Low-Rank Bandits via Tight Two-to-Infinity Singular Subspace Recovery

Yassir Jedra [* 1]   **William Réveillard** [* 2]   **Stefan Stojanovic** [* 2]   **Alexandre Proutiere** [2]

Abstract

We study contextual bandits with low-rank structure where, in each round, if the (context, arm) pair $(i, j) \in [m] \times [n]$ is selected, the learner observes a noisy sample of the (i, j)-th entry of an unknown low-rank reward matrix. Successive contexts are generated randomly in an i.i.d. manner and are revealed to the learner. For such bandits, we present efficient algorithms for policy evaluation, best policy identification and regret minimization. For policy evaluation and best policy identification, we show that our algorithms are nearly minimax optimal. For instance, the number of samples required to return an ε-optimal policy with probability at least $1 - \delta$ typically[1] scales as $\frac{m+n}{\varepsilon^2} \log(1/\delta)$. Our regret minimization algorithm enjoys minimax guarantees typically[1] scaling as $r^{5/4}(m + n)^{3/4}\sqrt{T}$, which improves over existing algorithms. All the proposed algorithms consist of two phases: they first leverage spectral methods to estimate the left and right singular subspaces of the low-rank reward matrix. We show that these estimates enjoy tight error guarantees in the two-to-infinity norm. This in turn allows us to reformulate our problems as a misspecified linear bandit problem with dimension roughly $r(m + n)$ and misspecification controlled by the subspace recovery error, as well as to design the second phase of our algorithms efficiently.

1. Introduction

Stochastic Multi-Armed Bandits (MABs) (Lai & Robbins, 1985) provide an efficient and natural framework for the analysis of the exploration-exploitation trade-off in sequential decision making and have been used in a variety of applications ranging from recommendation systems to clinical trials. When designing bandit algorithms, it is crucial to leverage as much as possible all information initially available to the learner. This information often comes as structural properties satisfied by the mapping of the arms to their rewards, see e.g. linear (Dani et al., 2008), convex (Agarwal et al., 2011), or unimodal (Combes & Proutiere, 2014). In this paper, we investigate bandit problems where the arm-to-reward mapping exhibits a *low-rank structure*. Such a structure has received a lot of attention recently (Kveton et al., 2017; Jun et al., 2019; Jang et al., 2021; Bayati et al., 2022; Kang et al., 2022; Pal & Jain, 2022; Lee et al., 2023). Despite these efforts, it remains unclear how one may optimally exploit low-rank structures and how much gains in terms of performance they bring. We make progress towards answering these questions.

We consider the following contextual low-rank bandit model, similar to that studied in (Bayati et al., 2022; Pal & Jain, 2022; Lee et al., 2023). In each round, the learner first observes a random context $i \in [m]$, and then selects an arm $j \in [n]$. When the pair (i, j) is selected, the learner observes a noisy sample of the (i, j)-th entry of a rank-r matrix M, where r is typically assumed to be much smaller than m and n. For this low-rank bandit model, we address three learning tasks: policy evaluation, best policy identification, and regret minimization. Our contributions are as follows:

(a) We propose a generic method to design two-phase algorithms for low-rank bandit problems. In the first phase, we estimate the singular subspaces of the reward matrix M using simple spectral methods. We establish tight upper bounds on the subspace recovery error in the two-to-infinity norm. In turn, these new error bounds allow us to reformulate our low-rank bandits as a misspecified contextual linear bandit problem with dimension $r(m+n) - r^2$ and controlled misspecification. In the second phase, the algorithm solves the resulting misspecified contextual linear bandit problem. The main contribution of this paper is to demonstrate that the above method, based on tight subspace recovery guarantees in the two-to-infinity norm, yields computationally efficient algorithms that either are nearly minimax optimal or outperform existing algorithms for the three considered learning tasks.

---

[*]Equal contribution [1]Laboratory for Information and Decision Systems, MIT, Cambridge, MA, USA [2]Division of Decision and Control Systems, KTH, Stockholm, Sweden. Correspondence to: Yassir, Jedra <jedra@mit.edu>.

[1]By "typically" here, we mean when the matrix M and the (context, arm) distribution are homogeneous (refer to Definition 3.1 for details). Note that under homogeneity of M, $m = \Theta(n)$.

(b) Policy Evaluation (PE). For this task, the objective is to estimate the average reward of a given *target* policy based on data generated from a fixed behavior policy. We derive instance-specific and minimax lower bounds on the sample complexity satisfied by any (ε, δ)-PAC algorithms[2]. The latter typically[1] scales as $\frac{m+n}{\varepsilon^2} \log(1/\delta)$. We leverage our method to devise SIPS (Spectral Importance Propensity Score) and RS-PE (Recover Subspace for Policy Evaluation), two PE algorithms with nearly minimax optimal sample complexity. The second phase of RS-PE refines the estimate of the reward matrix, using a regularized least-square-estimator applied to the misspecified linear bandit model obtained in the first phase.

(c) Best Policy Identification (BPI). Here, the goal is to return an approximately optimal policy based on data gathered using a fixed or adaptive sampling strategy. Using the same estimates of the reward matrix as those used in the PE algorithms, we devise SBPI (Spectral Best Policy Identification) and RS-BPI (RS for Best Policy Identification), two (ε, δ)-PAC algorithms with nearly minimax optimal sample complexity, again typically scaling as $\frac{m+n}{\varepsilon^2} \log(1/\delta)$. This significantly improves over existing algorithms whose sample complexity scales as $\frac{m+n}{\varepsilon^{2+r}} \log(1/\delta)$ (Lee et al., 2023).

(d) Regret minimization. The two-phase design method yields RS-RMIN (RS for Regret MINimization), an algorithm with minimax regret guarantees typically scaling as $(m+n)^{3/4}\sqrt{T}$ over T rounds. Surprisingly and to the best of our knowledge, RS-RMIN is the first algorithm enjoying minimax guarantees that are always strictly tighter than those achieved in unstructured bandits (in this case, the best guarantees scale as $\sqrt{mnT}$). In its second phase, RS-RMIN uses an extension of SupLin-UCB (Chu et al., 2011) to solve the misspecified linear bandit problem derived at the end of the first phase.

