition of the incoherence parameter and of the condition number. □

C Comparison inequalities and the Poisson approximation argument

In this section, we state and prove the results related to the Poisson approximations used to handle the noise correlations in the data. We start by presenting some of the key tools behind the Poisson approximation argument. This argument comes in the form of comparison inequalities. The latter are applied and specified first to Model I (reward matrix estimation), and then to Model II (transition matrix estimation).

C.1 Preliminaries on Poisson approximation

The Poisson approximation argument comes in the form of an inequality, which is presented in Lemma 19. However, the key idea behind the argument is, roughly speaking, the equality in distribution between a multinomial distribution with t trials and n outcomes, and the joint distribution of n independent Poisson random variables with properly chosen parameters, conditioned on some particular event. This equality of distribution is powerful for our purposes precisely because of the independence between the Poisson random variables. Below, we present Lemma 18 that represents this idea.

Lemma 18. *(Heterogeneous analogue of Theorem 5.2 in [45]) Let* $Y_i^{(t)} \sim \mathrm{Poisson}(tp_i)$, $i = 1, \dots, n$, *be independent random variables with* $\sum_{i=1}^n p_i = 1$. *Moreover, let* $(Z_1^{(t)}, Z_2^{(t)}, \dots, Z_n^{(t)}) \sim \mathrm{Multinomial}(t, (p_1, \dots, p_n))$. *Then distribution of* $(Y_1^{(t)}, \dots, Y_n^{(t)})$ *conditioned on* $\sum_{i=1}^n Y_i^{(t)} = s$ *is the same as* $(Z_1^{(s)}, \dots, Z_n^{(s)})$ *irrespective of t.*

Proof. The proof follows similar steps as the proof of Theorem 5.2 in [45], but we provide it here for the sake of completeness. First, note that from the definition of multinomial distributions:

$$\mathbb{P}((Z_1^{(s)}, \dots, Z_n^{(s)}) = (a_1, \dots, a_n)) = \frac{s!}{a_1! \cdots a_n!} p_1^{a_1} \cdots p_n^{a_n} \tag{10}$$

if $\sum_{i=1}^n a_i = s$, and 0 otherwise. Since the sum of Poisson random variables is a Poisson random variable with parameter equal to the sum of parameters of the initial random variables, we get that the random variable $\sum_{i=1}^n Y_i^{(t)} \sim \mathrm{Poisson}(\sum_{i=1}^n tp_i) = \mathrm{Poisson}(t)$. Hence we have:

$$\begin{aligned}
\mathbb{P}\left((Y_1^{(t)}, \dots, Y_n^{(t)}) = (a_1, \dots, a_n) \middle| \sum_{i=1}^n Y_i^{(t)} = s \right) &= \frac{\mathbb{P}((Y_1^{(t)}, \dots, Y_n^{(t)}) = (a_1, \dots, a_n))}{\mathbb{P}(\sum_{i=1}^n Y_i^{(t)} = s)} \\
&= \frac{s!}{\exp(-t) t^s} \prod_{i=1}^n \frac{(tp_i)^{a_i} \exp(-tp_i)}{a_i!} \\
&= \frac{s!}{a_1! \cdots a_n!} p_1^{a_1} \cdots p_n^{a_n} \qquad (11)
\end{aligned}$$

where in the last step we used the independence of $Y_i^{(t)}$'s and $\sum_{i=1}^n a_i = s$. Note that equations (10) and (11) are exactly the same, which concludes the proof. □

Lemma 19. *Consider the setting of Lemma 18 and let* $f : \mathbb{R}^p \to \mathbb{R}_+$ *be any non-negative function. Then:*

$$\mathbb{E}\left[f(Z_1^{(t)}, \dots, Z_p^{(t)}) \right] \le e\sqrt{t}\,\mathbb{E}\left[f(Y_1^{(t)}, \dots, Y_p^{(t)}) \right]$$

Proof. The proof is essentially the same as that of Theorem 5.7 in [45] with the exception that we use Lemma 18 instead of Theorem 5.6 in [45] and we repeat it here for the sake of completeness.

$$\begin{aligned}
\mathbb{E}[f(Y_1^{(t)}, \dots, Y_p^{(t)})] &= \sum_{k=0}^{\infty} \mathbb{E}\left[f(Y_1^{(t)}, \dots, Y_p^{(t)}) \middle| \sum_{i=1}^p Y_i^{(t)} = k \right] \mathbb{P}\left(\sum_{i=1}^p Y_i^{(t)} = k \right) \\
&\ge \mathbb{E}\left[f(Y_1^{(t)}, \dots, Y_p^{(t)}) \middle| \sum_{i=1}^p Y_i^{(t)} = t \right] \mathbb{P}\left(\sum_{i=1}^p Y_i^{(t)} = t \right) \\
&= \mathbb{E}[f(Z_1^{(t)}, \dots, Z_p^{(t)})] \mathbb{P}\left(\sum_{i=1}^p Y_i^{(t)} = t \right) \qquad (12)
\end{aligned}$$

where in the second line we used non-negativeness of f, and in the last line we used Lemma 18. Now, since $\sum_{i=1}^{p} Y_i^{(t)}$ is a Poisson random variable with mean t we have $\mathbb{P}(\sum_{i=1}^{p} Y_i^{(t)} = t) = \frac{t^t \exp(-t)}{t!}$ and using simple inequality $t! < e\sqrt{t}(\frac{t}{e})^t$ we can rewrite inequality (12) as follows:

$$\mathbb{E}[f(Y_1^{(t)}, \dots, Y_p^{(t)})] \geq \mathbb{E}[f(Z_1^{(t)}, \dots, Z_p^{(t)})] \frac{1}{e\sqrt{t}} \tag{13}$$

which gives statement of the lemma. □

C.2 Poisson approximation for reward matrices – Model I

We recall from Section 4 that the definition of the empirical reward matrix $\widetilde{M}$ is given as follows

$$\forall (i,j) \in [n] \times [m], \qquad \widetilde{M}_{i,j} = \frac{nm}{T} \sum_{t=1}^{T} (M_{i_t, j_t} + \xi_t) \mathbb{1}_{\{(i_t, j_t) = (i,j)\}} \tag{14}$$

where (i_t, j_t) are sampled uniformly at random from $[n] \times [m]$. Due to independence between $(i_1, j_1), \dots, (i_T, j_T)$ and $\xi_1, \dots, \xi_T$, we note that the observation model (14) is equivalent in distribution to the following one

$$\forall (i,j) \in [n] \times [m], \qquad \widetilde{M}_{i,j} = \frac{nm}{T} \sum_{t=1}^{Z_{i,j}} (M_{i,j} + \xi'_{i,j,t})$$

where we for all $(i,j) \in [n] \times [j]$, $(\xi'_{i,j,t})_{t \geq 1}$ is a sequence of i.i.d. random variables copies, say of ξ_1, and

$$Z_{i,j} = \sum_{t=1}^{T} \mathbb{1}_{\{(i_t, j_t) = (i,j)\}}.$$

Observe that $Z = (Z_{i,j})_{(i,j)}$ is a multinomial random variable whose parameters are defined by the fact that for all $t \in [T]$, $\mathbb{P}((i_t, j_t) = (i,j)) = 1/nm)$. We denote $\mathbb{P}$ the joint probability of the entries of Z and sequences $(\xi_{i,j,t})_{t \geq 1}$, $(i,j) \in [n] \times [m]$.

Compound Poisson random matrix model. We define a random matrix $Y \in \mathbb{R}^{n \times m}$ generated by a Poisson model as follows:

$$Y_{i,j} \sim \text{Poisson}\,(T/nm)\,, \qquad (i,j) \in [n] \times [m]$$

and denote $\mathbb{P}'$ the joint probability of the entries of Y and the sequences $(\xi'_{i,j,t})_{t \geq 1}$, for $(i,j) \in [n] \times [m]$. We may then consider the matrix model

$$X_{i,j} = \sum_{t=1}^{Y_{i,j}} (M_{i,j} + \xi'_{i,j,t}). \tag{15}$$

We note that the entries of the matrix X are distributed according to compound Poisson distributions. Below, we precise the Poisson approximation argument for the reward matrix model.

Lemma 20 (Poisson Approximation)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ (resp. $(\Omega, \mathcal{F}, \mathbb{P}')$) be the probability space under the matrix-plus-noise model* (14) *(resp.* (15)*). Then for any event $\mathcal{E} \in \mathcal{F}$, we have*

$$\mathbb{P}\left(\mathcal{E}\right) \leq e\sqrt{T}\,\mathbb{P}'\left(\mathcal{E}\right).$$

Proof of Lemma 20. For convenience, we denote $X = ((\xi_{i,j,t})_{t \geq 1})_{i,j \in [n] \times [m]}$. We set $f(Z, X) = \mathbb{1}_{\{\mathcal{E}\}}$. Thanks to Lemma 19, given that Z is independent of X, we have

$$\mathbb{E}\left[f(Z,X) | X\right] \leq e\sqrt{T}\mathbb{E}\left[f(Y,X)|X\right].$$

We further take the expectation on X and write

$$\mathbb{P}(\mathcal{E}) = \mathbb{E}\left[f(Z,X)\right] \leq e\sqrt{T}\mathbb{E}\left[f(Y,X)\right] = e\sqrt{T}\,\mathbb{P}'(\mathcal{E}).$$

□

C.3 Approximations for transition matrices – Model II

We restrict our attention to the forward model, Model II(b). The results for the generative model are simpler and can be easily deduced from those for the forward model. Recall from Section 3.3 definition of matrix $\widetilde{M}^{(k)}$ and in the following discussion we fix value of $k \in [\tau]$. Define a matrix $N = T_\tau \widetilde{M}^{(k)}$ and note that it is equal to:

$$N_{i,j} = \sum_{l=0}^{T_\tau - 1} \mathbb{1}_{\{(x_{k+l\tau}, x_{k+1+l\tau})=(i,j)\}}, \qquad i,j = 1,2,\dots,n \tag{16}$$

Furthermore, let $\mathbb{P}_1$ be joint probability distribution of entries of N.

C.3.1 Multinomial approximation

Here we define a matrix $Z \in \mathbb{R}^{n\times n}$ with entries:

$$Z_{i,j} = \sum_{t=0}^{T_\tau - 1} \mathbb{1}_{\{(i_t, j_t)=(i,j)\}}, \qquad i,j = 1,2,\dots,n, \tag{17}$$

where $\mathbb{P}((i_t, j_t) = (i,j)) = \nu_i P_{i,j}$ independently over $i, j \in [n]$ and $t \in [T_\tau]$. Denote by $\mathbb{P}_2$ joint probability distribution of entries of Z. Then we have:

Lemma 21. *Let N and Z be matrices obtained under the models* (16) *and* (17)*, respectively. Then, for any subset $\mathcal{Z}$ of $\{z \in \mathbb{N}^{n^2} : \sum_{(i,j)} z_{i,j} = T_\tau\}$, we have $\mathbb{P}(N \in \mathcal{Z}) \le 3\mathbb{P}(Z \in \mathcal{Z})$.*

Proof. Note that by subsampling as explained in Section 3.3, for each k we obtain a Markov chain with transition kernel

$$P_\tau((y,y')|(x,x')) = P^\tau(y|x')P(y'|y)$$

and initial distribution $\nu_0^{(k)}(x,x') = \nu_0^{(k)}(x)P(x'|x)$ with

$$\nu_0^{(1)}(x) = \nu_0(x) \qquad \text{and} \qquad \nu_0^{(k)}(x) = \sum_{y\in[n]} \nu_0(y)P^{k-1}(x|y) \quad \text{for } k = 2,\dots,\tau.$$

Moreover, all chains share the same stationary distribution given by $\Pi \in \mathbb{R}^{n\times n}$ with $\Pi_{x,x'} = \nu(x)P(x'|x)$, $x, x' \in [n]$. Now, recall definition of τ from Theorem 5 and note that according to Lemma 10 with $\delta = \frac{1}{4}$ and $\varepsilon = \nu_{\min}/(eT)$ we have $\tau(\varepsilon) \le \tau$ and thus:

$$\max_{1\le i\le n} \|P^\tau_{i,:} - \nu^\top\|_1 \le \frac{\nu_{\min}}{eT}. \tag{18}$$

Now, let $z = (z_{i,j})_{i,j=1}^n \in \{z \in \mathbb{N}^{n^2} : \sum_{i,j} z_{i,j} = T_\tau\}$ be a tuple of fixed integers. Define a set:

$$\mathcal{S}(z) := \{(a_{2l+1}, a_{2l+2})_{l=0}^{T_\tau-1} \in ([n]\times[n])^{T_\tau} : \sum_{l=0}^{T_\tau-1} \mathbb{1}_{\{(a_{2l+1},a_{2l+2})=(i,j)\}} = z_{i,j}, \forall i,j \in [n]\}$$

and note that $|\mathcal{S}(z)| = T_\tau!(\prod_{i,j=1}^n z_{i,j}!)^{-1}$. By definition of Markovian and multinomial models, we have:

$$\mathbb{P}(N = z) = \sum \nu_0^{(k)}(x_{k-1}, x_k) \prod_{l=1}^{T_\tau - 1} P_\tau((x_{k-1+l\tau}, x_{k+l\tau})|(x_{k-1+(l-1)\tau}, x_{k+(l-1)\tau}))$$

where the sum is over $(x_{k-1+l\tau}, x_{k+l\tau})_{l=0}^{T_\tau-1} \in \mathcal{S}(z)$, and

$$\mathbb{P}(Z = z) = \frac{T_\tau!}{\prod_{i,j=1}^n z_{i,j}!} \prod_{i,j=1}^n \Pi_{i,j}^{z_{i,j}}.$$

Now we fix arbitrarily one of the summands in the expression for $\mathbb{P}(N = z)$ and note that:

$$\begin{aligned}
&\left|\nu_0^{(k)}(x_{k-1}, x_k) \prod_{l=1}^{T_\tau - 1} P_\tau((x_{k-1+l\tau}, x_{k+l\tau})|(x_{k-1+(l-1)\tau}, x_{k+(l-1)\tau})) - \prod_{i,j=1}^{n} \Pi_{i,j}^{z_{i,j}}\right| \\
&= \left(\prod_{l=0}^{T_\tau - 1} P(x_{k+l\tau}|x_{k-1+l\tau})\right) \left|\nu_0^{(k)}(x_{k-1}) \prod_{l=1}^{T_\tau - 1} P^\tau(x_{k-1+l\tau}|x_{k+(l-1)\tau}) - \prod_{l=0}^{T_\tau - 1} \nu(x_{k-1+l\tau})\right| \\
&\leq \left(\prod_{l=0}^{T_\tau - 1} P(x_{k+l\tau}|x_{k-1+l\tau})\right) \left(\prod_{l=0}^{T_\tau - 1} (\nu(x_{k-1+l\tau}) + \epsilon) - \prod_{l=0}^{T_\tau - 1} \nu(x_{k-1+l\tau})\right) \\
&\leq \left(\prod_{i,j=1}^{n} \Pi_{i,j}^{z_{i,j}}\right) \sum_{j=1}^{T_\tau} \left(\frac{\epsilon}{\nu_{\min}}\right)^j \binom{T_\tau}{j} \leq \left(\prod_{i,j=1}^{n} \Pi_{i,j}^{z_{i,j}}\right) \sum_{j=1}^{T_\tau} \left(\frac{eT_\tau \epsilon}{j\nu_{\min}}\right)^j \leq 2\left(\prod_{i,j=1}^{n} \Pi_{i,j}^{z_{i,j}}\right)
\end{aligned}$$

where in first inequality we used Equation (18), where we then used the bound on binomial coefficients $\binom{T_\tau}{j} \leq (eT_\tau/j)^j$, and where in the last inequality, we used definition of ϵ. Since this upper bound holds irrespective of the summand, we deduce that:

$$|\mathbb{P}(N = z) - \mathbb{P}(Z = z)| \leq 2\frac{T_\tau!}{\prod_{i,j=1}^{n} z_{i,j}!} \left(\prod_{i,j=1}^{n} \Pi_{i,j}^{z_{i,j}}\right) = 2\mathbb{P}(Z = z).$$

Now, let $\mathcal{Z}$ be any subset of $\{z \in \mathbb{N}^{n^2} : \sum_{(i,j)} z_{i,j} = T_\tau\}$. Then we have:

$$\mathbb{P}(N \in \mathcal{Z}) = \sum_{z \in \mathcal{Z}} \mathbb{P}(N = z) \leq 3 \sum_{z \in \mathcal{Z}} \mathbb{P}(Z = z) = 3\mathbb{P}(Z \in \mathcal{Z})$$

as claimed in the lemma.

□

C.3.2 Poisson approximation

We define a matrix $Y \in \mathbb{R}^{n \times n}$ generated by the Poisson model as follows:

$$Y_{i,j} \sim \text{Poisson}(T_\tau M_{i,j}), \qquad i, j = 1, 2, \dots, n. \tag{19}$$

We show that rare random events occur with approximately equal probability for the Poisson and multinomial models:

Lemma 22. *Let Z and Y be matrices obtained under the models* (17) *and* (19)*, respectively. Then for any $\mathcal{Z} \subset \mathbb{N}^{n^2}$, we have $\mathbb{P}(Z \in \mathcal{Z}) \leq e\sqrt{T_\tau}\mathbb{P}(Y \in \mathcal{Z})$.*

Proof. Proof of the lemma is a straightforward consequence of Lemma 19 with parameters T_τ, n^2 and $f = \mathbb{1}_{\{\mathcal{Z}\}}$. □

D Concentration of matrices with Poisson and compound Poisson entries

As mentioned in Appendix C, our analysis relies on a Poisson approximation argument. As a result, we will require tight concentration bounds for random matrices with entries distributed according to compound Poisson distributions (when estimating the reward matrix) and Poisson distributions (when estimating the transition matrices). In §D.1, we present a few simple facts about Poisson and compound Poisson random variables, together with some other useful tools. In §D.2, we present two concentration results, required for the model with compound Poisson entries. Similarly, in §D.3, we present two concentration results, required for the model with Poisson entries. These concentration results will be extensively used in the forthcoming analysis for the subspace recovery.