Notation. For a given matrix $M \in \mathbb{R}^{m \times n}$, we denote its i-th row by $M_{i,:}$, its j-th column by $M_{:,j}$, and its (i,j)-th entry by $M_{i,j}$. We denote by $\|M\|_{\mathrm{op}}$ its Euclidean operator norm, by $\|M\|_{\mathrm{F}}$ its Frobenius norm, by $\|M\|_{2\to\infty} = \max_{i \in [m]} \|M_{i,:}\|_2$ its two-to-infinity norm, by $\|M\|_{\max} = \max_{(i,j) \in [m] \times [n]} |M_{i,j}|$ its max-norm, by $\lambda_{\max}(M)$ (resp. $\lambda_{\min}(M)$) its maximal (resp. minimal) eigenvalue. For a given vector x, we denote its Euclidean norm by $\|x\|_2$, and by $\|x\|_{\Lambda} = \sqrt{x^\top \Lambda x}$ its Euclidean norm weighted by some positive definite matrix Λ. The notation $f(x) \lesssim g(x)$ (resp. $f(x) \gtrsim g(x)$) means that there exists a universal constant $C > 0$ such that $f(x) \leq Cg(x)$ (resp. $f(x) \geq Cg(x)$) for all x. We write $f(x) = \Theta(g(x))$ when $f(x) \lesssim g(x)$ and $f(x) \gtrsim g(x)$. We write $\mathrm{poly}(x)$ to denote a quantity that is upper bounded by a polyno-

---

[2]These algorithms return an ε-accurate estimate of the value of the policy with certainty level $1 - \delta$.

mial function of x. We also use $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$. Finally, for any $p, q \in (0, 1)$, we define $\mathrm{kl}(p, q) = p \log(p/q) + (1 - p) \log((1 - p)/(1 - q))$ as the KL-divergence between two Bernoulli distributions with mean p and q, respectively.

2. Related Work

In this section, we discuss existing results for the three investigated learning tasks in contextual low-rank bandits, as well as for singular subspace recovery for low-rank matrices. Additional related work can be found in Appendix A (there, we discuss misspecified linear bandits and other models related to low-rank bandits).

Contextual low-rank bandits. There has been some interest in settings similar to ours (Gentile et al., 2014; Gopalan et al., 2016; Pal & Jain, 2022; Lee et al., 2023; Pal et al., 2023), although mostly in the context of regret minimization. Nonetheless, progress on minimax guarantees has remained surprisingly limited. For example, (Pal & Jain, 2022) proposed OCTAL, an algorithm that achieves a regret of order $O(\mathrm{polylog}(m + n)\sqrt{T})$ for rank-1 bandits, and a simple ETC-based algorithm achieving a regret of $O(\mathrm{polylog}(m + n)T^{3/4})$ for general rank r. However, their results assume that the learner observes m entries per round, while in our setting, only one entry is observed. Recently, (Lee et al., 2023) considered a setting closely related to ours, and referred to as *context-lumpable bandits*. There, the rows of the reward matrix can be clustered into r groups, within which the rows are identical. They establish minimax regret upper bounds of order $\widetilde{O}(\sqrt{r^3(m + n)T})$. They further extend their results to contextual low-rank bandits, but for these bandits, prove a regret upper bound of order $\widetilde{O}((m + n)^{\frac{1}{3r+2}} T^{\frac{3r+1}{3r+2}})$ (see Theorem 26 in (Lee et al., 2023)). A cluster-like structure comparable to that of (Lee et al., 2023) was also considered in (Pal et al., 2023), where the authors proposed an algorithm that attains a regret of order $\tilde{O}(\sqrt{\mathrm{poly}(r)(m + n)T})$. These works leave the existence of an algorithm with minimax regret of order $\widetilde{O}(\sqrt{\mathrm{poly}(r)(m + n)T})$ in contextual low-rank bandit as an open question. We do not fully answer this question, but propose an algorithm with regret scaling as $\tilde{O}(r^{5/4}(m + n)^{3/4}\sqrt{T})$, which improves over existing algorithms.

For the BPI task, (Lee et al., 2023) exhibits sample complexity guarantees of order $\widetilde{O}(r(m + n)/\varepsilon^2)$ in the context-lumpable case. For contextual low-rank bandits, the authors present an algorithm with sample complexity guarantees of order $\widetilde{O}((m + n)/\varepsilon^{2+r})$ (see their Theorem 25 and the discussion thereafter). We get guarantees similar to those they obtain for context-lumpable bandits, but for more general low-rank bandits. We finally mention the work of (Xi et al.,

2022) addressing the PE task in low-rank MDPs. When applied to contextual low-rank bandits (see their Subsection 4.3.2), the authors obtain a sharp estimation error term with an additional bias that does not vanish unless the behavior and target policy are identical; an assumption that we do not require.

Singular subspace recovery in the two-to-infinity norm. It is not surprising that estimating the singular subspaces of M is useful in the context of bandit problems with low-rank structure as it has been showcased in (Jun et al., 2019; Lale et al., 2019; Lu et al., 2021; Kang et al., 2022). However, to the best of our knowledge, only Frobenius norm guarantees have been used, and guarantees in the two-to-infinity norm have remained largely unexplored. The latter guarantees are often harder to obtain and progress towards obtaining them have only emerged recently (Eldridge et al., 2018; Fan et al., 2018; Cape et al., 2019; Abbe et al., 2020). Typically, under standard assumptions, such as bounded incoherence (Candès & Tao, 2010; Recht, 2011), these guarantees state that, for a matrix of size $m \times n$, the subspace estimation error in the two-to-infinity norm is smaller by a factor of $\sqrt{m+n}$ than the achievable bounds on the estimation error in the Frobenius norm. For this reason, we say that the subspace recovery error exhibits a *delocalization* phenomenon (Rudelson & Vershynin, 2015), i.e., the estimation error is spread out across $m + n$ directions (see also the survey (Chen et al., 2021) and references therein).

The major difficulty towards using two-to-infinity norm guarantees is that their derivation requires stringent independence assumptions. Recently, (Stojanovic et al., 2024) provided tools to relax such assumptions and to accommodate scenarios that are suitable for sequential decision making problems such as Markov decision processes and bandits. We make use of these tools to obtain tight guarantees on the singular subspace recovery in the two-to-infinity norm for low-rank bandits. These guarantees enable an effective reduction of the contextual low-rank bandit problem to the misspecified contextual linear bandit problem.