It is worth noting that our results in §D.3 are sharper than those in §D.2 thanks to Bennett's inequality. As a consequence, our results for estimating the reward matrix exhibit a dependence in $\log^3(n+m)$ while in the estimation of the transitions, our results exhibit a dependence in $\log^2(n)$ and even $\log(n)$ in some regimes.

D.1 Preliminaries

We first present Theorem 23, which can be seen as a version of matrix Bernstein inequality. The theorem is borrowed from [26] and relies on a truncation trick. The proofs of our concentration results in §D.2 and §D.3 rely on this theorem.

Theorem 23. *(Proposition A.3 in [26]) Let $\{Z_t\}_{t=1}^T$ be a sequence of $m \times n$ independent zero-mean real random matrices. Suppose that for all $1 \le t \le T$,*

$$(i)\ \ \mathbb{P}\left(\|Z_t\| \ge \beta\right) \le p, \qquad \textit{and} \qquad (ii)\ \ \left\|\mathbb{E}[Z_t \mathbb{1}_{\{\|Z_t\| > \beta\}}]\right\| \le q, \tag{20}$$

hold for some quantities $p \in (0,1)$, and $q \ge 0$. Furthermore, assume there exists $v \ge 0$, such that

$$(iii)\ \ \max\left\{\left\|\sum_{t=1}^T \mathbb{E}\left[Z_t Z_t^\top\right]\right\|, \left\|\sum_{t=1}^T \mathbb{E}\left[Z_t^\top Z_t\right]\right\|\right\} \le v. \tag{21}$$

Then, for all $u > 0$,

$$\mathbb{P}\left(\left\|\sum_{t=1}^T Z_t\right\| \ge Tq + u\right) \le Tp + (n+m)\exp\left(-\frac{u^2/2}{v + \beta u/3}\right). \tag{22}$$

To apply Theorem 23, we need control of the tails of the entries of the random matrix we study. In the case of Poisson entries, we will simply use the following standard fact about Poisson random variables. It is a simple consequence of Bennett's inequality [8].

Lemma 24. *Let Y be a Poisson random variable with mean λ. Then for, all $\theta \in \mathbb{R}$, we have $\mathbb{E}[e^{\theta Y}] \le \exp(\lambda(e^\theta - 1))$. Furthermore, we have for all $u > 0$*

$$\mathbb{P}(|Y - \lambda| > u) \le 2\exp\left(-\lambda h(u/\lambda)\right) \le 2\exp\left(-\frac{u^2/2}{\lambda + u/3}\right),$$

where $h(u) = (1+u)\log(1+u) - u$.

In the case of compound Poisson entries, we do not have any result similar to Bennett's inequality. Instead, we derive a Bernstein-type concentration result on these random variables.

Lemma 25. *Let $(\xi_t)_{t\ge1}$ be a sequence of zero-mean, σ^2-subgaussian, i.i.d. random variables. Let Y be a Poisson random variables with mean λ. Let M be a positive constant. Then, the moment generating function of the compound Poisson random variable $Z = \sum_{i=1}^Y (M + \xi_i)$ satisfies the following:*

$$\forall u > 0, \qquad \mathbb{P}(|Z - \lambda M| > u) \le 2\exp\left(-\min\left(\frac{u^2}{16e\lambda L^2}, \frac{u}{4L}\right)\right),$$
$$\mathbb{E}\left[|Z - \lambda M|^2\right] \le 18\lambda L^2,$$

where $L = \max(M, \sigma)$.

Proof of Lemma 25. First, we upper bound the moment generating function of $\sum_{i=10}^{Y}(M+\xi_i)$. Let $\theta > 0$, we have

$$
\begin{aligned}
I_Z(\theta) \triangleq \mathbb{E}\left[e^{\theta(\sum_{i=1}^{Y}(M+\xi_i))}\right] &\leq \sqrt{\mathbb{E}\left[e^{2\theta MY}\right]\mathbb{E}[e^{2\theta\sum_{i=0}^{Y}\xi_i}]} \\
&\leq \exp\left(\frac{\lambda(e^{2\theta M}-1)}{2}\right)\sqrt{\mathbb{E}[e^{2\theta\sum_{i=0}^{Y}\xi_i}]} \\
&\leq \exp\left(\frac{\lambda}{2}\left((2\theta M)^2 e^{2\theta M} + 2\theta M\right)\right)\sqrt{\mathbb{E}[e^{2\theta\sum_{i=0}^{Y}\xi_i}]},
\end{aligned}
$$

where in the first inequality, we use Cauchy-Schwarz inequality, in the second inequality, we use the well known bound on the moment generating function of a Poisson random variable (if Y is a Poisson random variable with mean λ, then for all $\theta > 0$, $\mathbb{E}[e^{\theta Y}] \leq \exp(\lambda(e^{\theta}-1))$), and in the last inequality, we use the elementary fact that $e^x - 1 \leq x^2 e^x + x$ for all $x \in \mathbb{R}$. Next, we have

$$
\begin{aligned}
\mathbb{E}\left[e^{2\theta\sum_{i=1}^{Y}\xi_i}\right] &= \mathbb{E}\left[\sum_{k=1}^{\infty}\mathbb{1}_{\{Y=k\}}\exp\left(2\theta\sum_{i=1}^{k}\xi_i\right)\right] \\
&= \sum_{k=1}^{\infty}\mathbb{P}(Y=k)\mathbb{E}\left[\exp\left(2\theta\sum_{i=1}^{k}\xi_i\right)\right] \\
&\leq \sum_{k=1}^{\infty}\mathbb{P}(Y=k)\exp(2k\theta^2\sigma^2) \\
&\leq \exp\left(\lambda(e^{2\theta^2\sigma^2}-1)\right) \\
&\leq \exp\left(\lambda\left(2\theta^2\sigma^2 e^{2\theta^2\sigma^2}\right)\right),
\end{aligned}
$$

where we use the fact that the ξ_i are σ^2-subgaussian r.v., and the elementary inequality $e^{x^2}-1 \leq x^2 e^{x^2}$ for all $x \in \mathbb{R}$. We conclude that for all $\theta > 0$,

$$
I_Z(\theta) \leq \exp\left(\lambda\left(2\theta^2 M^2 e^{2\theta M} + 2\theta^2\sigma^2 e^{2\theta^2\sigma^2}\right) + \lambda\theta M\right).
$$

Next, we introduce $L = \max(M, \sigma)$. Then, for all $\alpha > 0$, we deduce that

$$
I_Z(\theta) \leq \exp\left(2\lambda\theta^2 L^2\left(e^{\alpha} + e^{\alpha^2}\right) + \lambda\theta M\right), \qquad \forall|\theta| \leq \frac{\alpha}{2L}.
$$

By Markov inequality, and fixing $\alpha = 1$, we have

$$
\mathbb{P}(Z - \lambda M > u) \leq \inf_{|\theta|\leq 1/(2L)} I_Z(\theta)e^{-\lambda\theta M - \theta u} \leq \exp\left(-\min\left(\frac{u^2}{16e\lambda L^2}, \frac{u}{4L}\right)\right).
$$

Similarly, we have

$$
\mathbb{P}(\lambda M - Z > u) \leq \exp\left(-\min\left(\frac{u^2}{16e\lambda L^2}, \frac{u}{4L}\right)\right).
$$

The final tail bound follows from a union bound. Finally, straightforward computations yield an upper bound on $\mathbb{E}[|\lambda M - Z|^2]$. Indeed, we have

$$
\mathbb{E}[|\lambda M - Z|^2] \leq 2\mathbb{E}[|Y-\lambda|^2]M^2 + 2\mathbb{E}\left[\left(\sum_{i=1}^{Y}\xi_i\right)^2\right] \leq 2\lambda M^2 + 16\lambda\sigma^2 \leq 18\lambda L^2.
$$

□

D.2 Random matrices with compound Poisson entries

We list below the two main concentration results that we need for the forthcoming analysis. In Proposition 26, we provide a high probability guarantee on the error between the empirical mean reward matrix and the true matrix in operator norm. In Proposition 27, we establish another concentration result that will be instrumental in the subspace recovery analysis. The proofs of the two results are similar with slight differences and they both rely on Theorem 23. The proofs are presented at the end of this subsection.

Proposition 26. *Under the random matrix model* (15) *with compound Poisson entries, for all* $\delta \in (0,1)$*, for all* $T \geq 13(n+m)\log^3((n+m)/\delta)$*, the following statement*

$$\|\widetilde{M} - M\| \leq 36\sqrt{2}L\sqrt{\frac{nm}{T}}\left(\sqrt{(n+m)\log\left(\frac{n+m}{\delta}\right)} + \log^{3/2}\left(\frac{n+m}{\delta}\right)\right)$$

holds with probability at least $1-\delta$*, where* $L = \max(\|M\|_\infty, \sigma)$*.*

Proposition 27. *Let* A *be a* $m \times 2r$ *nonrandom matrix, and* B *be a* $n \times 2r$ *nonrandom matrix. Then, under the random matrix model* (15) *with compound Poisson entries, and denoting* $L = \max(\|M\|_\infty, \sigma)$*, we have:*

(i) for all $\ell \in [m]$*, for all* $\delta \in (0,1)$*, for all* $T \geq m\log^3(en/\delta)$*, the following event*

$$\|(\widetilde{M}_{\ell,:} - M_{\ell,:})A\| \leq 73\sqrt{2}L\|A\|_{2\to\infty}\sqrt{\frac{nm}{T}}\left(\sqrt{n\log\left(\frac{en}{\delta}\right)} + \log^{3/2}\left(\frac{en}{\delta}\right)\right) \tag{23}$$

holds with probability at least $1-\delta$*;*

(ii) for all $k \in [n]$*, for all* $\delta \in (0,1)$*, for all* $T \geq n\log^3(em/\delta)$*, the following event*

$$\|(\widetilde{M}_{:,k} - M_{:,k})^\top B\| \leq 73\sqrt{2}L\|B\|_{2\to\infty}\sqrt{\frac{nm}{T}}\left(\sqrt{m\log\left(\frac{em}{\delta}\right)} + \log^{3/2}\left(\frac{em}{\delta}\right)\right) \tag{24}$$

holds with probability at least $1-\delta$*.*

Proof of Proposition 26. To simplify the notation, introduce the matrices $Z_{i,j} = (\widetilde{M}_{i,j} - M_{i,j})e_i e_j^\top$, for all $(i,j) \in [m]\times[n]$, $\lambda = T/mn$, and $L = \max(\|M\|_\infty, \sigma)$. We remark that we can write

$$\widetilde{M} - M = \sum_{(i,j)\in[m]\times[n]} Z_{i,j}.$$

Starting from the above expression, we will apply Theorem 23 to obtain the desired result. First, we note that for all $(i,j) \in [m]\times[n]$, $\|Z_{i,j}\| = |\widetilde{M}_{i,j} - M_{i,j}|$ and $\widetilde{M}_{i,j} - M_{i,j}$ is a centered and normalized compound Poisson random variable. Thus, we have by Lemma 25, for all $\delta \in (0,1)$, $\mathbb{P}(\|Z_{i,j}\| > \beta) \leq \delta/(2n^2m^2)$, where we define

$$\begin{aligned}\beta &= 4L\max\left(\sqrt{\frac{e}{\lambda}\log\left(\frac{4n^2m^2}{\delta}\right)}, \frac{1}{\lambda}\log\left(\frac{4n^2m^2}{\delta}\right)\right),\\ &\leq 4L\max\left(\sqrt{\frac{4e}{\lambda}\log\left(\frac{n+m}{\delta}\right)}, \frac{4}{\lambda}\log\left(\frac{n+m}{\delta}\right)\right).\end{aligned}$$

Moreover, we have

$$\begin{aligned}\mathbb{E}\left[\|Z_{i,j}\|\mathbb{1}_{\{\|Z_{i,j]}\|>\beta\}}\right] &\leq \sqrt{\mathbb{E}\left[\|Z_{i,j}\|^2\right]\mathbb{E}\left[\mathbb{1}_{\{\|Z_{i,j]}\|>\beta\}}\right]}\\ &\leq \sqrt{\mathbb{E}[|\widetilde{M}_{i,j} - M_{i,j}|^2]\mathbb{P}(\|Z_{i,j}\| > \beta)}\\ &\leq \sqrt{\frac{9L^2\delta}{\lambda n^2m^2}},\end{aligned}$$

where the first inequality follows from Cauchy-Schwarz inequality, the second inequality follows from the expression of $Z_{i,j}$, and the third inequality follows from Lemma 25. Next, we have

$$\begin{aligned}\left\|\sum_{(i,j)\in[m]\times[n]}\mathbb{E}\left[Z_{i,j}Z_{i,j}^\top\right]\right\| &= \left|\sum_{i\in[m]}\left(\sum_{j\in[n]}\mathbb{E}\left[\left(\widetilde{M}_{i,j} - M_{i,j}\right)^2\right]\right)e_ie_i^\top\right|\\ &= \max_{i\in[m]}\sum_{j\in[n]}\mathbb{E}\left[\left(\widetilde{M}_{i,j} - M_{i,j}\right)^2\right]\\ &\leq \frac{18nL^2}{\lambda}.\end{aligned}$$

By symmetry, we obtain similarly

$$\left\| \sum_{(i,j)\in[m]\times[n]} \mathbb{E}\left[Z_{i,j}^\top Z_{i,j}\right] \right\| \leq \frac{18mL^2}{\lambda}.$$

Let us set $v = 18(n \wedge m)L^2/\lambda$. We conclude using Theorem 23 that, for all $u > 0$,

$$\begin{aligned}
\mathbb{P}\left(\|\widetilde{M} - M\| > \sqrt{\frac{9L^2\delta}{\lambda}} + u \right) &\leq \frac{\delta}{2(nm)} + (n+m)\exp\left(-\frac{u^2/2}{v + \beta u/3} \right) \\
&\leq \frac{\delta}{2(nm)} + (n+m)\exp\left(-\frac{1}{4}\min\left(\frac{u^2}{v}, \frac{3u}{\beta} \right) \right).
\end{aligned}$$

We re-parametrize by choosing $\delta = 2(n+m)\exp(-(1/4)\min(u^2/v, 3u/\beta))$, and write

$$\mathbb{P}\left(\|\widetilde{M} - M\| > \frac{3L\sqrt{\delta}}{\sqrt{\lambda}} + u \right) \leq \delta \tag{25}$$

with

$$\begin{aligned}
u &= \max\left(\sqrt{4v\log\left(\frac{2(n+m)}{\delta}\right)}, \frac{4\beta}{3}\log\left(\frac{2(n+m}{\delta}\right) \right) \\
&\leq \max\left(\sqrt{8v\log\left(\frac{n+m}{\delta}\right)}, \frac{8\beta}{3}\log\left(\frac{n+m}{\delta}\right) \right).
\end{aligned}$$

By inspecting the definition of β and v, we note that under the condition

$$\lambda = \frac{T}{nm} \geq \frac{4^5}{3^4}\frac{1}{n \wedge m}\log^3\left(\frac{n+m}{\delta}\right) \tag{26}$$

then

$$\begin{aligned}
u &\leq \max\left(\sqrt{8v\log\left(\frac{n+m}{\delta}\right)}, \frac{16L\sqrt{2e}}{3\sqrt{\lambda}}\log^{3/2}\left(\frac{2(n+m)}{\delta}\right) \right) \\
&\leq \frac{L}{\sqrt{\lambda}}\max\left(\sqrt{3^2 4^2 (n \wedge m)\log\left(\frac{n+m}{\delta}\right)}, \frac{4^2 2\sqrt{e}}{3}\log\left(\frac{n+m}{\delta}\right) \right).
\end{aligned}$$

After using the upper bound on u in (25), and after upper bounding δ by 1, we obtain, under the condition (26),

$$\begin{aligned}
\|\widetilde{M} - M\| &> \frac{L}{\sqrt{\lambda}}\left(3 + 12\sqrt{2(n+m)\log\left(\frac{n+m}{\delta}\right)} + \frac{4^3\sqrt{e}}{3}\log^{3/2}\left(\frac{n+m}{\delta}\right) \right) \\
&> \frac{L}{\sqrt{\lambda}}\left(36\sqrt{2(n+m)\log\left(\frac{n+m}{\delta}\right)} + 36\log^{3/2}\left(\frac{n+m}{\delta}\right) \right) \\
&> \frac{36\sqrt{2}L}{\sqrt{\lambda}}\left(\sqrt{(n+m)\log\left(\frac{n+m}{\delta}\right)} + \log^{3/2}\left(\frac{n+m}{\delta}\right) \right)
\end{aligned}$$

with probability at most δ. Noting that a stricter condition than (26) is

$$T \geq 13(n+m)\log^3\left(\frac{n+m}{\delta}\right),$$

we complete the proof. □

Proof of Proposition 27. To simplify the notation, let us denote $Z_j = (\widetilde{M}_{\ell,j} - M_{\ell,j})A_{j,:}$, $\lambda = mn/T$, and $L = \max(\|M\|_\infty, \sigma)$. We remark that we can write

$$(\widetilde{M}_{\ell,:} - M_{\ell,:})A = \sum_{j\in[n]} (\widetilde{M}_{\ell,j} - M_{\ell,j})A_{j,:} = \sum_{j\in[n]} Z_j.$$

Starting from the above expression, we will apply 23 to obtain the desired result. First, we note that for all $j \in [n]$, $\|Z_j\| = |\widetilde{M}_{\ell,j} - M_{\ell,j}|\|A_{j,:}\|$, and $\widetilde{M}_{\ell,j} - M_{\ell,j}$ is a centered and normalized compound Poisson random variable. Thus, we have by Lemma 25, for all $\delta \in (0,1)$, $\mathbb{P}\left(\|Z_j\| > \|A\|_{2\to\infty}\beta\right) \le \mathbb{P}\left(\|Z_j\| > \|A_{j,:}\|\beta\right) \le \delta/(2n^2)$, where we define

$$\begin{aligned}\beta &= 4L\max\left(\sqrt{\frac{e}{\lambda}\log\left(\frac{4n^2}{\delta}\right)}, \frac{1}{\lambda}\log\left(\frac{4n^2}{\delta}\right)\right)\\ &\le 4L\max\left(\sqrt{\frac{2e}{\lambda}\log\left(\frac{en}{\delta}\right)}, \frac{2}{\lambda}\log\left(\frac{en}{\delta}\right)\right).\end{aligned}$$

Moreover, we have

$$\begin{aligned}\mathbb{E}\left[\|Z_j\|\mathbb{1}_{\{\|Z_j\|>\|A\|_{2\to\infty}\beta\}}\right] &\le \sqrt{\mathbb{E}[\|Z_j\|^2]\mathbb{P}\left(\|Z_j\| > \|A\|_{2\to\infty}\beta\right)}\\ &\le \|A\|_{2\to\infty}\sqrt{\frac{\mathbb{E}[\|\widetilde{M}_{\ell,:} - M_{\ell,:}\|^2]\delta}{2n^2}}\\ &\le \|A\|_{2\to\infty}\sqrt{\frac{9L^2\delta}{\lambda n^2}},\end{aligned}$$

where in the first inequality, we use Cauchy-Schwarz inequality, and in the third inequality, the result of Lemma 25 to upper bound the variances. Next, we have

$$\begin{aligned}\left\|\mathbb{E}\left[\sum_{j\in[n]} Z_j Z_j^\top\right]\right\| &\le \sum_{j\in[n]} \mathbb{E}\left[(\widetilde{M}_{\ell,j} - M_{\ell,j})^2\right]\|A_{j,:}\|^2\\ &\le \frac{18L^2\|A\|_F^2}{\lambda}\\ &\le \frac{18L^2 n\|A\|_{2\to\infty}^2}{\lambda},\end{aligned}$$

where we simply used the expressions of Z_j, $j \in [n]$, the triangular inequality, and Lemma 25 to upper bound the variances. Similarly, we have

$$\left\|\mathbb{E}\left[\sum_{j\in[n]} Z_j^\top Z_j\right]\right\| \le \frac{18L^2 n\|A\|_{2\to\infty}^2}{\lambda}.$$

We set $v = 18L^2 n\|A\|_{2\to\infty}^2/\lambda$. Now we are ready to apply Theorem 23. We get:

$$\mathbb{P}\left(\|(\widetilde{M}_{\ell,:} - M_{\ell,:})A\| > \|A\|_{2\to\infty}\sqrt{\frac{9L^2\delta}{\lambda}} + u\right) \le \frac{\delta}{2n} + n\exp\left(-\frac{1}{4}\min\left(\frac{u^2}{v}, \frac{3u}{\|A\|_{2\to\infty}\beta}\right)\right).$$

We re-parametrize by choosing $\delta = 2n\exp(-(1/4)\min(u^2/v, 3u/(\|A\|_{2\to\infty}\beta)))$ and we write

$$\mathbb{P}\left(\|(\widetilde{M}_{\ell,:} - M_{\ell,:})A\| > \|A\|_{2\to\infty}\sqrt{\frac{9L^2\delta}{\lambda}} + u\right) \le \delta$$

with

$$\begin{aligned}u &= \max\left(\sqrt{4v\log\left(\frac{2n}{\delta}\right)}, \frac{4\|A\|_{2\to\infty}\beta}{3}\log\left(\frac{2n}{\delta}\right)\right)\\ &\le \max\left(\sqrt{4v\log\left(\frac{en}{\delta}\right)}, \frac{4\|A\|_{2\to\infty}\beta}{3}\log\left(\frac{en}{\delta}\right)\right).\end{aligned}$$

By inspecting the definition of β and v, we note that when the condition

$$\lambda = \frac{T}{mn} \geq \frac{4^3}{3^4}\frac{1}{n}\log^3\left(\frac{en}{\delta}\right) \tag{27}$$

holds, then

$$\begin{aligned} u &\leq \max\left(\sqrt{4v\log\left(\frac{en}{\delta}\right)}, \frac{16\sqrt{2e}L\|A\|_{2\to\infty}}{3\sqrt{\lambda}}\log^{3/2}\left(\frac{en}{\delta}\right)\right) \\ &\leq \frac{L\|A\|_{2\to\infty}}{\sqrt{\lambda}}\max\left(\sqrt{2^3 3^2 n\log\left(\frac{en}{\delta}\right)}, \frac{16\sqrt{2e}}{3}\log^{3/2}\left(\frac{en}{\delta}\right)\right) \\ &\leq \frac{36\sqrt{2}L\|A\|_{2\to\infty}}{\sqrt{\lambda}}\max\left(\sqrt{n\log\left(\frac{en}{\delta}\right)}, \log^{3/2}\left(\frac{en}{\delta}\right)\right). \end{aligned} \tag{28}$$

After using the upper bound in (28), and upper bounding δ by 1, we obtain that, under the condition (27),

$$\|(\widetilde{M}_{\ell,:} - M_{\ell,:})A\| > \frac{73\sqrt{2}L\|A\|_{2\to\infty}}{\sqrt{\lambda}}\left(\sqrt{n\log\left(\frac{en}{\delta}\right)} + \log^{3/2}\left(\frac{en}{\delta}\right)\right)$$

holds with probability at most δ. We can also refine the condition (27) as follows

$$T \geq m\log^3\left(\frac{en}{\delta}\right).$$

This concludes the proof of the statement (23) in the proposition. The statement (24) follows similarly. Therefore, we omit it. □

D.3 Random matrices with Poisson entries

Recall from Section B.2, the definition of the function g_δ from (9) and that $\mathcal{A} = \frac{1}{\sqrt{T}}\sqrt{\|M\|_{1\to\infty} + \|M^\top\|_{1\to\infty}}$. First we show the following lemma that provides an upper bound of the spectral norm. This lemma is used to derive Lemma 13.

Lemma 28. *Let $Y \in \mathbb{R}^{n\times n}$ be a matrix with independent entries $Y_{i,j} \sim T^{-1}\mathrm{Poisson}(TM_{ij})$, $i,j \in [n]$, and let $0 \leq \delta \leq 1$. Then, w.p. at least $1-\delta$, $\|Y - M\| \leq C\mathcal{A} + \frac{C}{T}g_\delta(TM)\sqrt{\log(\frac{ne}{\delta}})$.*

Proof. The proof follows from that of Lemma 29 and that of Lemma 4 in [42], which is based on a spectral bound from [6]. We use that the random variables $|Y_{i,j} - M_{i,j}|$ concentrate well around $L = L_1\mathbb{1}_{\{\exists\ell:T\|M_{\ell,:}\|_\infty\leq 1\}} + L_2\mathbb{1}_{\{\forall\ell:T\|M_{\ell,:}\|_\infty>1\}}$ where $L_1 = 4T^{-1}\log^{-1}(1 + (T\|M\|_\infty)^{-1} \wedge n\delta^{-1})\log(\frac{ne}{\delta})$ and $L_2 = 4\sqrt{T^{-1}\|M\|_\infty}\log\left(T\|M\|_\infty\frac{ne}{\delta}\right)$ using exactly the same argument as in the first step of Lemma 29. Moreover, we use upper bound on $|\mathbb{E}[(Y_{i,j} - M_{i,j})\mathbb{1}_{\{|Y_{i,j}-M_{i,j}|<L\}}]|$ derived in the second step of Lemma 29. □

We also derive upper bounds in the $\ell_{2\to\infty}$ norm. These bounds are used in the analysis of the singular subspace recovery in Lemma 32, and therefore in the proofs of Theorems 3 and 5.

Lemma 29. *Let $Y \in \mathbb{R}^{n\times n}$ be a matrix with independent entries $Y_{i,j} \sim T^{-1}\mathrm{Poisson}(TM_{ij})$, $i,j \in [n]$, for an arbitrary integer $T > 0$. Let $0 \leq \delta \leq 1$. Then, for any $1 \leq l \leq n$ and any matrix $A \in \mathbb{R}^{n\times p}$, with $p \leq n$, and independent of $Y_{l,:}$ we have, if $T\|M_{l,:}\|_\infty \leq 1$,*

$$\|(Y_{l,:} - M_{l,:})A\| \lesssim \|A\|_F\frac{\sqrt{\|M_{l,:}\|_\infty\log\left(\frac{ne}{\delta}\right)}}{\sqrt{T}} + \|A\|_{2\to\infty}\frac{\log^2\left(\frac{ne}{\delta}\right)}{T\log(1 + (T\|M_{l,:}\|_\infty)^{-1}\wedge n\delta^{-1}))}$$

else if $T\|M_{l,:}\|_\infty > 1$,

$$\|(Y_{l,:} - M_{l,:})A\| \lesssim \|A\|_F\frac{\sqrt{\|M_{l,:}\|_\infty\log\left(\frac{ne}{\delta}\right)}}{\sqrt{T}} + \|A\|_{2\to\infty}\frac{\sqrt{\|M_{l,:}\|_\infty}}{\sqrt{T}}\log\left(T\|M_{l,:}\|_\infty\frac{ne}{\delta}\right)\log\left(\frac{ne}{\delta}\right)$$

with probability at least $1 - \delta/n$.

Proof of Lemma 29. The lemma is an application of the truncated matrix Bernstein theorem i.e. Theorem 23. In this theorem, T corresponds to n in Lemma 29, n in Theorem 23 corresponds to 1 in Lemma 29, and m in Theorem 23 corresponds to n in Lemma 29. First note that for any l, we have $(Y_{l,:} - M_{l,:})A = \sum_{i=1}^{n}(Y_{l,i} - M_{l,i})A_{i,:}$. Moreover, since each of these n summands are independent, zero-mean random vectors, we can identify Z_i's from Theorem 23 with $(Y_{l,i} - M_{l,i})A_{i,:} \in \mathbb{R}^{1\times n}$ for $i \in [n]$. To apply Theorem 23, we need to verify its assumptions. This is done below.

Step 1: Showing *(i)* in (20) First, recall Bennett's concentration inequality from Lemma 24, which in our case implies that for any $i, j \in [n]$:

$$\mathbb{P}(|Y_{i,j} - M_{i,j}| \geq tM_{i,j}) \leq 2\exp\left(-h(t)TM_{i,j}\right). \tag{29}$$

Note that $\|Z_i\|$ in Theorem 23 in our case corresponds to:

$$\|(Y_{l,i} - M_{l,i})A_{i,:}\| = |Y_{l,i} - M_{l,i}|\|A_{i,:}\| \leq |Y_{l,i} - M_{l,i}|\|A\|_{2\to\infty}.$$

We consider two different cases:

1. $T\|M_{l,:}\|_\infty \leq 1$: We let $\beta_1 = 4T^{-1}\|A\|_{2\to\infty}\log^{-1}(1 + (T\|M_{l,:}\|_\infty)^{-1} \wedge n\delta^{-1})\log(\frac{ne}{\delta})$ and note that $h(t) \geq \frac{1}{2}t\log t$ for $t \geq 1$. Thus, from Equation (29), we have:

$$\mathbb{P}\left(|Y_{l,i} - M_{l,i}| \geq \frac{\beta_1}{\|A\|_{2\to\infty}}\right) \leq 2\exp\Bigg(-2\log(\frac{ne}{\delta})\log^{-1}(1 + (T\|M_{l,:}\|_\infty)^{-1} \wedge n\delta^{-1}) \cdot \log\left(\frac{4\log(\frac{ne}{\delta})}{T\|M_{l,:}\|_\infty \log(1 + (T\|M_{l,:}\|_\infty)^{-1} \wedge n\delta^{-1})}\right)\Bigg) \leq \frac{\delta}{2n^2}.$$

where, in the second inequality, we show using simple algebra that $\log^{-1}(1 + (T\|M_{l,:}\|_\infty)^{-1} \wedge n\delta^{-1})\log\left(\frac{4\log(\frac{ne}{\delta})}{T\|M_{l,:}\|_\infty\log(1+(T\|M_{l,:}\|_\infty)^{-1}\wedge n\delta^{-1})}\right) \geq 1$ for $\delta \leq 1$ and $T\|M_{l,:}\|_\infty \leq 1$.

2. $T\|M_{l,:}\|_\infty > 1$: Here we define $\beta_2 := 4\|A\|_{2\to\infty}\sqrt{T^{-1}\|M_{l,:}\|_\infty}\log\left(T\|M_{l,:}\|_\infty\frac{ne}{\delta}\right)$. Then, according to Equation (29) and the approximation $h(t) \geq \min\{t^2/4, t\}$ for $t \geq 0$, we have:

$$\begin{aligned}
&\mathbb{P}\left(|Y_{l,i} - M_{l,i}| \geq \frac{\beta_2}{\|A\|_{2\to\infty}}\right) \\
&\leq 2\exp\left(-4\min\left\{\log^2(T\|M_{l,:}\|_\infty\frac{ne}{\delta}), \sqrt{T\|M_{l,:}\|_\infty}\log(T\|M_{l,:}\|_\infty\frac{ne}{\delta})\right\}\right) \\
&\leq 2\exp\left(-4\log(T\|M_{l,:}\|_\infty\frac{ne}{\delta})\right) \leq \frac{1}{2T\|M_{l,:}\|_\infty}\frac{\delta}{n^2}.
\end{aligned}$$

where, in the second inequality, we used that $\delta \leq 1$ and $T\|M_{l,:}\|_\infty > 1$. Finally, we define $\beta = \beta_1\mathbb{1}_{\{T\|M_{l,:}\|_\infty\leq 1\}} + \beta_2\mathbb{1}_{\{T\|M_{l,:}\|_\infty>1\}}$ and $p = \frac{\delta}{2n}\mathbb{1}_{\{T\|M_{l,:}\|_\infty\leq 1\}} + \frac{1}{2T\|M_{l,:}\|_\infty}\frac{\delta}{n}\mathbb{1}_{\{T\|M_{l,:}\|_\infty>1\}}$ (since we took union bound over $i \in [n]$).

Step 2: Showing *(ii)* in (20) In our case the l.h.s. corresponds to $\|\mathbb{E}[(Y_{l,i} - M_{l,i})A_{i,:}\mathbb{1}_{\{\|(Y_{l,i}-M_{l,i})A_{i,:}\|>\beta\}}]\| = \|\mathbb{E}[(Y_{l,i}-M_{l,i})A_{i,:}\mathbb{1}_{\{\|(Y_{l,i}-M_{l,i})A_{i,:}\|\leq\beta\}}]\|$ which can be upper bounded by $\|A\|_{2\to\infty}|\mathbb{E}[(Y_{l,i} - M_{l,i})\mathbb{1}_{\{|Y_{l,i}-M_{l,i}|\leq\frac{\beta}{\|A_{i,:}\|}\}}]|$. For some integers $\kappa_{\min}, \kappa_{\max}$, let $Y_{l,i} \in \frac{1}{T}[\kappa_{\min}, \kappa_{\max}]$ be interval of $Y_{l,i}$ for which indicator $\mathbb{1}_{\{|Y_{l,i}-M_{l,i}|\leq\frac{\beta}{\|A_{i,:}\|}\}}$ is active and note that this is a superset of interval for which $\mathbb{1}_{\{|Y_{l,i}-M_{l,i}|\leq\frac{\beta}{\|A\|_{2\to\infty}}\}}$ is active. Then from the definition of Poisson random variables and the bounds derived previously, we obtain:

$$\begin{aligned}
\left|\mathbb{E}[(Y_{l,i} - M_{l,i})\mathbb{1}_{\{|Y_{l,i}-M_{l,i}|\leq\frac{\beta}{\|A_{i,:}\|}\}}]\right| &= \frac{1}{T}\Big|\sum_{k=\kappa_{\min}}^{\kappa_{\max}}(k - TM_{l,i})\frac{\exp(-TM_{l,i})(TM_{l,i})^k}{k!}\Big| \\
&= M_{l,i}\Big|\sum_{k=\kappa_{\min}-1}^{\kappa_{\max}-1}\frac{\exp(-TM_{l,i})(TM_{l,i})^k}{k!} - \sum_{k=\kappa_{\min}}^{\kappa_{\max}}\frac{\exp(-TM_{l,i})(TM_{l,i})^k}{k!}\Big| \\
&\leq M_{l,i}(\mathbb{P}(TY_{l,i} = \kappa_{\min} - 1) + \mathbb{P}(TY_{l,i} = \kappa_{\max})) \leq \frac{2\delta}{Tn^2}\min\{T\|M_{l,:}\|_\infty, 1\},
\end{aligned}$$

where we assumed that $\kappa_{\min} \geq 1$, otherwise we keep just the second probability term above. Thus, using previous two inequalities, we have:

$$\|\mathbb{E}[(Y_{l,i} - M_{l,i})A_{i,:}\mathbb{1}_{\{\|(Y_{l,i}-M_{l,i})A_{i,:}\|>\beta\}}]\| \leq \|A\|_{2\to\infty} \frac{2\delta}{Tn^2} \min\{T\|M_{l,:}\|_\infty, 1\}$$

Step 3: Showing *(iii)* in (21) Using our definition $Z_i = (Y_{l,i} - M_{l,i})A_{i,:} \in \mathbb{R}^{1\times n}$, we have that

$$\begin{aligned} Z_i Z_i^\top &= (Y_{l,i} - M_{l,i})^2 \|A_{i,:}\|^2, \\ Z_i^\top Z_i &= (Y_{l,i} - M_{l,i})^2 A_{i,:}^\top A_{i,:}. \end{aligned}$$

Since A and $Y_{l,:}$ are independent, we have:

$$\|\sum_{i=1}^n \mathbb{E}[Z_i Z_i^\top]\| = \sum_{i=1}^n \mathbb{E}[Z_i Z_i^\top] = \sum_{i=1}^n \|A_{i,:}\|^2 \mathbb{E}(Y_{l,i} - M_{l,i})^2 \leq \|A\|_F^2 \max_i \mathbb{E}(Y_{l,i} - M_{l,i})^2$$

and

$$\begin{aligned} \|\sum_{i=1}^n \mathbb{E}[Z_i^\top Z_i]\| = \|\sum_{i=1}^n \mathbb{E}(Y_{l,i} - M_{l,i})^2 A_{i,:}^\top A_{i,:}\| &\leq \sum_{i=1}^n \mathbb{E}(Y_{l,i} - M_{l,i})^2 \|A_{i,:}^\top A_{i,:}\| \\ &\leq \|A\|_F^2 \max_i \mathbb{E}(Y_{l,i} - M_{l,i})^2. \end{aligned}$$

Now note that for $Y_{l,i} \sim T^{-1}\text{Poisson}(TM_{l,i})$, $\text{Var}(Y_{l,i}) = \mathbb{E}(Y_{l,i} - M_{l,i})^2 = T^{-1}M_{l,i}$. Thus, by setting $v = T^{-1}\|A\|_F^2 \|M_{l,:}\|_\infty$, we get *(iii)*.