3. Model and Learning Objectives

We consider a stochastic bandit problem with low-rank structure. Specifically, we assume that the expected rewards are parametrized by a matrix $M \in \mathbb{R}^{m \times n}$ which has low rank $r \ll \min(m,n)$. Its SVD is $M = U\Sigma V^\top$, where $U \in \mathbb{R}^{m \times r}$ (resp. $V \in \mathbb{R}^{n \times r}$) contains the left (resp. right) singular vectors, and where $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_r)$ with $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r$. The incoherence parameters of M are defined as $\mu(U) = \sqrt{m/r}\|U\|_{2\to\infty}$ and $\mu(V) = \sqrt{n/r}\|V\|_{2\to\infty}$. We define $\mu = \mu(U) \vee \mu(V)$, and denote by $\kappa = \sigma_1/\sigma_r$ the condition number of M.

Contextual low-rank bandits. In each round $t \geq 1$, the learner first observes a context i_t, selected in an i.i.d. manner and with distribution ρ over $[m]$. She then selects an arm $j_t \in [n]$ potentially based on previous observations, and receives a reward $r_t = M_{i_t,j_t} + \xi_t$. The reward matrix M is a priori unknown, and $(\xi_t)_{t\geq 1}$ is a sequence of i.i.d. zero-mean and σ-subgaussian random variables. We denote the (context, arm) distribution by ω.

For a given randomized policy π, we define its policy value as $v^\pi := \sum_{i,j} \omega_{i,j}^\pi M_{i,j}$ with $\omega_{i,j}^\pi := \rho_i \pi(j|i)$ ($\pi(j|i)$ is the probability to select arm j when observing context i). We further define the optimal policy $\pi^\star := \arg\max_\pi v^\pi$ and denote its value by $v^\star$. We consider three learning tasks:

(i) Policy evaluation. Given a target policy π, and assuming data is gathered under a policy[3] π^b, we aim at designing an efficient estimator of its policy value v^π. We say that a PE estimator $\hat{v}^\pi$ is (ε, δ)-PAC if $\mathbb{P}_M(|v^\pi - \hat{v}^\pi| \leq \varepsilon) \geq 1 - \delta$ for every rank-r matrix $M \in \mathbb{R}^{m \times n}$.

(ii) Best policy identification. We aim to design an efficient algorithm to identify an ε-optimal policy. We say that a BPI algorithm is (ε, δ)-PAC if it outputs a policy $\hat{\pi} : [m] \to [n]$ such that $\mathbb{P}_M(v^\star - v^{\hat{\pi}} \leq \varepsilon) \geq 1 - \delta$ for every rank-r matrix $M \in \mathbb{R}^{m \times n}$.

We define the *sample complexity* of a PE estimator or a BPI algorithm as the number of samples required to achieve an (ε, δ)-PAC guarantee.

(iii) Regret minimization. Here, the objective is to minimize the regret. The regret up to round T of a sequential decision algorithm π is defined as $R^\pi(T) = \sum_{t=1}^T \mathbb{E}[M_{i_t, \pi^\star(i_t)} - M_{i_t, j_t^\pi}]$ where j_t^π is the arm selected under algorithm π in round t.

Throughout the paper, we make the assumption that the learner is aware of upper bounds on $\kappa, \mu, \|M\|_{\max}$ and $1/(m \min_{i\in[m]} \rho_i) > 0$. Our results exhibit precise dependencies on all these parameters. Sometimes, to simplify the exposition of our results, we consider that the reward matrix M, the context distribution ρ or the (context, arm) distribution ω are *homogeneous* in the following sense:

Definition 3.1 (Homogeneity)**.** The reward matrix M is *homogeneous* when $m = \Theta(n)$, $r = \Theta(1)$, $\kappa = \Theta(1)$, $\mu = \Theta(1)$. A distribution p on a finite set $\mathcal{I}$ is *homogeneous* when $p_{\min} = \Theta(p_{\max})$ for $p_{\min} = \min_{i\in\mathcal{I}} p_i$, $p_{\max} = \max_{i\in\mathcal{I}} p_i$ [4].

The notion of homogeneity allows us to have meaningful discussions but also constitutes a reasonable assumption.

---

[3] The arm j is selected with probability $\pi^b(j|i)$ when the context i is observed.

[4] This implies $p_{\min} = \Theta(1/|\mathcal{I}|)$, $p_{\max} = \Theta(1/|\mathcal{I}|)$.

In particular, assuming that the matrix M has bounded incoherence entails that the matrix M can be recovered using fewer samples than mn (Candes & Recht, 2012). The bounded incoherence property is also related to the notion of spikiness of a matrix, which has to do with assuming that $\|M\|_{\max}/\sigma_1 = \Theta(1/\sqrt{mn})$ (Negahban & Wainwright, 2012; Mackey et al., 2015)). Indeed, we can verify (see Lemma B.8) that $\frac{1}{\sqrt{mn}} \leq \frac{\|M\|_{\max}}{\sigma_1} \leq \frac{\mu^2 r}{\sqrt{mn}}$. Note that this inequality does not mean that $\|M\|_{\max} = \Theta(1/\sqrt{mn})$ when $\mu = \Theta(1)$, because for most interesting settings σ_1 would scale as $\sqrt{mn}$. A naive example would be to take a matrix M that has all entries equal to 1. Observe then that $\text{rank}(M) = 1$, $\|M\|_{\max} = 1$, $\sigma_1 = \sqrt{mn}$, and M satisfies the bounded incoherence assumption.

4. Singular Subspace Recovery

In the first phase of our algorithms for low-rank bandits, the singular subspaces of M are estimated. In this section, we study the performance of this estimation procedure. We assume that the learner observes T samples of noisy entries of M chosen in an i.i.d. manner according to some distribution ω over $[m] \times [n]$, and that from these observations, she builds estimates of the singular subspaces of M. For simplicity, we assume that ω is known to the learner. This is without loss of generality in our settings[5]. We also assume that $\omega_{\min} = \min_{(i,j) \in [m] \times [n]} \omega_{i,j} > 0$.

4.1. Subspace Estimation

To estimate the singular subspaces of M, we construct $\widetilde{M}$ as: for all $(i, j) \in [m] \times [n]$,

$$\widetilde{M}_{i,j} = \frac{1}{T\omega_{i,j}} \sum_{t=1}^{T} (M_{i_t,j_t} + \xi_t) \mathbb{1}_{\{(i_t,j_t)=(i,j)\}}. \quad (1)$$

Then, we let $\widehat{M}$ be the best r-rank approximation of $\widetilde{M}$. More precisely, we write from the SVD of $\widetilde{M}$:

$$\widehat{M} = \widehat{U}\widehat{\Sigma}\widehat{V}^\top, \quad (2)$$

where $\widehat{\Sigma}$ is the diagonal matrix containing the r largest singular values of $\widetilde{M}$. $\widehat{U} \in \mathbb{R}^{m \times r}$ contains their corresponding r left singular vectors, and $\widehat{V} \in \mathbb{R}^{n \times r}$ their corresponding r right singular vectors. By Eckart-Young-Mirsky's theorem, the matrix $\widehat{M}$ is the best r-rank approximation of $\widetilde{M}$. We use $\widehat{U}$ and $\widehat{V}$ as our estimates of the singular subspaces spanned by U and V, respectively.