Plugging in all obtained quantities into Equation (22) finishes proof of the lemma. □

E Singular subspace recovery via the leave-one-out argument

In this section, we present Lemma 30 and Lemma 32 providing sharp guarantees for the singular subspace recovery in two-to-infinity norm. Obtaining such guarantees is not trivial and requires the use of a rather technical analysis, namely the leave-one-out technique [2, 14]. However, such technique heavily relies on independence between entries of the observed random matrix. We use the Poisson approximation argument to address this, which in turn requires to reproduce the leave-one-out analysis under a different random matrix observation models (see (15) and (19)).

We wish to highlight that Farias et al. [23], like us, have also used the leave-one-out argument to obtain entry-wise guarantees for matrix estimation with sub-exponential noise. In our case, we use this argument as a sub-step of our analysis after performing the Poisson approximation. However, we believe that, our final results are richer, more precise and actually needed for our RL applications. Indeed, we are able to obtain guarantees in the norms $\|\cdot\|_{2\to\infty}$ and $\|\cdot\|_{1\to\infty}$ (these are not provided in [23]). Moreover, the entry-wise guarantees in [23] are only expressed in terms of the matrix dimensions m and n. Our guarantees on the other hand exhibit dependencies on the dimensions m, n, the number of observation T and the confidence level δ. Having guarantees with an explicit dependence for all $T \geq 1$ and $\delta \in (0,1)$ is crucial in the design of our algorithm for low-rank bandits.

E.1 Subspace recovery for reward matrices

Lemma 30. *Let $\delta \in (0,1)$. Define:*

$$\mathcal{B} = \sqrt{\frac{nm}{T}}\left(\sqrt{(n+m)\log\left(\frac{e(n+m)T}{\delta}\right)} + \log^{3/2}\left(\frac{e(n+m)T}{\delta}\right)\right).$$

For all $T \geq c(\mu^4\kappa^2 r^2 + 1)(m+n)\log^3\left(e^2(m+n)T/\delta\right)$, the event

$$\max(\|U - \widehat{U}(\widehat{U}^\top U)\|, \|V - \widehat{V}(\widehat{V}^\top V)\|) \leq C\frac{\|M\|\|M\|_\infty}{\sigma_r(M)^2}\max(\|V\|_{2\to\infty}\|U\|_{2\to\infty})\mathcal{B}$$

holds with probability at least $1-\delta$, for some universal constants $c, C > 0$.

Proof of Lemma 30. The proof follows similar steps as that of Theorem 4.2 in [14], which is based on the leave-one-out analysis.

Step 1: Dilation trick. In order to apply the leave-one-out analysis, we first use a dilation trick [58] to reduce the problem to that of symmetric matrices. Define:

$$S = \begin{bmatrix} 0 & M \\ M^\top & 0 \end{bmatrix}$$

and note that for matrix M with SVD $M = U\Sigma V^\top$, we have:

$$S = \frac{1}{\sqrt{2}}\begin{bmatrix} U & U \\ V & -V \end{bmatrix}\begin{bmatrix} \Sigma & 0 \\ 0 & -\Sigma \end{bmatrix}\frac{1}{\sqrt{2}}\begin{bmatrix} U & U \\ V & -V \end{bmatrix}^\top := QDQ^\top.$$

We define, in a similar way, $\widetilde{S}$ using $\widetilde{M}$, and let $\widehat{Q} \in \mathbb{R}^{(n+m)\times 2r}$ be the matrix of eigenvectors of the best $2r$-rank approximation of $\widetilde{S}$. Note that:

$$\|Q - \widehat{Q}(\widehat{Q}^\top Q)\|_{2\to\infty} = \max\left\{\|U - \widehat{U}(\widehat{U}^\top U)\|_{2\to\infty}, \|V - \widehat{V}(\widehat{V}^\top V)\|_{2\to\infty}\right\}. \tag{30}$$

To keep the notation simple, we will define $W_{\widehat{Q}} = \widehat{Q}^\top Q$. Further note that

$$\|\widetilde{S} - S\| = \|\widetilde{M} - M\|, \qquad \sigma_1(S) = \sigma_1(M), \qquad \text{and} \qquad \sigma_{2r}(S) = \sigma_r(M). \tag{31}$$

We start the analysis under the model (15) and assume that $\widetilde{M}$ has independent entries with compound Poisson distributions. We will eventually invoke the Poisson approximation argument via Lemma 20 to deduce the final result.

Step 2: Error decomposition. We apply the decomposition in Lemma 33 to obtain:

$$\|Q-\widehat{Q}W_{\widehat{Q}}\|_{2\to\infty} \le \frac{1}{\sigma_{2r}(S)}\left(\frac{4\|\widetilde{S}Q\|_{2\to\infty}\|E\|}{\sigma_{2r}(S)} + \|EQ\|_{2\to\infty} + 2\|\widetilde{S}(Q-\widehat{Q}W_{\widehat{Q}})\|_{2\to\infty}\right),$$

where we set $E=\widetilde{S}-S$. We observe that when $\|E\|\le \sigma_{2r}(S)/2$, then

$$\|Q-\widehat{Q}W_{\widehat{Q}}\|_{2\to\infty} \le \frac{1}{\sigma_{2r}(S)}\left(\frac{4\|SQ\|_{2\to\infty}\|E\|}{\sigma_{2r}(S)} + 3\|EQ\|_{2\to\infty} + 2\|\widetilde{S}(Q-\widehat{Q}W_{\widehat{Q}})\|_{2\to\infty}\right). \tag{32}$$

Furthermore, we also have

$$\begin{aligned}\|\widetilde{S}(Q-\widehat{Q}W_{\widehat{Q}})\|_{2\to\infty} &\le \|E(Q-\widehat{Q}W_{\widehat{Q}})\|_{2\to\infty} + \|S(Q-\widehat{Q}W_{\widehat{Q}})\|_{2\to\infty}\\ &\le \|E(Q-\widehat{Q}W_{\widehat{Q}})\|_{2\to\infty} + \|SQ\|_{2\to\infty}\|\sin(Q,\widehat{Q})\|^2\\ &\le \|E(Q-\widehat{Q}W_{\widehat{Q}})\|_{2\to\infty} + \frac{\|SQ\|_{2\to\infty}\|E\|^2}{\sigma_{2r}(S)^2}\\ &\le \|E(Q-\widehat{Q}W_{\widehat{Q}})\|_{2\to\infty} + \frac{\|SQ\|_{2\to\infty}\|E\|}{2\sigma_{2r}(S)},\end{aligned}$$

where the first inequality follows from the triangular inequality, the second inequality follows by the relation between the two-to-infinity norm and the sin theorem (see e.g., [12]). The third inequality follows from Davis-Kahan's theorem. The fourth inequality follows under the condition $\|E\|\le\sigma_{2r}(S)/2$. We finally obtain

$$\|Q-\widehat{Q}W_{\widehat{Q}}\|_{2\to\infty} \le \frac{1}{\sigma_{2r}(S)}\left(\frac{5\|SQ\|_{2\to\infty}\|E\|}{\sigma_{2r}(S)} + 3\|EQ\|_{2\to\infty} + 2\|E(Q-\widehat{Q}W_{\widehat{Q}})\|_{2\to\infty}\right). \tag{33}$$

Note that in the above inequality, we can control $\|E\|$ using Proposition 26 and $\|EQ\|_{2\to\infty}$ using Proposition 27. However, the term $\|E(Q-\widehat{Q}W_{\widehat{Q}})\|_{2\to\infty}$ is not easy to control because E and $(Q-\widehat{Q}W_{\widehat{Q}})$ are dependent on each other in a non-trivial way. To control this term, we use the leave-one-out analysis.

Step 3: Leave-one-out analysis. We define a matrix $\widetilde{S}^{(\ell)}\in\mathbb{R}^{(n+m)\times(n+m)}$ as follows:

$$\widetilde{S}^{(\ell)}_{i,j} = \begin{cases}\widetilde{S}_{i,j}, & \text{if } i\neq\ell \text{ or } j\neq\ell\\ S_{i,j}, & \text{otherwise}\end{cases}$$

Then define $\widehat{Q}^{(\ell)}\in\mathbb{R}^{n\times 2r}$ as a matrix of eigenvectors corresponding to the $2r$ greatest (in absolute value) eigenvalues of matrix $\widetilde{S}^{(\ell)}$. Define $W_{\widetilde{U}^{(\ell)}}$ accordingly. We have

$$\|E(Q-\widehat{Q}W_{\widehat{Q}})\|_{2\to\infty} \le \max_{\ell\in[n+m]}\|E_{\ell,:}(Q-\widehat{Q}^{(\ell)}W_{\widehat{Q}^{(\ell)}})\|_2 + \|E\|_2\|\widehat{Q}W_{\widehat{Q}}-\widehat{Q}^{(\ell)}W_{\widehat{Q}^{(\ell)}}\|_F.$$

We have by Proposition 26 that

$$\mathbb{P}\left(\|E\|\lesssim\|M\|_\infty\mathcal{G}\right)\ge 1-\delta$$

provided that

$$\textbf{(C1)}\qquad\qquad T\ge c_1\frac{mn}{m+n}\log^3\left(\frac{e(m+n)}{\delta}\right)$$

and where we define

$$\mathcal{G}=\sqrt{\frac{mn}{T}}\left(\sqrt{(m+n)\log\left(\frac{e(m+n)}{\delta}\right)}+\log^{3/2}\left(\frac{e(m+n)}{\delta}\right)\right).$$

Let us now introduce the event $\mathcal{E}_1$ as follows

$$\mathcal{E}_1=\left\{\|E\|\le\|M\|_\infty\mathcal{G}\right\}.$$

Note that if the following condition holds

$$\textbf{(C2)} \qquad T \geq c_2(\mu\kappa r)^2\left((m+n)\log\left(\frac{e(m+n)}{\delta}\right)+\log^3\left(\frac{e(m+n)}{\delta}\right)\right)$$

for c_2 large enough then $16\|E\| \leq \sigma_r(M)$. Hence, under the event $\mathcal{E}_1$, using Lemma 31, we have

$$\|\widehat{Q}W_{\widehat{Q}} - \widehat{Q}^{(\ell)}W_{\widehat{Q}^{(\ell)}}\|_F \leq \frac{16\|E_{\ell,:}\widehat{Q}^{(\ell)}W_{\widehat{Q}^{(\ell)}}\|_2 + 16\|E\|\|\widehat{Q}W_{\widehat{Q}}\|_{2\to\infty}}{\sigma_{2r}(M)},$$

which further gives by triangular inequality

$$\begin{aligned}\|\widehat{Q}W_{\widehat{Q}} - \widehat{Q}^{(\ell)}W_{\widehat{Q}^{(\ell)}}\|_F \leq{}& \frac{16\|E_{\ell,:}(Q-\widehat{Q}^{(\ell)}W_{\widehat{Q}^{(\ell)}})\|_2}{\sigma_r(M)} \\ &+ \frac{16\left(\|E_{\ell,:}Q\|_2 + \|E\|\|Q-\widehat{Q}W_{\widehat{Q}}\|_{2\to\infty} + \|E\|\|Q\|_{2\to\infty}\right)}{\sigma_r(M)}.\end{aligned}$$

Now, by Proposition 27,

$$\mathbb{P}\left(\|E_{\ell,:}(Q-Q^{(\ell)}W_{\widehat{Q}^{(\ell)}})\|_2 \lesssim \|M\|_\infty\|Q-Q^{(\ell)}W_{\widehat{Q}^{(\ell)}}\|_{2\to\infty}\mathcal{G}\right) \geq 1-\delta$$

as long as the same condition **(C1)** holds with c_1 large enough. So let us introduce the event

$$\mathcal{E}_2 = \left\{\|E_{\ell,:}(Q-Q^{(\ell)}W_{\widehat{Q}^{(\ell)}})\|_2 \lesssim \|M\|_\infty\|Q-Q^{(\ell)}W_{\widehat{Q}^{(\ell)}}\|_{2\to\infty}\mathcal{G}\right\}.$$

We further upper bound under the event $\mathcal{E}_1 \cap \mathcal{E}_2$,

$$\|E_{\ell,:}(Q-\widehat{Q}^{(\ell)}W_{\widehat{Q}^{(\ell)}})\|_2 \lesssim \|M\|_\infty\left(\|Q-\widehat{Q}W_{\widehat{Q}}\|_{2\to\infty} + \|\widehat{Q}W_{\widehat{Q}} - \widehat{Q}^{(\ell)}W_{\widehat{Q}^{(\ell)}}\|_F\right)\mathcal{G}.$$

Note that, under the condition **(C2)** with c_2 large enough, we can also obtain

$$\frac{16\|M\|_\infty}{\sigma_r(M)}\mathcal{G} \leq \frac{1}{2}$$

which entails that

$$\begin{aligned}\|\widehat{Q}W_{\widehat{Q}} - \widehat{Q}^{(\ell)}W_{\widehat{Q}^{(\ell)}}\|_F \leq{}& \frac{32\|M\|_\infty\|Q-\widehat{Q}W_{\widehat{Q}}\|_{2\to\infty}}{\sigma_r(M)}\mathcal{G} \\ &+ \frac{32(\|E_{\ell,:}Q\| + \|E\|\|Q-\widehat{Q}W_{\widehat{Q}}\|_{2\to\infty} + \|E\|\|Q\|_{2\to\infty})}{\sigma_r(M)}.\end{aligned}$$

To simplify the notation, let us define the three errors as

$$\begin{aligned}x &= \|Q-\widehat{Q}W_{\widehat{Q}}\|_{2\to\infty}, \\ y &= \|EQ\|_{2\to\infty} \geq \|E_{\ell,:}Q\|_2, \\ z &= \|E\|\|Q\|_{2\to\infty}.\end{aligned}$$

We have

$$\|\widehat{Q}W_{\widehat{Q}} - \widehat{Q}^{(\ell)}W_{\widehat{Q}^{(\ell)}}\|_F \lesssim \left(\frac{\|M\|_\infty\mathcal{G}}{\sigma_r(M)} + \frac{\|E\|}{\sigma_r(M)}\right)x + \frac{1}{\sigma_r(M)}(y+z).$$

By plugging the above in the previous inequality, we get

$$\|E_{\ell,:}(Q-\widehat{Q}^{(\ell)}W_{\widehat{Q}^{(\ell)}})\|_2 \lesssim \|M\|_\infty\mathcal{G}\left(\left(1+\frac{\|M\|_\infty\mathcal{G}}{\sigma_r(M)} + \frac{\|E\|}{\sigma_r(M)}\right)x + \frac{1}{\sigma_r(M)}(y+z)\right)$$

which entails finally

$$\begin{aligned}\|E(Q-\widehat{Q}W_{\widehat{Q}})\|_{2\to\infty} \lesssim{}& \left(\frac{\|E\|}{\sigma_r(M)} + \frac{\mathcal{G}\|M\|_\infty}{\sigma_1(M)}\right)(y+z) \\ &+ (\|E\| + \mathcal{G}\|M\|_\infty)\left(1+\frac{\mathcal{G}\|M\|_\infty}{\sigma_r(M)} + \frac{\|\mathcal{E}\|}{\sigma_r(M)}\right)x. \qquad (34)\end{aligned}$$

Step 4: Putting everything together. Combining the inequalities (33) and (34) gives

$$\begin{aligned} x \le C_1 \left(\frac{\|E\|}{\sigma_r(M)} + \frac{\|M\|_\infty \mathcal{G}}{\sigma_r(M)} \right) \left(1 + \frac{\|M\|_\infty \mathcal{G}}{\sigma_r(M)} + \frac{\|E\|}{\sigma_r(M)} \right) x \\ + \frac{C_2}{\sigma_r(M)} \left(1 + \frac{\|E\|}{\sigma_r(M)} + \frac{\mathcal{G}\|M\|_\infty}{\sigma_r(M)} \right) y \\ + \frac{C_3}{\sigma_r(M)} \left(\frac{\|M\|}{\sigma_r(M)} + \frac{\|E\|}{\sigma_r(M)} + \frac{\mathcal{G}\|M\|_\infty}{\sigma_r(M)} \right) z. \end{aligned}$$

Under the events $\mathcal{E}_1$ and $\mathcal{E}_2$ and provided that the conditions **(C1)** and **(C2)** hold, for c_1 and c_2 are large enough, we have

$$\begin{aligned} C_1 \left(\frac{\|E\|}{\sigma_r(M)} + \frac{\|M\|_\infty \mathcal{G}}{\sigma_r(M)} \right) \left(1 + \frac{\|M\|_\infty \mathcal{G}}{\sigma_r(M)} + \frac{\|E\|}{\sigma_r(M)} \right) &\le \frac{1}{2}, \\ \left(1 + \frac{\|E\|}{\sigma_r(M)} + \frac{\mathcal{G}\|M\|}{\sigma_r(M)} \right) &\le 3, \\ \left(\frac{\|M\|}{\sigma_r(M)} + \frac{\|E\|}{\sigma_r(M)} + \frac{\mathcal{G}\|M\|_\infty}{\sigma_r(M)} \right) &\le \left(\frac{\|M\|}{\sigma_r(M)} + 2 \right). \end{aligned}$$

Thus, we obtain

$$x \le \frac{1}{\sigma_r(M)} \left(y + \frac{\|M\|}{\sigma_r(M)} z \right).$$

We note that, under a similar conditions as before , we also have by Proposition 26 and Proposition 27 that

$$\begin{aligned} y &\lesssim \|M\|_\infty \|Q\|_{2\to\infty} \mathcal{G} \\ z &\lesssim \|M\|_\infty \|Q\|_{2\to\infty} \mathcal{G} \end{aligned}$$

with probability at least $1-\delta$. Thus, we conclude after further simplifications that for some $C > 0$ large enough, we have

$$\mathbb{P}\left(\|Q - \widehat{Q} W_{\widehat{Q}}\|_{2\to\infty} \le C \frac{\|M\|\|M\|_\infty}{\sigma_r(M)^2} \|Q\|_{2\to\infty} \mathcal{G} \right) \ge 1 - \delta$$

provided

$$T \ge c(\mu^4 \kappa^2 r^2 + 1)(m+n) \log^3 \left(\frac{e(m+n)}{\delta} \right),$$

with

$$\mathcal{G}(n,m,T,\delta) = \sqrt{\frac{nm}{T}} \left(\sqrt{(n+m)\log\left(\frac{e(n+m)}{\delta}\right)} + \log^{3/2}\left(\frac{e(n+m)}{\delta}\right) \right).$$

Step 5: Poisson approximation. To conclude, we now invoke Lemma 20 which entails that under the true model (14), we have

$$\mathbb{P}\left(\|Q - \widehat{Q} W_{\widehat{Q}}\|_{2\to\infty} > C \frac{\|M\|\|M\|_\infty}{\sigma_r(M)^2} \|Q\|_{2\to\infty} \mathcal{G}(n,m,T,\delta) \right) \le e\sqrt{T}\delta$$

provided $T \ge c(\mu^4\kappa^2 r^2 + 1)(m+n)\log^3(e(m+n)/\delta)$. By re-parametrizing with $\delta' = e\sqrt{T}\delta$, we obtain

$$\mathbb{P}\left(\|Q - \widehat{Q} W_{\widehat{Q}}\|_{2\to\infty} > C \frac{\|M\|\|M\|_\infty}{\sigma_r(M)^2} \|Q\|_{2\to\infty} \mathcal{G}(n,m,T,\delta'/e\sqrt{T}) \right) \le \delta',$$

again provided that $T \ge c(\mu^4\kappa^2 r^2 + 1)(m+n)\log^3\left(e^2(m+n)\sqrt{T}/\delta'\right)$. Recalling that $\|Q\|_{2\to\infty} = \max(\|V\|_{2\to\infty}, \|U\|_{2\to\infty})$, we immediately obtain the final result. □

Lemma 31. *Under the notation used in the proof of Lemma 30, provided the condition $\|E\| \leq \sigma_{2r}(S)/16$, the following inequality holds:*

$$\|\widehat{Q}W_{\widehat{Q}} - \widehat{Q}^{(\ell)}W_{\widehat{Q}^{(\ell)}}\|_F \leq \frac{16\|E_{\ell,:}\widehat{Q}^{(\ell)}W_{\widehat{Q}^{(\ell)}}\|_2 + 16\|E\|\|\widehat{Q}W_{\widehat{Q}}\|_{2\to\infty}}{\sigma_{2r}(S)}$$

Proof of Lemma 31. We have

$$\|\widehat{Q}W_{\widehat{Q}} - \widehat{Q}^{(\ell)}W_{\widehat{Q}^{(\ell)}}\|_F \leq \|\widehat{Q}\widehat{Q}^\top - \widehat{Q}^{(\ell)}(\widehat{Q}^{(\ell)})^\top\|_F\|Q\| \leq \frac{2\|(\widetilde{S} - \widetilde{S}^{(\ell)})\widehat{Q}^{(\ell)}\|_F}{|\sigma_{2r}(\widetilde{S}^{(\ell)}) - \sigma_{2r+1}(\widetilde{S}^{(\ell)})|}$$

where the first inequality follows the elementary fact that $\|AB\|_F \leq \|A\|_F\|B\|$, and the second inequality follows by Davis-Kahan. Now, by Weyl's inequality, we have for all $k \in [n+m]$, $|\sigma_k(\widetilde{S}^{(\ell)}) - \sigma_k(S)| \leq \|E^{(\ell)}\| \leq \|E\|$, where the error matrix $E^{(\ell)} = \widetilde{S}^\ell - S$, and more precisely is defined as follows:

$$E^{(\ell)}_{i,j} = \begin{cases} E_{i,j} & \text{if } i \neq \ell \text{ or } j \neq \ell, \\ 0 & \text{otherwise.} \end{cases}$$

The crude inequality $\|E^{(\ell)}\| \leq \|E\|$ follows from the fact that $\|E^{(\ell)}\|$ is equal to the operator norm of a submatrix of E which will always be smaller than $\|E\|$. Therefore, under the condition that $\|E\| \leq \sigma_{2r}(S)/4$, we have $|\sigma_{2r}(\widetilde{S}^{(\ell)}) - \sigma_{2r+1}(\widetilde{S}^{(\ell)})| \geq \sigma_{2r}(S)/2$. In summary, we obtain that

$$\|\widehat{Q}W_{\widehat{Q}} - \widehat{Q}^{(\ell)}W_{\widehat{Q}^{(\ell)}}\|_F \leq \frac{4\|(\widetilde{S} - \widetilde{S}^{(\ell)})\widehat{Q}^{(\ell)}\|_F}{\sigma_{2r}(S)}.$$

Now, we further have by triangular inequality and by definition of $\widetilde{S}^{(\ell)}$:

$$\begin{aligned} \|(\widetilde{S} - \widetilde{S}^{(\ell)})\widehat{Q}^{(\ell)}\|_F &= \|(e_\ell E_{\ell,:} + (E_{:,\ell} - E_{\ell,\ell}e_\ell)e_\ell^\top)\widehat{Q}^{(\ell)}\|_F \\ &\leq \|E_{\ell,:}\widehat{Q}^{(\ell)}\|_2 + \|E_{:,\ell} - E_{\ell,\ell}e_\ell\|_2\|\widehat{Q}^{(\ell)}\|_{2\to\infty} \\ &\leq \|E_{\ell,:}\widehat{Q}^{(\ell)}\|_2 + \|E\|_2\|\widehat{Q}^{(\ell)}\|_{2\to\infty} \\ &\leq \|E_{\ell,:}\widehat{Q}^{(\ell)}\|_2 + 2\|E\|_2\|\widehat{Q}^{(\ell)}(\widehat{Q}^{(\ell)})^\top Q\|_{2\to\infty} \end{aligned}$$

where the last inequality follows under the condition that $\|E\| \leq 2\sigma_{2r}(S)$. Indeed, we have under such condition that $\|\widehat{Q}^{(\ell)}\|_{2\to\infty} = \|\widehat{Q}^{(\ell)}(\widehat{Q}^{(\ell)})^\top Q\|_{2\to\infty} + \|\widehat{Q}^{(\ell)}(\text{sgn}((\widehat{Q}^{(\ell)})^\top Q^\top) - (\widehat{Q}^{(\ell)})^\top Q)\|_{2\to\infty}$, and by Davis-Kahan's inequality $\|\text{sgn}((\widehat{Q}^{(\ell)})^\top Q^\top) - (\widehat{Q}^{(\ell)})^\top Q\| \leq \frac{2\|E^{(\ell)}\|^2}{(\sigma_{2r}(S))^2} \leq \frac{2\|E\|^2}{(\sigma_{2r}(S))^2} \leq \frac{1}{2}$. Similarly, we also have $\|E_{\ell,:}\widehat{Q}^{(\ell)}\|_2 \leq 2\|E_{\ell,:}\widehat{Q}^{(\ell)}(\widehat{Q}^{(\ell)})^\top Q\|_2$.

Hence, we obtain:

$$\|(\widetilde{S} - \widetilde{S}^{(\ell)})\widehat{Q}^{(\ell)}\|_F \leq 2\|E_{\ell,:}\widehat{Q}^{(\ell)}W_{\widehat{Q}^{(\ell)}}\|_2 + 2\|E\|(\|\widehat{Q}W_{\widehat{Q}}\|_{2\to\infty} + \|\widehat{Q}W_{\widehat{Q}} - \widehat{Q}^{(\ell)}W_{\widehat{Q}^{(\ell)}}\|_{2\to\infty})$$

Which entails under the condition that $\|E\| \leq \sigma_{2r}(S)/16$ that

$$\|\widehat{Q}W_{\widehat{Q}} - \widehat{Q}^{(\ell)}W_{\widehat{Q}^{(\ell)}}\|_F \leq \frac{8\|E_{\ell,:}\widehat{Q}^{(\ell)}W_{\widehat{Q}^{(\ell)}}\|_2 + 8\|E\|\|\widehat{Q}W_{\widehat{Q}}\|_{2\to\infty}}{\sigma_{2r}(S)} + \frac{\|\widehat{Q}W_{\widehat{Q}} - \widehat{Q}^{(\ell)}W_{\widehat{Q}^{(\ell)}}\|_{2\to\infty}}{2}$$

After rearranging, we obtain

$$\|\widehat{Q}W_{\widehat{Q}} - \widehat{Q}^{(\ell)}W_{\widehat{Q}^{(\ell)}}\|_F \leq \frac{16\|E_{\ell,:}\widehat{Q}^{(\ell)}W_{\widehat{Q}^{(\ell)}}\|_2 + 16\|E\|\|\widehat{Q}W_{\widehat{Q}}\|_{2\to\infty}}{\sigma_{2r}(S)}$$

□

E.2 Subspace recovery for transition matrices

Lemma 32. *Let $Y \in \mathbb{R}^{n\times n}$ be a matrix of independent Poisson entries with $Y_{i,j} \sim \frac{1}{T}\text{Poisson}(TM_{i,j})$, and let $\widehat{U}, \widehat{V}$ be the matrices of left and right singular vectors of best r-rank* *approximation of Y. Let g_δ be the function defined in (9). Conditioned on the events where $\|Y-M\| \le c_1\sigma_r(M), g_\delta(TM)\log(ne/\delta) \le c_2 T\sigma_r(M), \sqrt{\|M\|_\infty \log(ne/\delta)} \le c_3\sqrt{T}\sigma_r(M)$ for some sufficiently small universal constants $c_1, c_2, c_3 > 0$, we have, with probability at least $1-\delta$,*

$$\max\left\{\|U-\widehat{U}(\widehat{U}^\top U)\|_{2\to\infty}, \|V-\widehat{V}(\widehat{V}^\top V)\|_{2\to\infty}\right\}$$
$$\lesssim \frac{1}{\sigma_r(M)}\left[\mu\sqrt{\frac{r}{n}}\left(\frac{\sigma_1(M)}{\sigma_r(M)}\|Y-M\| + \frac{1}{T}g_\delta(TM)\log\left(\frac{ne}{\delta}\right)\right) + \sqrt{\frac{r\|M\|_\infty}{T}\log\left(\frac{ne}{\delta}\right)}\right].$$

Proof. The proof follows similar steps as the proof of Theorem 4.2 in [14]. In order to apply the leave-one-out technique, we first repeat the symmetric dilation trick as in Step 1 of proof of Lemma 30. We define

$$S = \begin{bmatrix} 0 & M \\ M^\top & 0 \end{bmatrix} \tag{35}$$

and note that for matrix M with SVD $M = U\Sigma V^\top$, we have:

$$S = \frac{1}{\sqrt{2}}\begin{bmatrix} U & U \\ V & -V \end{bmatrix}\begin{bmatrix} \Sigma & 0 \\ 0 & -\Sigma \end{bmatrix}\frac{1}{\sqrt{2}}\begin{bmatrix} U & U \\ V & -V \end{bmatrix}^\top := QDQ^\top.$$

We define $\widetilde{S}$ as the symmetrized version of matrix Y, and let $\widehat{Q} \in \mathbb{R}^{n\times 2r}$ be the matrix of eigenvectors of the best $2r$-rank approximation of $\widetilde{S}$. Note that:

$$\|Q-\widehat{Q}(\widehat{Q}^\top Q)\|_{2\to\infty} = \max\left\{\|U-\widehat{U}(\widehat{U}^\top U)\|_{2\to\infty}, \|V-\widehat{V}(\widehat{V}^\top V)\|_{2\to\infty}\right\}.$$

We will also repeatedly use the properties (31). To keep the notation simple, define $W_{\widehat{Q}} = \widehat{Q}^\top Q$. Thus, proving Lemma 32 is equivalent to showing:

$$\|Q-\widehat{Q}W_{\widehat{Q}}\|_{2\to\infty}$$
$$\lesssim \frac{1}{\sigma_r(M)}\left[\|Q\|_{2\to\infty}(\frac{\sigma_1(M)}{\sigma_r(M)}\|\widetilde{S}-S\| + \frac{1}{T}g_\delta(TM)\log\left(\frac{ne}{\delta}\right)) + \sqrt{\frac{r\|M\|_\infty}{T}\log\left(\frac{ne}{\delta}\right)}\right]$$

with high probability. Define $E = \widetilde{S} - S$. Now, as in Lemma 33, we have:

$$\|Q-\widehat{Q}W_{\widehat{Q}}\|_{2\to\infty} \le \frac{1}{\sigma_r(M)}\left(\frac{4\|\widetilde{S}Q\|_{2\to\infty}\|E\|}{\sigma_r(M)} + \|EQ\|_{2\to\infty} + 2\|\widetilde{S}(Q-\widehat{Q}W_{\widehat{Q}})\|_{2\to\infty}\right) \tag{36}$$

under the assumption that $\|E\| \le c_1\sigma_r(M)$. Indeed, it is straightforward to show the same bounds as in Lemma 4.14 in [14] - note that the boundedness assumption is not used in these lemmas. We bound the three terms in Equation (36) as follows:

1. To bound the first term, we use:

$$\|\widetilde{S}Q\|_{2\to\infty} \le \|SQ\|_{2\to\infty} + \|EQ\|_{2\to\infty} \le \|Q\|_{2\to\infty}\|S\| + \|EQ\|_{2\to\infty} \tag{37}$$

where we first used the triangle inequality and then $\|SQ\|_{2\to\infty} = \|QD\|_{2\to\infty} \le \|Q\|_{2\to\infty}\|S\|$.

2. For the second term, according to Lemma 29, we obtain with probability at least $1-\delta$:

$$\|EQ\|_{2\to\infty} \lesssim \frac{1}{T}\left[\|Q\|_F\sqrt{T\|M\|_\infty\log(ne/\delta)} + g_\delta(TM)\log(ne/\delta)\|Q\|_{2\to\infty}\right]. \tag{38}$$

Moreover, we will use $\|Q\|_F \le \sqrt{2r}$ and $\|Q\|_{2\to\infty} \le \mu\sqrt{\frac{r}{n}}$, which follow from the low-rank and incoherence assumptions.