In what follows, we denote by $\widehat{U}_\perp \in \mathbb{R}^{m \times (m-r)}$ a matrix made of orthonormal vectors completing $\widehat{U}$ so as to obtain

---

[5]Indeed, we can estimate the context distribution ρ using only $\widetilde{O}(1/\rho_{\min})$ samples, and select j uniformly at random (refer to Appendix I.1)

an orthonormal basis of $\mathbb{R}^m$. $\widehat{V}_\perp \in \mathbb{R}^{n \times (n-r)}$ is constructed similarly.

4.2. Guarantees in the Two-to-Infinity Norm

We wish to examine the performance of our subspace recovery method in the two-to-infinity norm. More precisely, we denote the left and right singular subspace recovery errors in this norm by

$$d_{2\to\infty}(U, \widehat{U}) = \|UU^\top - \widehat{U}\widehat{U}^\top\|_{2\to\infty},$$
$$d_{2\to\infty}(V, \widehat{V}) = \|VV^\top - \widehat{V}\widehat{V}^\top\|_{2\to\infty}.$$

Now, we present the main guarantee on these errors, which is adapted from (Stojanovic et al., 2024).

Theorem 4.1 (Subspace recovery in $\|\cdot\|_{2\to\infty}$). *Let us define* $\epsilon_{\text{Sub-Rec}} := \max(d_{2\to\infty}(U, \widehat{U}), d_{2\to\infty}(V, \widehat{V}))$ *and* $L := \|M\|_{\max} \vee \sigma$. *For any* $\delta \in (0,1)$, *the following event:*

$$\epsilon_{\text{Sub-Rec}} \lesssim \sqrt{\frac{L^2\mu^2\kappa^2 r(m+n)}{\sigma_r^2 T\omega_{\min}(m \wedge n)} \log^3\left(\frac{(m+n)T}{\delta}\right)} \quad (3)$$

holds with probability at least $1 - \delta$, *provided that*

$$T \gtrsim \frac{L^2(m+n)}{\sigma_r^2\omega_{\min}} \log^3\left(\frac{(m+n)T}{\delta}\right). \quad (4)$$

The proof of Theorem 4.1 is sketched in Appendix B. The analysis is technically involved and relies on combining the so-called leave-one-out argument (see (Chen et al., 2021) and references therein) together with a Poisson approximation technique (Stojanovic et al., 2024).

From Theorem 4.1, we can immediately verify that if the matrix M and the sampling distribution ω are homogeneous, then the subspace recovery error bound in the two-to-infinity norm is smaller by a factor of $\sqrt{m+n}$ than any achievable subspace recovery error bound in the Frobenius norm with our estimates of the singular subspaces. Indeed, assume additionally and only for simplicity that $\sigma \lesssim \|M\|_{\max}$, then after only $\widetilde{\Omega}(m+n)$ observations, we can recover the singular subspaces with an error rate in the $\|\cdot\|_{2\to\infty}$ norm that scales as $\widetilde{O}(1/\sqrt{T})$ with high probability. To see that, note that when $\sigma \lesssim \|M\|_{\max}$, then according to Lemma B.8, we have $L/\sigma_r = \Theta(1/\sqrt{mn})$. The recovery error in $\|\cdot\|_F$ would typically scale as $\widetilde{O}(\sqrt{(m+n)/T})$ (see Lemma B.7 which follows from classical arguments as in (Jun et al., 2019)). This suggests that the subspace recovery error (seen as a matrix) is *delocalized*, i.e., spread out along $m + n$ directions. Thus, the provided two-to-infinity norm guarantees offer a finer and more precise control over the subspace recovery error than the Frobenius norm guarantees. We take advantage of this when devising algorithms for contextual low-rank bandits.

It is worth mentioning that the obtained error rate (3) in Theorem 4.1 exhibits a dependence on $L = \|M\|_{\max} \vee \sigma$ instead of only σ. This means that the proposed guarantees do not suggest exact subspace recovery when $\sigma \to 0$. This is to be expected because we rely on a spectral method that consists in truncating the SVD of $\widetilde{M}$. Indeed, such a method typically suffers from a bias that comes from setting unobserved entries to zero, as can be remarked in (1) (see also (Chen et al., 2021)). As we shall see in the coming sections, this affects the final bounds we obtain in all three learning tasks. Nonetheless, we provide a discussion in Appendix I.4 on how we may improve this dependence, which may be of independent interest. Indeed, one may resort to ideas based on nuclear norm penalization which have been shown recently to enjoy tight entry-wise matrix estimation guarantees (Chen et al., 2020). However, one has to take further care to ensure that the results based on such ideas would extend to our dependent noise setting. Finally, let us also mention that the dependence on σ, through L, in the sample complexity (4) provided in Theorem 4.1 is to be expected and cannot be further improved.

5. Reduction to Misspecified Linear Bandits

In this section, we assume that we use T_1 reward observations to estimate the singular subspaces U and V of M as described in the previous section. These T_1 first observations will constitute the first phase of our low-rank bandit algorithms. Based on the estimates of U and V, one could, as in (Jun et al., 2019; Kang et al., 2022), reduce the problem faced by the learner in the remaining $(T - T_1)$ rounds to an *almost low-dimensional linear bandit*; see Appendix H for a detailed discussion. The second phase of low-rank bandit algorithms would then just consist in applying an algorithm for such linear bandits. We explain in Appendix H that for regret minimization, this approach would however not lead to the tightest regret guarantees possible. Instead, we establish that the second phase of low-rank bandit algorithms can be reduced to solving a *misspecified* linear bandit with misspecification controlled using our subspace recovery error guarantees in the two-to-infinity norm. This new approach yields improved regret upper bounds, but also allows us to devise efficient algorithms for PE and BPI. In what follows, we make the reduction to a misspecified linear bandit precise.