3. Finally, regarding the last term in Equation (36), we split it using the triangle inequality as follows:

$$\|\widetilde{S}(Q-\widehat{Q}W_{\widehat{Q}})\|_{2\to\infty} \le \|S(Q-\widehat{Q}W_{\widehat{Q}})\|_{2\to\infty} + \|E(Q-\widehat{Q}W_{\widehat{Q}})\|_{2\to\infty},$$

and from Step 3 of proof of Theorem 4.2 in [14] we have:

$$\|S(Q-\widehat{Q}W_{\widehat{Q}})\|_{2\to\infty} \le \|Q\|_{2\to\infty}\|S\|\|Q^\top(Q-\widehat{Q}W_{\widehat{Q}})\| \lesssim \|Q\|_{2\to\infty}\|S\|\frac{\|E\|^2}{\sigma_r^2(M)}, \tag{39}$$

where we used $\|Q^\top(Q - \widehat{Q}W_{\widehat{Q}})\| = \|\sin\Theta(Q, \widehat{Q})\|^2$. The remaining of the proof consists in bounding $\|E(Q - \widehat{Q}W_{\widehat{Q}})\|_{2\to\infty} = \max_{\ell=1,\dots,n}\|E_{\ell,:}(Q - \widehat{Q}W_{\widehat{Q}})\|$. First note that the matrix $Q - \widehat{Q}W_{\widehat{Q}}$ depends on E and thus we cannot apply Lemma 29 immediately. Instead, we will use the leave-one-out method, and define a matrix $\widetilde{S}^{(\ell)} \in \mathbb{R}^{n\times n}$ as follows:

$$\widetilde{S}^{(\ell)}_{i,j} = \begin{cases} \widetilde{S}_{i,j}, & \text{if } i \neq \ell \text{ or } j \neq \ell \\ S_{i,j}, & \text{otherwise} \end{cases}$$

Then define $\widehat{Q}^{(\ell)} \in \mathbb{R}^{n\times 2r}$ as a matrix of eigenvectors corresponding to $2r$ greatest (in absolute value) eigenvalues of matrix $\widetilde{S}^{(\ell)}$. Define $W_{\widehat{Q}^{(\ell)}}$ accordingly. Then we have:

$$\|E(Q - \widehat{Q}W_{\widehat{Q}})\|_{2\to\infty} \leq 2\max_{1\leq\ell\leq n}\left\{\|E_{\ell,:}(Q - \widehat{Q}^{(\ell)}W_{\widehat{Q}^{(\ell)}})\|, \|E_{\ell,:}(\widehat{Q}^{(\ell)}W_{\widehat{Q}^{(\ell)}} - \widehat{Q}W_{\widehat{Q}})\|\right\}. \quad (40)$$

(3a) Since $E_{\ell,:}$ is statistically independent of $Q - \widehat{Q}^{(\ell)}W_{\widehat{Q}^{(\ell)}}$, the first term from (40) can be bounded according to Lemma 29 for any $1 \leq \ell \leq n$ as follows:

$$\begin{aligned}\|E_{\ell,:}(Q - \widehat{Q}^{(\ell)}W_{\widehat{Q}^{(\ell)}})\| \lesssim \frac{1}{T}\Big[&\|Q - \widehat{Q}^{(\ell)}W_{\widehat{Q}^{(\ell)}}\|_F\sqrt{T\|M\|_\infty \log\left(\frac{ne}{\delta}\right)} \\ &+ g_\delta(TM)\log\left(\frac{ne}{\delta}\right)\|Q - \widehat{Q}^{(\ell)}W_{\widehat{Q}^{(\ell)}}\|_{2\to\infty}\Big]\end{aligned}$$

with probability at least $1-\delta$. After applying the triangle inequality to the second term and using $\|\cdot\|_{2\to\infty} \leq \|\cdot\|_F$, we get:

$$\|Q - \widehat{Q}^{(\ell)}W_{\widehat{Q}^{(\ell)}}\|_{2\to\infty} \leq \|Q - \widehat{Q}W_{\widehat{Q}}\|_{2\to\infty} + \|\widehat{Q}W_{\widehat{Q}} - \widehat{Q}^{(\ell)}W_{\widehat{Q}^{(\ell)}}\|_F. \quad (41)$$

Thus, combining the last two inequalities yields:

$$\begin{aligned}&\|E_{\ell,:}(Q - \widehat{Q}^{(\ell)}W_{\widehat{Q}^{(\ell)}})\| \\ &\lesssim \|Q - \widehat{Q}W_{\widehat{Q}}\|_F\sqrt{T^{-1}\|M\|_\infty\log\left(\frac{ne}{\delta}\right)} + T^{-1}g_\delta(TM)\log\left(\frac{ne}{\delta}\right)\|Q - \widehat{Q}W_{\widehat{Q}}\|_{2\to\infty} \\ &+ \|\widehat{Q}W_{\widehat{Q}} - \widehat{Q}^{(\ell)}W_{\widehat{Q}^{(\ell)}}\|_F\left(\sqrt{T^{-1}\|M\|_\infty\log\left(\frac{ne}{\delta}\right)} + T^{-1}g_\delta(TM)\log\left(\frac{ne}{\delta}\right)\right) \quad (42)\end{aligned}$$

for all $1 \leq \ell \leq n$.

(3b) The second term from (40) can be bounded very roughly as follows:

$$\|E_{\ell,:}(\widehat{Q}^{(\ell)}W_{\widehat{Q}^{(\ell)}} - \widehat{Q}W_{\widehat{Q}}))\| \leq \|E\|\|\widehat{Q}^{(\ell)}W_{\widehat{Q}^{(\ell)}} - \widehat{Q}W_{\widehat{Q}}\|_F. \quad (43)$$

Similar to the Step 2.2 in the proof of Theorem 4.2 in [14], we have:

$$\begin{aligned}\|\widehat{Q}^{(\ell)}W_{\widehat{Q}^{(\ell)}} - \widehat{Q}W_{\widehat{Q}}\|_F \leq \frac{16}{\sigma_r(M)}(&\|E_{\ell,:}(Q - \widehat{Q}^{(\ell)}W_{\widehat{Q}^{(\ell)}})\| + \|EQ\|_{2\to\infty} \\ &+ \|E\|\|Q - \widehat{Q}^{(\ell)}W_{\widehat{Q}^{(\ell)}}\|_{2\to\infty} + \|E\|\|Q\|_{2\to\infty}).\end{aligned}$$

Applying again the inequality (41) and moving the term $\|\widehat{Q}^{(\ell)}W_{\widehat{Q}^{(\ell)}} - \widehat{Q}W_{\widehat{Q}}\|_F$ to the left side of inequality, we get under assumption $\|E\| \leq c_1\sigma_r(M)$ that:

$$\begin{aligned}\|\widehat{Q}^{(\ell)}W_{\widehat{Q}^{(\ell)}} - \widehat{Q}W_{\widehat{Q}}\|_F \lesssim \frac{1}{\sigma_r(M)}(&\|E_{\ell,:}(Q - \widehat{Q}^{(\ell)}W_{\widehat{Q}^{(\ell)}})\| + \|EQ\|_{2\to\infty} \\ &+ \|E\|\|Q - \widehat{Q}W_{\widehat{Q}}\|_{2\to\infty} + \|E\|\|Q\|_{2\to\infty}). \quad (44)\end{aligned}$$

After substitution of the results from (37), (38), (39), (40), (42), (43) and (44) into Equation (36) and using assumptions stated in the lemma, we obtain the statement of the lemma. □

E.3 Error decomposition in the two-to-infinity norm

Below, we present a decomposition for the error of subspace recovery in the norm $\|\cdot\|_{2\to\infty}$. We borrow this result from [14] and provide its proof for completeness.

Lemma 33 (Lemma 4.16 in [14])**.** *Let $S, \widetilde{S} \in \mathbb{R}^{n\times n}$ be symmetric matrices and assume that S and $\widetilde{S}$ are of rank r. Let $Q, \widehat{Q} \in \mathcal{O}_{n\times r}$ be the corresponding r singular vectors of S and $\widetilde{S}$, respectively. Denote $E = \widetilde{S} - S$. Under the condition $\|E\| \le \sigma_r(S)/2$, we have:*

$$\|Q - \widehat{Q}\widehat{Q}^\top Q\|_{2\to\infty} \le \frac{4\|\widetilde{S}Q\|_{2\to\infty}\|E\|}{(\sigma_r(S))^2} + \frac{\|EQ\|_{2\to\infty}}{\sigma_r(S)} + \frac{2\|\widetilde{S}(Q - \widehat{Q}\widehat{Q}^\top Q)\|_{2\to\infty}}{\sigma_r(S)}$$

Proof of Lemma 33. Since S is a symmetric matrix of rank r, by SVD we write $S = Q\Sigma Q^\top$, where the matrix $\Sigma = \text{diag}(\sigma_1(S), \dots, \sigma_r(S))$. For ease of notations, let us further denote $W = \widehat{Q}Q^\top$. We have

$$\begin{aligned}\|Q - \widehat{Q}\widehat{Q}^\top Q\|_{2\to\infty} &= \|SQ\Sigma^{-1} - \widehat{Q}W\|_{2\to\infty} \\ &\le \|\widetilde{S}Q\Sigma^{-1} - \widehat{Q}W\|_{2\to\infty} + \|EQ\Sigma^{-1}\|_{2\to\infty} \\ &\le \frac{\|\widetilde{S}Q - \widehat{Q}W\Sigma\|_{2\to\infty}}{\sigma_r(S)} + \frac{\|EQ\|_{2\to\infty}}{\sigma_r(S)}\end{aligned}$$

Now, we focus on the term $\|\widetilde{S}Q - \widehat{Q}W\Sigma\|_{2\to\infty}$. To that end, we first establish the identity

$$\begin{aligned}\widehat{Q}W\Sigma &= \widehat{Q}\widehat{Q}^\top Q\Sigma \\ &= \widehat{Q}\widehat{Q}^\top SQ \\ &= \widehat{Q}\widehat{Q}^\top \widetilde{S}Q + \widehat{Q}\widehat{Q}^\top EQ \\ &= \widehat{Q}\widehat{\Sigma}\widehat{Q}^\top Q + \widehat{Q}\widehat{Q}^\top EQ \\ &= \widetilde{S}\widehat{Q}\widehat{Q}^\top Q + \widehat{Q}\widehat{Q}^\top EQ\end{aligned}$$

where we use the identities $SQ = Q\Sigma$, $\widehat{Q}^\top\widetilde{S} = \widehat{\Sigma}\widehat{Q}^\top$, and $\widehat{Q}\widehat{\Sigma}\widehat{Q}^\top = \widetilde{S}\widehat{Q}\widehat{Q}^\top$. Then, we observe that

$$\begin{aligned}\|\widetilde{S}Q - \widehat{Q}W\Sigma\|_{2\to\infty} &= \|\widetilde{S}Q - \widetilde{S}\widehat{Q}\widehat{Q}^\top Q + \widehat{Q}\widehat{Q}^\top EQ\|_{2\to\infty} \\ &\le \|\widetilde{S}(Q - \widehat{Q}\widehat{Q}^\top Q)\|_{2\to\infty} + \|\widehat{Q}\widehat{Q}^\top EQ\|_{2\to\infty}\end{aligned}$$

Next, we note that when $\|E\| \le \sigma_r(S)/2$, we have

$$\begin{aligned}\|\widehat{Q}\widehat{Q}^\top EQ\|_{2\to\infty} &= \|\widetilde{S}\widehat{Q}\widehat{\Sigma}^{-1}\widehat{Q}^\top EQ\|_{2\to\infty} \\ &\le \|\widetilde{S}\widehat{Q}\|_{2\to\infty}\|\widehat{\Sigma}^{-1}\|\|\widehat{Q}^\top\|\|E\|\|Q\| \\ &\le \frac{\|\widetilde{S}\widehat{Q}\text{sgn}(\widehat{Q}^\top Q)\|_{2\to\infty}\|E\|}{\sigma_r(\widetilde{S})}.\end{aligned}$$

At this point, we try to bound $\|\widetilde{S}\widehat{Q}\|_{2\to\infty}$ and $\sigma_r(\widetilde{S})$, under the condition $\|E\| \le \sigma_r(S)/2$. First, we can easily see by Weyl's inequality we have $|\sigma_r(\widetilde{S}) - \sigma_r(S)| \le \|E\|$, which entails under the assumed condition that $\sigma_r(\widetilde{S}) \ge \sigma_r(S)/2$. Next, we observe:

$$\begin{aligned}\|\widetilde{S}\widehat{Q}\|_{2\to\infty} &= \|\widetilde{S}\widehat{Q}\text{sgn}(\widehat{Q}^\top Q)\|_{2\to\infty} \\ &\le \|\widetilde{S}\widehat{Q}\widehat{Q}^\top Q\|_{2\to\infty} + \|\widetilde{S}\widehat{Q}\|_{2\to\infty}\|\text{sgn}(\widehat{Q}^\top Q) - \widehat{Q}^\top Q\| \\ &\le \|\widetilde{S}\widehat{Q}\widehat{Q}^\top Q\|_{2\to\infty} + \frac{2\|\widetilde{S}\widehat{Q}\|_{2\to\infty}\|E\|^2}{\sigma_r(S)^2} \\ &\le \|\widetilde{S}\widehat{Q}\widehat{Q}^\top Q\|_{2\to\infty} + \frac{\|\widetilde{S}\widehat{Q}\|_{2\to\infty}}{2}\end{aligned}$$

where we used the Davis-Kahan's inequality and properties of the $\mathrm{sgn}(\cdot)$, to upper bound $\|\mathrm{sgn}(\widehat{Q}^\top Q) - \widehat{Q}^\top Q\| \leq \|\sin(\widehat{Q}, Q)\|^2 \leq 2\|E\|^2/(\sigma_r(S))^2$. Thus, leading to:

$$\|\widetilde{S}\widehat{Q}\|_{2\to\infty} \leq 2\|\widetilde{S}\widehat{Q}\widehat{Q}^\top Q\|_{2\to\infty}$$

Moving forward we obtain:

$$\begin{aligned}
\|\widehat{Q}\widehat{Q}^\top EQ\|_{2\to\infty} &\leq \frac{4\|\widetilde{S}\widehat{Q}\widehat{Q}^\top Q\|_{2\to\infty}\|E\|}{\sigma_r(S)} \\
&\leq \frac{4\left(\|\widetilde{S}(\widehat{Q}\widehat{Q}^\top Q - Q)\|_{2\to\infty} + \|\widetilde{S}Q\|_{2\to\infty}\right)\|E\|}{\sigma_r(S)} \\
&\leq 2\|\widetilde{S}(\widehat{Q}\widehat{Q}^\top Q - Q)\|_{2\to\infty} + \frac{4\left(\|\widetilde{S}Q\|_{2\to\infty}\right)\|E\|}{\sigma_r(S)}
\end{aligned}$$

Now, putting everything together we conclude that:

$$\|Q - \widehat{Q}\widehat{Q}^\top Q\|_{2\to\infty} \leq \frac{2\|\widetilde{S}(\widehat{Q}\widehat{Q}^\top Q - Q)\|_{2\to\infty}}{\sigma_r(S)} + \frac{4\|\widetilde{S}Q\|_{2\to\infty}\|E\|}{(\sigma_r(S))^2} + \frac{\|EQ\|_{2\to\infty}}{\sigma_r(S)}$$

□

F Row-wise and entry-wise matrix estimation errors

In this appendix, we provide a series of results about quantifying the matrix estimation error using different norms. It is important to note that all these results require a control of the error in the two-to-infinity norm, which in turn requires a subspace recovery guarantee in the two-to-infinity norm. Lemmas 35 and 37 are specific to our analysis for the estimation of the transition matrices. Lemmas 34 and 36 are common to the analysis of both the estimation of reward matrices and transition matrices. The results presented in this appendix are used in the proofs of the main results, presented in Appendix B.

F.1 Bounding $\|M - \widehat{M}\|_{2\to\infty}$

Lemma 34. *Let $M, \widehat{M}$ be as in §2. Assume that there exists a sufficiently small universal constant $c_1 > 0$ such that $\|M - \widetilde{M}\| \le c_1 \sigma_r(M)$. Then, there exists a universal constant $c_2 > 0$ such that*

$$\|\widehat{M} - M\|_{2\to\infty} \le c_2 \sigma_1(M) \left[\|U - \widehat{U}(\widehat{U}^\top U)\|_{2\to\infty} + \|U\|_{2\to\infty} \frac{\|\widetilde{M} - M\|}{\sigma_r(M)} \right].$$

Proof. We start by using definition of $\widehat{M}$ as a projection of matrix $\widetilde{M}$, and then use the triangle inequality and the inequality (3) to obtain:

$$\begin{aligned}
\|\widehat{M} - M\|_{2\to\infty} &= \|\Pi_{\widehat{U}} \widetilde{M} - \Pi_U M\|_{2\to\infty} \\
&= \|(\Pi_{\widehat{U}} - \Pi_U)(\widetilde{M} - M) + \Pi_U(\widetilde{M} - M) + (\Pi_{\widehat{U}} - \Pi_U)M\|_{2\to\infty} \\
&\le \|(\Pi_{\widehat{U}} - \Pi_U)(\widetilde{M} - M)\|_{2\to\infty} + \|\Pi_U(\widetilde{M} - M)\|_{2\to\infty} + \|(\Pi_{\widehat{U}} - \Pi_U)M\|_{2\to\infty} \\
&\le \|\Pi_{\widehat{U}} - \Pi_U\|_{2\to\infty}(\|\widetilde{M} - M\| + \|M\|) + \|\Pi_U\|_{2\to\infty}\|\widetilde{M} - M\|. \qquad (45)
\end{aligned}$$

Moreover, we note that $\|\Pi_U\|_{2\to\infty} = \|U\|_{2\to\infty}$ (refer to Proposition 6.6 in [12]). In the remaining of the proof, we upper bound $\|\Pi_{\widehat{U}} - \Pi_U\|_{2\to\infty}$ from (45). For any orthogonal matrix $R \in \mathcal{O}^{r\times r}$, we have

$$\begin{aligned}
\|\Pi_{\widehat{U}} - \Pi_U\|_{2\to\infty} &= \|\widehat{U}\widehat{U}^\top - UU^\top\|_{2\to\infty} \\
&= \|\widehat{U}RR^\top\widehat{U}^\top - UR^\top\widehat{U}^\top + UR^\top\widehat{U}^\top - UU^\top\|_{2\to\infty} \\
&\le \|\widehat{U}RR^\top\widehat{U}^\top - UR^\top\widehat{U}^\top\|_{2\to\infty} + \|UR^\top\widehat{U}^\top - UU^\top\|_{2\to\infty} \\
&\le \|\widehat{U}R - U\|_{2\to\infty}\|R^\top\widehat{U}^\top\| + \|U\|_{2\to\infty}\|R^\top\hat{U}^\top - U^\top\| \\
&\le \|U - \widehat{U}R\|_{2\to\infty} + \|U\|_{2\to\infty}\|U - \widehat{U}R\|. \qquad (46)
\end{aligned}$$

Recall the definition of sgn function given in the notation presented in §1 and choose the matrix R as $R = \mathrm{sgn}(\widehat{U}^\top U)$. For this choice of R we have according to Davis-Kahan's theorem (Corollary 2.8 in [14]):

$$\|U - \widehat{U}R\| \le \sqrt{2}\|\sin\Theta(\widehat{U}, U)\| \le \frac{2\|M - \widetilde{M}\|}{\sigma_r(M)}. \qquad (47)$$

Define the matrix $W_{\widehat{U}} = \widehat{U}^\top U$. We use the facts that

$$\|U - \widehat{U}R\|_{2\to\infty} \le \|U - \widehat{U}W_{\widehat{U}}\|_{2\to\infty} + \|\widehat{U}\|_{2\to\infty}\|W_{\widehat{U}} - R\| \qquad (48)$$

and that $W_{\widehat{U}}$ is very close to R according to the proof of Lemma 4.15 in [14] to show:

$$\|W_{\widehat{U}} - R\| = \|\widehat{U}^\top U - \mathrm{sgn}(\widehat{U}^\top U)\| = \|\sin\Theta(\widehat{U}, U)\|^2 \le \frac{2\|M - \widetilde{M}\|^2}{\sigma_r^2(F)}. \qquad (49)$$

We also have $\sigma_i(R) = 1$ for $i \in [r]$ and according to Weyl's inequality $\sigma_{\min}(W_{\widehat{U}}) \ge \sigma_{\min}(R) - \|W_{\widehat{U}} - R\| = 1 - \|W_{\widehat{U}} - R\|$. Combining these results under assumption $\|M - \widetilde{M}\| < \sigma_r(M)/\sqrt{2}$ we obtain:

$$\|W_{\widehat{U}}^{-1}\| = \frac{1}{\sigma_{\min}(W_{\widehat{U}})} \le \frac{1}{1 - \|W_{\widehat{U}} - R\|} \le \frac{1}{1 - \frac{2\|M - \widetilde{M}\|^2}{\sigma_r^2(M)}}.$$

Thus:

$$\|\widehat{U}\|_{2\to\infty} \le \|\widehat{U}W_{\widehat{U}}\|_{2\to\infty}\|W_{\widehat{U}}^{-1}\| \le \frac{1}{1-\frac{2\|M-\widetilde{M}\|^2}{\sigma_r^2(M)}}(\|U\|_{2\to\infty} + \|U-\widehat{U}W_{\widehat{U}}\|_{2\to\infty}). \tag{50}$$

Combining Equations (48), (49), (50) we get:

$$\|U-\widehat{U}R\|_{2\to\infty} \le \frac{1}{1-\frac{2\|M-\widetilde{M}\|^2}{\sigma_r^2(M)}}(\|U-\widehat{U}W_{\widehat{U}}\|_{2\to\infty} + \frac{2\|M-\widetilde{M}\|^2}{\sigma_r^2(M)}\|U\|_{2\to\infty})$$

and combining the last equality with (46) and (47), we have

$$\|\Pi_{\widehat{U}} - \Pi_U\|_{2\to\infty} \le \frac{\|U-\widehat{U}W_{\widehat{U}}\|_{2\to\infty}}{1-\frac{2\|M-\widetilde{M}\|^2}{\sigma_r^2(M)}} + \left(\frac{\frac{2\|M-\widetilde{M}\|^2}{\sigma_r^2(M)}}{1-\frac{2\|M-\widetilde{M}\|^2}{\sigma_r^2(M)}} + \frac{2\|M-\widetilde{M}\|}{\sigma_r(M)}\right)\|U\|_{2\to\infty}. \tag{51}$$

Finally, substituting the obtained bound into Equation (45) and using assumption $\|M-\widetilde{M}\| \le c_1\sigma_r(M)$ for simplification, we obtain the statement of the lemma. □

F.2 Bounding $\|P-\widehat{P}\|_{1\to\infty}$

Lemma 35. *Let $P, \widehat{P}$ be as in Model II in §2. We have:*

$$\|\widehat{P}-P\|_{1\to\infty} \le 2\frac{\sqrt{n}\|\widehat{M}-M\|_{2\to\infty}}{\min_{j\in[n]}\|M_{j,:}\|_1}.$$

Proof. Starting with the definition of $\widehat{P}$, we get:

$$\begin{aligned}\|\widehat{P}-P\|_{1\to\infty} = \max_{i\in[n]}\|\widehat{P}_{i,:}-P_{i,:}\|_1 &= \max_{i\in[n]}\left\|\frac{(\widehat{M}_{i,:})_+}{\|(\widehat{M}_{i,:})_+\|_1} - \frac{M_{i,:}}{\|M_{i,:}\|_1}\right\|_1 \\ &\le 2\max_{i\in[n]}\frac{\|\widehat{M}_{i,:}-M_{i,:}\|_1}{\|M_{i,:}\|_1} \le 2\frac{\sqrt{n}\max_{i\in[n]}\|\widehat{M}_{i,:}-M_{i,:}\|}{\min_{j\in[n]}\|M_{j,:}\|_1},\end{aligned}$$

where the first inequality follows from Lemma 2 in [63] and the second by equivalence of norms. Moreover, note that the above inequality holds even in the case when $\|(\widehat{M}_{i,:})_+\|_1 = 0$ (and thus $\widehat{P}_{i,:} = \frac{1}{n}\mathbf{1}_n$), but the bound is vacuous in this case. □

F.3 Bounding $\|M-\widehat{M}\|_\infty$

Lemma 36. *Let $M, \widehat{M}$ be as in §2. Assume that there exists a sufficiently small universal constant $c_1 > 0$ such that $\|M-\widetilde{M}\| \le c_1\sigma_r(M)$. Then, there exists a universal constant $c_2 > 0$ such that*

$$\begin{aligned}\|\widehat{M}-M\|_\infty \le c_2\|M\|_{2\to\infty}&\left(\frac{\|M-\widetilde{M}\|}{\sigma_r(M)}\|V\|_{2\to\infty} + \|V-\widehat{V}W_{\widehat{V}}\|_{2\to\infty}\right) \\ &+ c_2\|M-\widehat{M}\|_{2\to\infty}(\|V\|_{2\to\infty} + \|V-\widehat{V}W_{\widehat{V}}\|_{2\to\infty}).\end{aligned}$$

Proof. Similarly to the decomposition leading to Equation (45), we can upper bound the infinity norm error easily from the following decomposition:

$$\begin{aligned}\|\widehat{M}-M\|_\infty &= \|\widehat{M}\Pi_{\widehat{V}} - M\Pi_V\|_\infty \\ &\le \|\widehat{M}\|_{2\to\infty}\|\Pi_{\widehat{V}}-\Pi_V\|_{2\to\infty} + \|\widehat{M}-M\|_{2\to\infty}\|\Pi_V\|_{2\to\infty} \\ &\le (\|\widehat{M}-M\|_{2\to\infty} + \|M\|_{2\to\infty})\|\Pi_{\widehat{V}}-\Pi_V\|_{2\to\infty} + \|\widehat{M}-M\|_{2\to\infty}\|V\|_{2\to\infty},\end{aligned}$$

where we used the inequality (4) together with the triangle inequalities and the fact that projection matrices are symmetric. To bound $\|\Pi_{\widehat{V}} - \Pi_V\|_{2\to\infty}$, we use the same approach as that used in (51) (just replacing U by V), and we obtain:

$$\|\Pi_{\widehat{V}} - \Pi_V\|_{2\to\infty} \le \frac{\|V - \widehat{V}W_{\widehat{V}}\|_{2\to\infty}}{1 - \frac{2\|M-\widetilde{M}\|^2}{\sigma_r^2(M)}} + \left(\frac{\frac{2\|M-\widetilde{M}\|^2}{\sigma_r^2(M)}}{1 - \frac{2\|M-\widetilde{M}\|^2}{\sigma_r^2(M)}} + \frac{2\|M - \widetilde{M}\|}{\sigma_r(M)} \right) \|V\|_{2\to\infty}.$$

□

F.4 Bounding $\|P - \widehat{P}\|_\infty$

Lemma 37. *Let $P, \widehat{P}$ be as in Model II in §2. Assume that $\mathcal{D} = \min_{i\in[n]} \|M_{i,:}\|_1 > 0$. If $\|\widehat{M} - M\|_{1\to\infty} \le \frac{1}{2}\mathcal{D}$, then*

$$\|\widehat{P} - P\|_\infty \le 2\frac{\|\widehat{M} - M\|_\infty}{\mathcal{D}} + 2\frac{\sqrt{n}\|M\|_\infty}{\mathcal{D}^2}\|\widehat{M} - M\|_{2\to\infty}.$$

Proof. First note that for any $i \in [n]$:

$$\Big| \|(\widehat{M}_{i,:})_+\|_1 - \|M_{i,:}\|_1 \Big| \le \|(\widehat{M}_{i,:})_+ - M_{i,:}\|_1 \le \|\widehat{M}_{i,:} - M_{i,:}\|_1 \le \frac{\|M_{i,:}\|_1}{2}, \tag{52}$$

where the first inequality follows from the reverse triangle inequality, the second from $|\max(0,x) - y| \le |x - y|$ for all $y > 0$ and $x \in \mathbb{R}$, and the last inequality follows from the assumption in the lemma. This implies that $\|(\widehat{M}_{i,:})_+\|_1 > 0$ for all $i \in [n]$, which further implies that $\widehat{P}$ is defined by: for all $i \in [n]$,

$$\widehat{P}_{i,:} = (\widehat{M}_{i,:})_+ / \|(\widehat{M}_{i,:})_+\|_1. \tag{53}$$

Furthermore, we have for all $i, j = 1, \dots, n$,

$$\begin{aligned}
|\widehat{P}_{i,j} - P_{i,j}| &= \left| \frac{(\widehat{M}_{i,j})_+}{\|(\widehat{M}_{i,:})_+\|_1} - \frac{M_{i,j}}{\|M_{i,:}\|_1} \right| \le \left| \frac{\widehat{M}_{i,j}}{\|(\widehat{M}_{i,:})_+\|_1} - \frac{M_{i,j}}{\|M_{i,:}\|_1} \right| \\
&\le \frac{1}{\|M_{i,:}\|_1} \left| \widehat{M}_{i,j} - M_{i,j} \right| + |\widehat{M}_{i,j}| \left| \frac{1}{\|(\widehat{M}_{i,:})_+\|_1} - \frac{1}{\|M_{i,:}\|_1} \right| \\
&= \frac{1}{\|M_{i,:}\|_1} \left| \widehat{M}_{i,j} - M_{i,j} \right| + \frac{|\widehat{M}_{i,j}|}{\|M_{i,:}\|_1} \left| \frac{1}{1 + \frac{\|(\widehat{M}_{i,:})_+\|_1 - \|M_{i,:}\|_1}{\|M_{i,:}\|_1}} - 1 \right| \\
&= \frac{1}{\|M_{i,:}\|_1} \left| \widehat{M}_{i,j} - M_{i,j} \right| + \frac{|\widehat{M}_{i,j}|}{\|M_{i,:}\|_1} \varphi\left(\frac{\|(\widehat{M}_{i,:})_+\|_1 - \|M_{i,:}\|_1}{\|M_{i,:}\|_1} \right)
\end{aligned}$$

where we define $\varphi(x) = |x/(1+x)|$ for all $x \in \mathbb{R}\backslash\{-1\}$. Note that if $|x| < 1/2$, then $\varphi(x) \le 2|x|$, which combined with (52) gives

$$\begin{aligned}
|\widehat{P}_{i,j} - P_{i,j}| &\le \frac{1}{\|M_{i,:}\|_1} \left| \widehat{M}_{i,j} - M_{i,j} \right| + \frac{2|\widehat{M}_{i,j}|}{\|M_{i,:}\|_1} \left| \frac{\|(\widehat{M}_{i,:})_+\|_1 - \|M_{i,:}\|_1}{\|M_{i,:}\|_1} \right| \\
&\le \frac{1}{\|M_{i,:}\|_1} \left| \widehat{M}_{i,j} - M_{i,j} \right| + \frac{2}{\|M_{i,:}\|_1^2} \left(\left| \widehat{M}_{i,j} - M_{i,j} \right| + |M_{i,j}| \right) \|\widehat{M}_{i,:} - M_{i,:}\|_1.
\end{aligned}$$

Using the assumption $\|\widehat{M} - M\|_{1\to\infty} \le \frac{1}{2}\min_{i\in[n]} \|M_{i,:}\|_1$ again, we can group first two terms, and then use $\|\widehat{M} - M\|_{1\to\infty} \le \sqrt{n}\|\widehat{M} - M\|_{2\to\infty}$ to get the statement of the lemma.

□

G Low-rank bandits: proofs of results from Section 4

G.1 Gap-dependent guarantees

Proof of Theorem 7. First, we prove the result corresponding the best entry identification problem. We proceed in several steps.

Step 1: entry-wise concentration. We can easily verify that for all $\ell \geq 1$, for all $(i,j) \in [m] \times [n]$, we have

$$|\widehat{\Delta}_{i,j}^{(\ell)} - \Delta_{i,j}| \leq 2\|\widehat{M}^{(\ell)} - M^\star\|_\infty.$$

Therefore, applying Theorem 1, we have, for $\delta > 0$, and $T_\ell \geq c_1(m+n)\log^3((e^2(m+n)(mn)/\delta_\ell)$,

$$\mathbb{P}\left(|\widehat{\Delta}_{i,j} - \Delta_{i,j}| > 2C_1\sqrt{\frac{e(m+n)}{T_\ell}\log^3\left(\frac{e(m+n)mnT_\ell}{\delta_\ell}\right)}\right) \leq \frac{\delta_\ell}{mn}$$

for some $c_1, C_1 > 0$ sufficiently large. In particular, we can choose $C_1 = C(\mu^{11/2}\kappa^2 r^{1/2} + \mu^3\kappa r^{3/2}(m+n)/\sqrt{mn})$, and $c_1 = c\mu^4\kappa^2 r^2$, but under a homogeneous reward matrix these constants are $\Theta(1)$. Thus, by a union bound and always under the same conditions, we have

$$\mathbb{P}\left(\max_{(i,j)\in[m]\times[n]}|\widehat{\Delta}_{i,j} - \Delta_{i,j}| > 2C_1\sqrt{\frac{e(m+n)}{T_\ell}\log^3\left(\frac{e(m+n)mnT_\ell}{\delta_\ell}\right)}\right) \leq \delta_\ell.$$

Next, we wish to choose T_ℓ so that we have

$$\mathbb{P}\left(\max_{i,j}|\widehat{\Delta}_{i,j} - \Delta_{i,j}| \leq 2^{-(\ell+2)}\right) \geq 1 - \delta_\ell. \tag{54}$$

Note that in order for the above guarantee to hold, it is sufficient to have:

$$T_\ell \geq c_1(m+n)\log^3\left(\frac{e^2(m+n)(mn)}{\delta_\ell}\right),$$
$$T_\ell \geq 2\sqrt{e}C_1^2(m+n)2^{2(\ell-2)}\log^3\left(\frac{e(m+n)(mn)}{\delta_\ell}\right).$$

This can be achieved if we choose

$$T_\ell = \left\lceil C_3 2^{2(\ell-2)}(m+n)\log^3\left(\frac{2^{2(\ell-2)}(m+n)}{\delta_\ell}\right)\right\rceil, \tag{55}$$

for some positive constant $C_3 > 0$ large enough which can be determined explicitly and only depend on c_1, C_1. Indeed, this can be deduced from the basic fact that if $T_\ell^{1/3} \geq 2a\log(2a) + 2b$, then $T_\ell^{1/3} \geq a\log(T_\ell^{1/3}) + b$. We spare the reader these tedious calculations and only argue that such C_3 exists and can be computed explicitly.