First observe that, following a similar transformation of the so-called left and right features as in (Jun et al., 2019), we may express any entry of the matrix M as:

$$M_{ij} = e_i^\top \widehat{U}(\widehat{U}^\top M \widehat{V})\widehat{V}^\top e_j + e_i^\top \widehat{U}(\widehat{U}^\top M \widehat{V}_\perp)\widehat{V}_\perp^\top e_j$$
$$+ e_i^\top \widehat{U}_\perp(\widehat{U}_\perp^\top M \widehat{V})\widehat{V}^\top e_j + e_i^\top \widehat{U}_\perp(\widehat{U}_\perp^\top M \widehat{V}_\perp)\widehat{V}_\perp^\top e_j. \quad (5)$$

From (5), we can write that for any $(i, j) \in [m] \times [n]$,

$$M_{i,j} = \phi_{i,j}^\top \theta + \epsilon_{i,j}, \quad (6)$$

where $\epsilon_{i,j} = e_i^\top \widehat{U}_\perp(\widehat{U}_\perp^\top M \widehat{V}_\perp)\widehat{V}_\perp^\top e_j$, and

$$\phi_{i,j} = \begin{bmatrix} \text{vec}(\widehat{U}^\top e_i e_j^\top \widehat{V}) \\ \text{vec}(\widehat{U}^\top e_i e_j^\top \widehat{V}_\perp) \\ \text{vec}(\widehat{U}_\perp^\top e_i e_j^\top \widehat{V}) \end{bmatrix}, \theta = \begin{bmatrix} \text{vec}(\widehat{U}^\top M \widehat{V}) \\ \text{vec}(\widehat{U}^\top M \widehat{V}_\perp) \\ \text{vec}(\widehat{U}_\perp^\top M \widehat{V}) \end{bmatrix}. \quad (7)$$

With this interpretation, we obtain a misspecified linear bandit of dimension $d := r(m+n) - r^2$. The benefit of this reduction is that we can characterize the misspecification $\epsilon_{\max} := \max_{(i,j) \in [m] \times [n]} |\epsilon_{i,j}|$ using our subspace recovery guarantees in the two-to-infinity norm. Indeed, we establish:

Corollary 5.1 (Misspecification error). *Let $\delta \in (0, 1)$, $T_1 > 0$ be the number of samples observed by the sampling distribution ω. Then, the event*

$$\epsilon_{\max} \lesssim \frac{L^2 \mu^2 \kappa^3 r(m+n)}{\sigma_r T_1 \omega_{\min}(m \wedge n)} \log^3\left(\frac{(m+n)T_1}{\delta}\right)$$

holds with probability at least $1 - \delta$, provided that

$$T_1 \gtrsim \frac{L^2(m+n)}{\sigma_r^2 \omega_{\min}} \log^3\left(\frac{(m+n)T_1}{\delta}\right).$$

Corollary 5.1 follows immediately from Theorem 4.1. To see that, observe that we have, for all $(i, j) \in [m] \times [n]$,

$$|\epsilon_{i,j}| = |e_i^\top \widehat{U}_\perp(\widehat{U}_\perp^\top M \widehat{V}_\perp)\widehat{V}_\perp^\top e_j|$$
$$\overset{(a)}{=} |e_i^\top(UU^\top - \widehat{U}\widehat{U}^\top)M(VV^\top - \widehat{V}\widehat{V}^\top)e_j|$$
$$\overset{(b)}{\leq} d_{2\to\infty}(U, \widehat{U})d_{2\to\infty}(V, \widehat{V})\|M\|_{\text{op}}$$
$$\overset{(c)}{\leq} \epsilon_{\text{Sub-Rec}}^2 \kappa \sigma_r(M) \quad (8)$$

where equality (a) follows from the fact that $\widehat{U}_\perp \widehat{U}_\perp^\top M = (I - \widehat{U}\widehat{U}^\top)M$ since $\widehat{U}\widehat{U}^\top + \widehat{U}_\perp \widehat{U}_\perp^\top = I$, and similarly for $M\widehat{V}_\perp \widehat{V}_\perp^\top$, inequality (b) follows from $|x^\top M y| \leq \|x\|_2 \|y\|_2 \|M\|_{\text{op}}$ for any vectors $x \in \mathbb{R}^m, y \in \mathbb{R}^n$, and inequality (c) follows from the definition of the condition number κ.

We remark that when M and ω are homogeneous, we have $\epsilon_{\max} = \widetilde{O}((m+n)/T_1)$ due to the quadratic dependence on $\epsilon_{\text{Sub-Rec}}$ appearing in (8).

6. Policy Evaluation

In this section, we present algorithms for PE and analyze their performance. The first algorithm, referred to as SIPS (Spectral Importance Propensity Score), consists of a single

phase and exploits all samples to construct a rank-r estimate of M. The value of the target policy is directly obtained using this estimate. We show that, when M and ω are homogeneous, SIPS exhibits a minimax optimal sample complexity w.r.t. the matrix size (m, n) and the accuracy level ε, but not w.r.t. the confidence level δ. The design of the second algorithm, RS-PE (Recover Subspace for Policy Evaluation), leverages our two-phase approach. In its second phase, RS-PE refines the estimate of M via a regularized least squares estimator applied to the misspecified linear bandit model obtained at the end of the first phase. Applying this two-phase approach has significant benefits: (i) RS-PE enjoys instance-dependent sample complexity guarantees; (ii) When M and ω are homogeneous, it is nearly minimax optimal w.r.t. the parameters m, n, ε and δ; (iii) as shown in Appendix J.1, it outperforms SIPS in numerical experiments. We conclude this section by deriving instance-dependent and minimax sample complexity lower bounds.

Throughout this section, we assume that the observations are gathered in an i.i.d. manner using a behavior policy π^b, and we wish to estimate the value v^π of a target policy π. To simplify the notation, we abbreviate the (context, arm) distribution ω^{π^b} by ω.

6.1. Spectral Importance Propensity Score

A natural approach to PE consists in directly constructing $\widehat{M}$, a rank-r estimate of M obtained via the spectral method described in (2), using all of the T available samples gathered under the (context, arm) distribution ω. We estimate the value of the target policy π by:

$$\hat{v}^\pi_{\texttt{SIPS}} := \sum_{(i,j) \in [m] \times [n]} \omega^\pi_{i,j} \widehat{M}_{i,j}. \quad (9)$$

The estimator $\hat{v}^\pi_{\texttt{SIPS}}$ can be interpreted as an Importance Propensity Score (IPS) estimator that takes into account the low-rank structure of M. Indeed, in the absence of such low-rank structure, i.e., $r = m \wedge n$, we see that our policy value estimator reduces to the classical IPS estimator (Wang et al., 2017).

Theorem 6.1. *Let $\varepsilon \in (0, \|M\|_{\max})$ and $\delta \in (0, 1)$. For any target policy π, $|v^\pi - \hat{v}^\pi_{\texttt{SIPS}}| \leq \varepsilon$ with probability larger than $1 - \delta$ as soon as*

$$T \gtrsim \frac{L^2 \mu^6 \kappa^4 r^3 (m+n)}{\omega_{\min}(m \wedge n)^2 \varepsilon^2} \log^3\left(\frac{(m+n)T}{\delta}\right).$$

The proof of this theorem, presented in Appendix D.1, relies on deriving entry-wise guarantees for the estimate $\widehat{M}$. These guarantees follow from the two-to-infinity guarantees provided in Theorem 4.1. Observe that when both M and ω are homogeneous, the sample complexity guarantees

derived in Theorem 6.1 scale as $\frac{L^2(m+n)}{\varepsilon^2} \log^3(1/\delta)$. The dependence in L is a consequence of the bias induced by setting the unobserved entries of M to 0 in the definition (1) of $\widetilde{M}$, as explained in Section 4.2.

Note that a naive control by the Frobenius norm would not be sufficient to obtain an error guarantee that scales with $\sqrt{m+n}$. Indeed, by the Cauchy-Schwartz inequality, $|v^\pi - \hat{v}^\pi_{\texttt{SIPS}}| \leq \|\omega^\pi\|_F \|M - \widehat{M}\|_F$. If ω is homogeneous, $\|M - \widehat{M}\|_F$ typically scales with $\sqrt{mn(m+n)/T}$, and $\|\omega^\pi\|_F = \Theta(1/\sqrt{m})$ when π is deterministic. We consequently obtain $|v^\pi - \hat{v}^\pi_{\texttt{SIPS}}| \lesssim \sqrt{n(m+n)/T}$, which does not improve over existing PE error bounds for contextual bandits with no structure (Yin & Wang, 2020).

6.2. Algorithm via the Two-Phase Approach

Next, we present RS-PE (Recover Subspace for Policy Evaluation), a two-phase algorithm that proceeds as follows. (i) In the first phase, we recover the subspaces by using the first T_1 samples to construct the estimates $\widehat{U}$ and $\widehat{V}$ as described in (5). (ii) In the second phase, using $\widehat{U}$ and $\widehat{V}$, we reduce the problem to a misspecified linear bandit as described in (6), and then run least squares estimation with the remaining $T - T_1$ samples to construct

$$\widehat{\theta} = \left(\sum_{t=t_1}^{T} \phi_{i_t,j_t} \phi_{i_t,j_t}^\top + \tau I_d\right)^{-1} \left(\sum_{t=t_1}^{T} r_t \phi_{i_t,j_t}\right), \quad (10)$$

where $\tau > 0$ is a regularization parameter, $t_1 = T_1 + 1$, and $d = r(m+n) - r^2$. Finally, the estimated value of the target policy π is:

$$\hat{v}^\pi_{\texttt{RS-PE}} := \sum_{(i,j) \in [m] \times [n]} w^\pi_{i,j} \phi^\top_{i,j} \hat{\theta}. \quad (11)$$

To analyze RS-PE, we introduce the following instance-dependent quantities. Define, for all $(i, j) \in [m] \times [n]$,

$$\psi_{i,j} = \begin{bmatrix} \text{vec}(U^\top e_i e_j^\top V) \\ \text{vec}(U^\top e_i e_j^\top V_\perp) \\ \text{vec}(U_\perp^\top e_i e_j^\top V) \end{bmatrix},$$

and $\psi_\pi = \sum_{(i,j) \in [m] \times [n]} \omega^\pi_{i,j} \psi_{i,j}$.

Theorem 6.2. *Let $\varepsilon \in (0, \|M\|_{\max})$ and $\delta \in (0, 1)$. With the choices $T_1 = \lfloor T/2 \rfloor$ and $\tau \leq r(m \wedge n)^{-1}(\omega_{\min}/\omega_{\max}) \log(16d/\delta)$, for any target policy π, $|v^\pi - \hat{v}^\pi_{\texttt{RS-PE}}| \leq \varepsilon$ with probability larger than $1 - \delta$ as soon as*

$$T \gtrsim \frac{\sigma^2 \|\psi_\pi\|_2^2}{\omega_{\min} \varepsilon^2} \log\left(\frac{e}{\delta}\right) + \frac{K}{\varepsilon^{4/3}} \log^3\left(\frac{(m+n)T}{\delta}\right),$$

where K depends polynomially on the model parameters[6] $\|M\|_{\max}, \sigma, \mu, \kappa, r, m, n, (\omega_{\min} mn)^{-1}$.

---

[6]Refer to (34) in Appendix D.2 for a complete expression.

We present in Appendix D.2 a stronger version of Theorem 6.2 and its proof. The latter relies on intermediate results pertaining to the PE task in contextual linear bandits, for which we present in Appendix C a complete analysis: we derive an instance-dependent sample complexity lower bound and an algorithm matching this limit.

Note that the guarantees for RS-PE are instance-dependent through ψ_π. By slightly altering the proof of Theorem 6.2 and leveraging the upper bound $\|\psi_\pi\|_2^2 \leq \mu^2 r \frac{m+n}{mn}$ (see Appendix D.3 for a proof), we can also derive minimax guarantees with a remainder term that scales with $1/\varepsilon$ instead of $1/\varepsilon^{4/3}$. We refer to Theorem D.9 for the precise statement.

Asymptotically, when ε tends to 0, these guarantees exhibit a better dependence in μ, κ, r, $\|M\|_{\max}$ and δ (and even in (m,n) for non-homogeneous matrices M) compared to those of SIPS: for ε small enough[7], the sample complexity guarantees of RS-PE scale as $\frac{\sigma^2 \mu^2 r(m+n)}{\omega_{\min} mn\varepsilon^2} \log(1/\delta)$. When, in addition, M and ω are homogeneous, they scale as $\frac{\sigma^2(m+n)}{\varepsilon^2} \log(1/\delta)$.

For large ε, the sample complexity guarantees of SIPS presented in Theorem 6.1 may be better than those of RS-PE (we believe however that the analysis of RS-PE can be improved for such ε). Yet, as our experiments (see Appendix J.1) suggest, RS-PE outperforms SIPS numerically even in the large ε regime.