Step 2: Good events. We define $S_\ell = \left\{(i,j) \in [n] \times [m] : \Delta_{i,j} \leq 2^{-\ell}\right\}$ and the good events under which we correctly find the best entry as

$$\mathcal{E}_\ell = \{\mathcal{A}_{\ell+1} \subseteq S_{\ell+1}\} \cap \{(i^\star, j^\star) \in \mathcal{A}_{\ell+1}\}.$$

We show that the good event $\mathcal{E}_\ell$ happens with high probability conditionally on $\mathcal{E}_1, \dots, \mathcal{E}_{\ell-1}$. Observe that by independence of the entries sampled at epoch ℓ from those of the previous epochs, we have based on (54)

$$\mathbb{P}\left(\max_{i,j}|\widehat{\Delta}_{i,j} - \Delta_{i,j}| \leq 2^{-(\ell+2)}\Big|\mathcal{E}_{\ell-1}, \dots, \mathcal{E}_1\right) \geq 1 - \delta_\ell$$

Now, conditionally on $\mathcal{E}_{\ell-1}, \dots, \mathcal{E}_1$, under the event that $\max_{i,j}|\widehat{\Delta}_{i,j} - \Delta_{i,j}| \leq 2^{-(\ell+2)}$, if $(i,j) \in S_{\ell+1}^c \cap \mathcal{A}_{\ell+1}$ then

$$\widehat{\Delta}_{i,j}^{(\ell)} \geq \Delta_{i,j} - 2^{-(\ell+2)} > 2^{-(\ell+1)} - 2^{-(\ell+2)} = 2^{-(\ell+2)}.$$

Thus, we have

$$\mathbb{P}\left(\mathcal{A}_{\ell+1} \subseteq S_{\ell+1} \middle| \mathcal{E}_{\ell-1}, \dots, \mathcal{E}_1\right) \geq \mathbb{P}\left(\max_{i,j} |\widehat{\Delta}_{i,j} - \Delta_{i,j}| \leq 2^{-(\ell+2)} \middle| \mathcal{E}_{\ell-1}, \dots, \mathcal{E}_1\right) \geq 1 - \delta_\ell.$$

Furthermore, note that under the event $\max_{i,j} |\widehat{\Delta}_{i,j} - \Delta_{i,j}| \leq 2^{-(\ell+2)}$, we clearly have that $\widehat{\Delta}_{i^\star, j^\star} \leq 2^{-(\ell+2)}$ and since $(i^\star, j^\star) \in \mathcal{A}_\ell$ conditionally on $\mathcal{E}_{\ell-1}$ we conclude that

$$\mathbb{P}\left(\mathcal{E}_\ell \middle| \mathcal{E}_{\ell-1}, \dots, \mathcal{E}_1\right) \geq \mathbb{P}\left(\max_{i,j} |\widehat{\Delta}_{i,j} - \Delta_{i,j}| \leq 2^{-(\ell+2)} \middle| \mathcal{E}_{\ell-1}, \dots, \mathcal{E}_1\right) \geq 1 - \delta_\ell$$

Step 3: Sample complexity. First, we remark that when $\ell = \lceil \log_2(1/\Delta_{\min}) \rceil$, we have $S_\ell = \{(i^\star, j^\star)\}$. Therefore, under the event

$$\mathcal{E}_1 \cap \dots \cap \mathcal{E}_{\lceil \log_2(1/\Delta_{\min}) \rceil}$$

the algorithm will stop after τ rounds, and recommend the optimal $(i^\star, j^\star)$, where

$$\begin{aligned}
\tau &\leq \sum_{\ell=1}^{\lceil \log_2(1/\Delta_{\min}) \rceil} T_\ell \\
&\leq \sum_{\ell=1}^{\lceil \log_2(1/\Delta_{\min}) \rceil} \left\lceil C_3 2^{2(\ell-2)} (m+n) \log^3\left(\frac{2^{2(\ell-2)}(m+n)}{\delta_\ell}\right) \right\rceil \\
&\leq \sum_{\ell=1}^{\lceil \log_2(1/\Delta_{\min}) \rceil} \left\lceil C_3 \frac{(m+n)}{\Delta_{\min}^2} \log^3\left(\frac{(m+n)\lceil \log_2(1/\Delta_{\min}) \rceil^2}{\Delta_{\min}^2 \delta}\right) \right\rceil \\
&\leq \log_2\left(\frac{1}{\Delta_{\min}}\right) \left\lceil C_3 \frac{(m+n)}{\Delta_{\min}^2} \log^3\left(\frac{(m+n)\lceil \log_2(1/\Delta_{\min}) \rceil^2}{\Delta_{\min}^2 \delta}\right) \right\rceil \\
&\leq \psi(n, m, \delta) := C_4 \frac{(m+n)\log(e/\Delta_{\min})}{\Delta_{\min}^2} \log^3\left(\frac{e(m+n)\log(e/\Delta_{\min})}{\Delta_{\min}\delta}\right)
\end{aligned}$$

where we recall the definition of T_ℓ in (55), that $\delta_\ell = \delta/\ell^2$, and where C_4 is a large enough universal constant. Hence, we have

$$\mathbb{P}\left((i_\tau, j_\tau) = (i^\star, j^\star), \tau \leq \psi(n, m, \delta)\right) \geq \mathbb{P}\left(\bigcap_{\ell=1}^{\lceil \log_2(1/\Delta_{\min}) \rceil} \mathcal{E}_\ell\right) \geq 1 - \delta. \tag{56}$$

This conclude the proof of the guarantee for the best entry identification. Note that we can immediately conclude from the above guarantee (56) that the sample complexity of SME-AE$(1/T^\alpha)$ for all $T \geq 1$, satisfies $\mathbb{E}[\tau \wedge T] \leq \psi(n, m, T^{-\alpha}) + T^{1-\alpha}$. Indeed, we have

$$\begin{aligned}
\mathbb{E}[\tau \wedge T] &= \mathbb{E}[(\tau \wedge T)\mathbb{1}_{\{\tau \leq \psi(n,m,T^{-\alpha})\}}] + \mathbb{E}[(\tau \wedge T)\mathbb{1}_{\{\tau > \psi(n,m,T^{-\alpha})\}}] \\
&\leq \psi(n, m, T^{-\alpha}) + T\mathbb{P}(\tau > \psi(n, m, T^{-\alpha})) \\
&\leq \psi(n, m, T^{-\alpha}) + T^{1-\alpha},
\end{aligned}$$

where the upper bound on the probability follows from (56) with $\delta = 1/T^\alpha$.

Next, we turn our attention to proving the regret upper bound. We define $\mathcal{E}_{good} = \{(\hat{\imath}_\tau, \hat{\jmath}_\tau) = (i^\star, j^\star), \tau \le \psi(n, m, 1/T^2)\}$. We have

$$\begin{aligned}
R^\pi(T) &= TM_{i^\star, j^\star} - \mathbb{E}\left[\sum_{t=1}^T M_{i_t^\pi, j_t^\pi}\right] \\
&= \mathbb{E}\left[\sum_{t=1}^T (M_{i^\star, j^\star} - M_{i_t^\pi, j_t^\pi})\mathbb{1}_{\{\mathcal{E}_{good}\}}\right] + \mathbb{E}\left[\sum_{t=1}^T (M_{i^\star, j^\star} - M_{i_t^\pi, j_t^\pi})\mathbb{1}_{\{\mathcal{E}_{good}^c\}}\right] \\
&\le \mathbb{E}\left[\sum_{t=1}^T (M_{i^\star, j^\star} - M_{i_t^\pi, j_t^\pi})\mathbb{1}_{\{\tau \le \psi(n,m,T^{-2})\}}\right] + \Delta_{\max} T \mathbb{P}(\mathcal{E}_{good}^c) \\
&\le \mathbb{E}\left[\sum_{t=1}^\infty (M_{i^\star, j^\star} - M_{i_t^\pi, j_t^\pi})\mathbb{1}_{\{\tau \wedge \psi(n,m,T^{-2}) > t\}}\right] + \frac{\Delta_{\max}}{T} \\
&\le \mathbb{E}\left[\sum_{t=1}^\infty \bar{\Delta}\mathbb{1}_{\{\tau \wedge \psi(n,m,T^{-2}) > t\}}\right] + \frac{\Delta_{\max}}{T} \\
&\le \bar{\Delta}\psi(n, m, T^{-2}) + \frac{\Delta_{\max}}{T}
\end{aligned}$$

where in the second to last inequality, we used the tower rule together with the observation that $\mathbb{E}[(M_{i^\star, j^\star} - M_{i_t^\pi, j_t^\pi})\mathbb{1}_{\{\tau \wedge \psi(n,m,T^{-2}) > t\}} | \mathcal{F}_{t-1}] = \bar{\Delta}\mathbb{1}_{\{\tau \wedge \psi(n,m,T^{-2}) > t\}}$ where $\mathcal{F}_{t-1}$ is the σ-algebra defined by the observations up to time $t-1$. This concludes the proof. □

G.2 Gap-independent guarantees

An immediate consequence of the regret bound in Theorem 7 is that we can have a gap-independent bound under some additional assumption. Let us define $\zeta = \Delta_{\max}/\Delta_{\min}$, then the regret bound becomes

$$R^\pi(T) \le \frac{\zeta C_4 (m+n)\log(e/\Delta_{\min})}{\Delta_{\min}} \log^3\left(\frac{e(m+n)\log(e/\Delta_{\min})\, T^2}{\Delta_{\min}}\right) + \frac{\Delta_{\max}}{T}. \tag{57}$$

At the same time, we also have the worst case bound

$$R^\pi(T) \le \zeta \Delta_{\min} T. \tag{58}$$

Taking the best of the two bounds (57) and (58) with the worst case choice for $\Delta_{\min}$, we get

$$R^\pi(T) = \tilde{O}\left(\zeta\sqrt{(n+m)T}\log^2((n+m)T)\right)$$

where the $\tilde{O}$ hides additional log-log terms in m, n and T.

H Related work

In this section, we first discuss the results for the estimation of a low-rank transition matrix presented in [63]. We then give a more detailed account of the related work for low-rank bandits.

H.1 Low-rank transition matrix estimation

In [63], the authors try to estimate a low-rank transition matrix from the data consisting of a single trajectory of the corresponding Markov chain. In a sense, this objective is similar to ours in Model II(b). The main results of [63] are presented in Theorem 1. First observe that our results are more precise since we manage to get entry-wise guarantees. Then it is also worth noting that, in the case of homogenous transition matrices, the upper bound on $\|\widehat{P} - P\|_{1\to\infty}$ stated in Theorem 1 in [63] are similar to the upper bounds we establish in Corollary 6. However, to obtain such bounds, we believe that it is necessary to first derive guarantees for the singular subspace recovery in the $\ell_{2\to\infty}$ norm, as we do. The authors of [63] do not present any step with such guarantees for the estimation of the singular subspaces. We explain below why this step is missing and where the analysis towards the upper bound $\|\widehat{P} - P\|_{1\to\infty}$ breaks in [63].

Proof of the guarantees for $\|\widehat{P} - P\|_{1\to\infty}$ in [63]. Note that in [63], the authors use F in lieu of M. We keep our notation M below to be consistent with the rest of the manuscript. In the proof of Theorem 1 in [63], the authors use the following decomposition:

$$\|\widehat{M}_{i,:} - M_{i,:}\| \le \|(\widehat{M}_{i,:} - M_{i,:})V\| + (\|\widehat{M}_{i,:} - M_{i,:}\| + \|M_{i,:}\|)\frac{C\|\widetilde{M} - M\|}{\sigma_r(M)}. \tag{59}$$

They apply concentration results on $\|(\widetilde{M} - M)V\|_{2\to\infty}$ (Lemma 8) and $\|M - \widetilde{M}\|$ (Lemma 7) to bound the two terms from above. More precisely, their proof includes (33) page 3217, a sequence of inequalities where these concentration results are used. In the fifth line of (33), the authors apply (31), the concentration result on $\|(\widetilde{M} - M)V\|_{2\to\infty}$, but to bound $\|(\widehat{M} - M)V\|_{2\to\infty}$ instead. Replacing $\widehat{M}$ by $\widetilde{M}$ is however not possible, and the analysis breaks here.

Is there a simple solution? We argue below that it is not easy to solve the aforementioned issue in the proof. We first claim that the two concentration bounds on $\|(\widetilde{M} - M)V\|_{2\to\infty}$ and $\|M - \widetilde{M}\|$ are not sufficient for bounding the first term from Equation (59). Specifically, for any row i:

$$\|(\widehat{M}_{i,:} - M_{i,:})V\| = \|(\widetilde{M}_{i,:}\widehat{V}\widehat{V}^\top - M_{i,:})V\| = \|(\widetilde{M}_{i,:} - M_{i,:})V + \widetilde{M}_{i,:}(\widehat{V}\widehat{V}^\top - VV^\top)V\|,$$

and in order to analyze the second term inside the norm, we need to deal with dependence between $\widetilde{M}$ and $\widehat{V}$. Doing this naively using the triangle inequality and Cauchy-Schwarz inequality yields:

$$\begin{aligned} \|(\widehat{M} - M)V\|_{2\to\infty} &\le \|(\widetilde{M} - M)V\|_{2\to\infty} + \|\widetilde{M}(\widehat{V}\widehat{V}^\top - VV^\top)V\|_{2\to\infty} \\ &\le \|(\widetilde{M} - M)V\|_{2\to\infty} + \|\widetilde{M}\|_{1\to\infty}\|V - \widehat{V}(\widehat{V}^\top V)\|_{2\to\infty}. \end{aligned} \tag{60}$$

It is not clear how bounds on $\|(\widetilde{M} - M)V\|_{2\to\infty}$ and $\|M - \widetilde{M}\|$ imply a bound on $\|(\widehat{M} - M)V\|_{2\to\infty}$ since term $\|V - \widehat{V}(\widehat{V}^\top V)\|_{2\to\infty}$ does not seem to be directly bounded by these two terms. We can think of bounding $\|V - \widehat{V}(\widehat{V}^\top V)\|_2$ using Davis-Kahan's inequality:

$$\|V - \widehat{V}(\widehat{V}^\top V)\|_{2\to\infty} \le \|V - \widehat{V}(\widehat{V}^\top V)\|_2 \lesssim \frac{\|M - \widetilde{M}\|}{\sigma_r(M)},$$

where we neglect the higher order term (see Equations (47),(48),(49)). Then, with the upper bound on $\|M - \widetilde{M}\|$, we may obtain an upper bound on $\|\widehat{P} - P\|_{1\to\infty}$ but that does not have a fast decaying rate as that claimed in Theorem 1 in [63] or in our main theorems.

It is also worth noting that assuming proof of Theorem 1 in [63] holds or that more specifically, the series of inequalities leading to Equation (33) holds, one could greatly simplify the singular subspace recovery problem. In particular, since

$$\|\widetilde{M}(V - \widehat{V}\widehat{V}^\top V)\|_{2\to\infty} = \|(\widetilde{M} - \widehat{M})V\|_{2\to\infty} \le \|(\widehat{M} - M)V\|_{2\to\infty} + \|EV\|_{2\to\infty}$$

we can rewrite Equation (36) (wlog for symmetric matrix M with eigenvector matrix V) as:

$$\|V - \widehat{V}(\widehat{V}^\top V)\|_{2\to\infty} \\ \leq \frac{1}{\sigma_r(M)} \left((2 + \frac{4\|E\|}{\sigma_r(M)})\|EV\|_{2\to\infty} + 2\|(\widehat{M} - M)V\|_{2\to\infty} + \frac{4\|MV\|_{2\to\infty}\|E\|}{\sigma_r(M)} \right). \quad (61)$$

Now if (33) in [63] was true, we could use the correspoding bound of the critical term $\|(\widehat{M} - M)V\|_{2\to\infty}$. This would not only greatly simplify proofs given in literature based on leave-one-out-technique, but also extend their work to Markov dependent random variables (which has not been done before). Lastly, note that we cannot skip estimation of singular subspaces by combining Equation (60) and (61) since inequality $2\|\widetilde{M}\|_{1\to\infty} < \sigma_r(M)$ does not hold in general.

H.2 Low rank bandits

Here we survey models for low-rank bandits that have emerged recently in the literature but that are not directly related to our model. Nonetheless our guarantees can be exported there.

[31] considers a bi-linear bandit model which seems more general than that of considered [7]. Indeed, they assume that the observed reward in round t after selecting a pair $(x, z) \in \mathcal{X} \times \mathcal{Z}$, is $x^\top \Theta z + \xi_t$ where $\mathcal{X} \subset \mathbb{R}^m$ and $\mathcal{Z} \subset \mathbb{R}^n$ are finite. They assume that $\Theta \in \mathbb{R}^{m\times n}$ is low rank. If we assume that $\mathcal{X} = \{e_1, \dots, e_m\}$ and $\mathcal{Z} = \{e_1, \dots, e_n\}$, we then recover our model and that of [7] with $M = \Theta$. However, we can also argue that if we restrict our attention to m vectors from $\mathcal{X}$, say $\mathcal{X}' = \{x_1, \dots, x_m\}$, that span $\mathbb{R}^m$, and n vectors from $\mathcal{Z}$, say $\mathcal{Z}' = \{z_1, \dots, z_n\}$, that span $\mathbb{R}^n$, then in our model and that of [7], $M_{i,j} = x_i^\top \Theta z_j$, for all $(i, j) \in [m] \times [n]$. Note that in this case, the rank of M is equal to that of Θ. In fact, in the first phase of the algorithm proposed by [31], the authors also restrict their attention to sets $\mathcal{X}'$ and $\mathcal{Z}'$ such that $\lambda_{\min}(\sum_{i=1}^m x_i x_i^\top)$ and $\lambda_{\min}(\sum_{i=1}^m z_i z_i^\top)$ are maximized. To simplify our exposition, we do not use the model presented by [31], instead we use that of [7].

[34] considers a generalized bandit framework with low rank structure which is rather different than the bandit framework we consider. There, the algorithm is based on the two stage idea introduced by [31], which consists in first estimating the subspace, then reducing the problem to a low-dimensional linear bandit with ambient dimension nm but with roughly $n + m$ relevant dimensions. They are able obtain a minimax regret scaling as $(n + m)\sqrt{T}$. It is worth noting that both these works do not have gap-dependent bounds.

[27] is another relevant work. There, the authors consider a low-rank bandit problem similar to ours but slightly more restrictive. At time t, they recommend an arm $\rho(j)_t$ for each user j, and they observe the corresponds rewards. In other words they observe m entries per round, while in our case we only observe one entry per round. They show that with an explore-then-commit algorithm, they attain a regret of order polylog$(n + m)T^{2/3}$. Their regret guarantees require an entry-wise matrix estimation guarantee with scaling comparable to ours. They use the result of Chen et al. [16] which again is only valid for independent entries and does not account for repetitive sampling. To remedy that they rely on ad-hoc pre-processing steps (see remarks 2, 3 and 4 in [27]). In our case, we believe that our matrix estimation guarantees can be immediately used in their setting and this would lead to a regret scaling of order $(n + m)^{1/3}T^{2/3}$ with the more reasonable constraint that we only observe one entry at each round. The authors also obtain an polylog$(n + m)\sqrt{T}$ guarantee but for rank-1 reward matrices only.