Observe further that in any case, the first term in the sample complexity guarantees of RS-PE scales with σ^2. In contrast, the guarantees of SIPS scale with $L^2 = (\|M\|_{\max} \vee \sigma)^2$ (see Theorem 6.1). The remainder term, however, still depends on L through the factor K.

We finally note that without accounting for the low-rank structure, the minimax sample complexity of any PE algorithm would scale as $\frac{mn}{\varepsilon^2} \log(1/\delta)$ (Yin & Wang, 2020). Exploiting the low-rank structure allows us to replace the factor mn by $(m+n)$.

6.3. Sample Complexity Lower Bounds

We first derive an instance-dependent lower bound on the sample complexity. Consider the following factorization of M: $M = PQ^\top$ with $P \in \mathbb{R}^{m \times r}$ and $Q \in \mathbb{R}^{n \times r}$ (this factorization is not unique). Let us define for every $i \in [m]$,

$$\Lambda_Q^i = \sum_{j=1}^n \omega_{i,j} Q_j Q_j^\top \quad \text{and} \quad Q_\pi^i = \sum_{j=1}^n \omega_{i,j}^\pi Q_j.$$

Similarly, for every $j \in [n]$, let

$$\Lambda_P^j = \sum_{i=1}^m \omega_{i,j} P_i P_i^\top \quad \text{and} \quad P_\pi^j = \sum_{i=1}^m \omega_{i,j}^\pi P_i.$$

[7]Refer to (38) in Appendix D.2 for a precise condition.

In the above definitions, Q_j (resp. P_i) denotes the j-th row vector of Q (resp. i-th row vector of P).

Theorem 6.3. *Let $\varepsilon > 0, \delta \in (0, 1/2)$ and assume that $\xi_t \sim \mathcal{N}\left(0, \sigma^2\right)$. The sample complexity of any (ϵ, δ)-PAC estimator of v^π must satisfy*

$$T \geq \frac{\sigma^2 L_{M,\pi}}{2\varepsilon^2} \, \mathrm{kl}\left(\delta, 1-\delta\right)$$

for $L_{M,\pi} := \max\left(\sum_i \|Q_\pi^i\|_{(\Lambda_Q^i)^{-1}}^2, \sum_j \|P_\pi^j\|_{(\Lambda_P^j)^{-1}}^2\right)$. Furthermore, $L_{M,\pi}$ is independent of the choice of rank factorization $M = PQ^\top$.

From Theorem 6.3, we can deduce a minimax lower bound. For some constants $c = (c_1, c_2)$ and c', we introduce the set of matrices $\mathcal{M}(c, c') = \{M \in \mathbb{R}^{m \times n} : \mathrm{rank}(M) = r \leq c', c_1 m \leq n \leq c_2 m\}$.

Proposition 6.4. *Let $\varepsilon > 0, \delta \in (0, 1/2)$ and assume that $\xi_t \sim \mathcal{N}\left(0, \sigma^2\right)$. There exists a target policy π and a homogeneous matrix $M \in \mathcal{M}(c, c')$ such that the sample complexity of any $(\varepsilon, \delta)-$PAC estimator of v^π must satisfy*

$$T \gtrsim \frac{\sigma^2(m+n)}{\omega_{\max} mn\varepsilon^2} \, \mathrm{kl}\left(\delta, 1-\delta\right).$$

Note that when ω is homogeneous, the sample complexity lower bound reduces to $T \gtrsim \frac{\sigma^2(m+n)}{\varepsilon^2} \, \mathrm{kl}\left(\delta, 1-\delta\right)$. Since $\mathrm{kl}\left(\delta, 1-\delta\right) \geq \log\left(1/\left(2.4\delta\right)\right)$ (Kaufmann et al., 2016), this lower bound nearly matches the sample complexity guarantees of RS-PE when M and ω are homogeneous. If in addition, $\sigma \gtrsim \|M\|_{\max}$, it also matches the sample complexity guarantees of SIPS.

7. Best Policy Identification

In this section, following the same approach as that used to address the PE task, we devise two BPI algorithms with nearly minmax optimal sample complexity. In what follows, we assume that the context distribution ρ is homogeneous. We explain how this condition can be dropped at the reasonable cost of additional logarithmic factors in our bounds in Appendix F.2.

As for the PE learning task, our first algorithm is directly based on the estimate $\widehat{M}$, and the second follows our two-phase approach. These algorithms, referred to as SBPI and RS-BPI, output the policy $\hat{\pi}_{\mathrm{SBPI}}$ and $\hat{\pi}_{\mathrm{RS-BPI}}$ respectively, defined by, for all $i \in [m]$,

$$\hat{\pi}_{\mathrm{SBPI}}(i) := \arg\max_{1 \leq j \leq n} \widehat{M}_{i,j},$$

$$\hat{\pi}_{\mathrm{RS-BPI}}(i) := \arg\max_{1 \leq j \leq n} \phi_{i,j}^T \hat{\theta},$$

where $\widehat{M}$, ϕ and $\hat{\theta}$ are defined in (2), (7) and (10), respectively. The pseudo-code of RS-BPI is presented in Algorithm 1.

Algorithm 1 RECOVER SUBSPACE FOR BEST POLICY IDENTIFICATION (RS-BPI)

Input: Budgets of rounds T and T_1, context distribution ρ, regularization parameter τ, confidence level δ,

Phase 1: Subspace recovery

Collect T samples according to the uniform policy

Use the first T_1 samples to construct $\widehat{U}, \widehat{V}$ as in (2)

Phase 2: Solving a misspecified linear bandit

Construct $\{\phi_{i,j} : (i,j) \in [m] \times [n]\}$ as in (7)

Use the remaining $T - T_1$ samples to construct the least square estimator $\hat{\theta}$ as in (10)

Set $\hat{\pi}_{\text{RS-BPI}}(i) = \arg\max_{1 \le j \le n} \phi_{i,j}^\top \hat{\theta}$ for all $i \in [m]$

Output: $\hat{\pi}_{\text{RS-BPI}}$

Theorem 7.1. *Let $\varepsilon \in (0, \|M\|_{\max})$, $\delta \in (0, 1)$ and assume that ρ is homogeneous.*

(i) $v^\star - v^{\hat{\pi}_{\text{SBPI}}} \le \varepsilon$ with probability larger than $1 - \delta$ as soon as

$$T \gtrsim \frac{L^2 \mu^6 \kappa^4 r^3 (m+n)^3}{(m \wedge n)^2 \varepsilon^2} \log^3\left(\frac{(m+n)T}{\delta}\right).$$

(ii) Choosing $\tau \le r(m \wedge n)^{-1} (\rho_{\min}/\rho_{\max}) \log(16d/\delta)$ and $T_1 = \lfloor T/2 \rfloor$, $v^\star - v^{\hat{\pi}_{\text{RS-BPI}}} \le \varepsilon$ with probability larger than $1 - \delta$ as soon as

$$T \gtrsim \frac{\sigma^2 \mu^2 r(m+n)}{\varepsilon^2} \log\left(\frac{m+n}{\delta}\right) + \frac{K}{\varepsilon} \log^3\left(\frac{(m+n)T}{\delta}\right),$$

where K depends polynomially on the model parameters [8] $\|M\|_{\max}, \sigma, \mu, \kappa, r, m, n$.

Note that under our homogeneity assumption on ρ, the sample complexity guarantees of RS-BPI scale as $\frac{\sigma^2 \mu^2 r(m+n)}{\varepsilon^2} \log(1/\delta)$ when ε is sufficiently small. In this case, it exhibits a better dependence on the model parameters than the sample complexity guarantees of SBPI when M is not homogeneous. For large ε, the guarantees of SBPI can be better than those of RS-BPI. Nonetheless, our experiments (see Appendix J.2) suggest that RS-BPI still outperforms SBPI in the large ε regime.

A known minimax sample complexity lower bound for BPI in contextual low-rank bandits scales as $\frac{r(m+n)}{\varepsilon^2} \log(1/\delta)$ (this is essentially proved by (Lee et al., 2023) for context-lumpable bandits, a subclass of contextual low-rank bandits). Hence, in view of Theorem 7.1, both SBPI and RS-BPI are nearly minimax optimal when the reward matrix and context distribution are homogeneous; the sample complexity of SBPI has an additional $\log^2(1/\delta)$ term, and RS-BPI is minimax optimal when ε is sufficiently small.

Finally, it is worth noting that for the two proposed algorithms, data is simply gathered using the uniform sampling

policy (there is no need to be adaptive to achieve a minimax optimal sample complexity when M and ρ are homogeneous).

8. Regret Minimization

In this section, we introduce RS-RMIN, our algorithm designed for regret minimization in contextual low-rank bandits. Presented as Algorithm 2, RS-RMIN runs in two phases. In the first phase, we collect T_1 samples by simply selecting arms uniformly at random. Based on the corresponding reward observations, we estimate the singular subspaces of M using the procedure outlined in Subsection 4.1. The problem is then reformulated as a misspecified contextual linear bandit, as described in Section 5.

In the second phase, to solve the misspecified linear bandit for the remaining $T - T_1$ rounds, RS-RMIN uses a variant of SUPLINUCB (Takemura et al., 2021), outlined in Algorithm 5 in Appendix G.1. In each round t, the set of feature vectors $\mathcal{X}_t = \{\phi_{i_t,j}, j \in [n]\}$ is constructed using the estimated singular subspaces as defined in (7), with features of dimension $d = r(m+n) - r^2$.

The choice of T_1 will be specified later. Let $T_2 = T - T_1$. We define the threshold $\beta(\delta)$ used in SUPLINUCB as:

$$\beta(\delta) = \sigma\left(1 + \sqrt{2 \log\left(\frac{T_2 mn}{\delta} \left\lceil \frac{1}{2} \log\left(\frac{T_2}{d}\right)\right\rceil\right)}\right). \quad (12)$$

Theorem 8.1. *Suppose[9] $T_1 = \widetilde{\Theta}(f(T, M, \rho))$, where*

$$f(T, M, \rho) := L \mu^2 \kappa^2 r^{5/4} \frac{(m+n)^{3/4} (mn)^{1/4}}{\sqrt{m \rho_{\min}} (m \wedge n)^{1/2}} \sqrt{T}.$$

Then, the regret of $\pi = $ RS-RMIN satisfies: $R^\pi(T) = \widetilde{O}(f(T, M, \rho))$, for all $T \ge 1$.

Here, $\widetilde{\Theta}(\cdot), \widetilde{O}(\cdot)$ may hide logarithmic factors in T, m, n and L/σ. Theorem 8.1 is proved in Appendix G. When the reward matrix M and context distribution ρ are homogeneous, the regret upper bound of RS-RMIN simplifies to:

$$R^\pi(T) = \widetilde{O}\left(L r^{5/4} n^{3/4} \sqrt{T}\right).$$

Observe that if we were to use an algorithm meant for contextual bandits with no structure, then the best regret guarantee we can obtain is $\widetilde{O}(\sqrt{mnT})$ (Lattimore & Szepesvári, 2020). In contrast, our regret upper bound scales as $\widetilde{O}(n^{3/4}\sqrt{T})$, and clearly $n^{3/4}\sqrt{T} \ll n\sqrt{T} = \Theta(\sqrt{mnT})$ for all $T \ge 1$ whenever $m = \Theta(n)$. In other words, our algorithm truly takes advantage of the low-rank structure. We finally remark that Lee et al. (2023) also provides non-trivial regret bounds for general low-rank reward

---

[8]Refer to (44) in Appendix F for a complete expression.

[9]Refer to (51) in Appendix G for an explicit expression.

Algorithm 2 RECOVER SUBSPACE FOR REGRET MINIMIZATION (RS-RMIN)

Input: Budget of rounds T, T_1, context distribution ρ

Phase 1: Subspace recovery

Collect T_1 samples according to the uniform policy

Construct $\widehat{U}, \widehat{V}$ as described in (2)

Phase 2: Solving a misspecified linear bandit

Construct $\{\phi_{i,j} : (i,j) \in [m] \times [n]\}$ as in (7)

Set $T_2 \leftarrow T - T_1$, $\delta \leftarrow 1/T$, $\beta(\delta)$ as in (12),
$\qquad \Lambda \leftarrow \sigma^2/(\|M\|_{\max}^2 mn) I_d$

Run SUPLINUCB for the remaining rounds with inputs T_2, $\beta(\delta)$ and Λ.

---

matrices, as discussed in the introduction, but only when $T = \widetilde{O}((m+n)^{2(1+(3r)^{-1})})$.

We also explain in Appendix H why reducing the problem to almost-low-dimensional linear bandits, a framework employed for regret minimization in previous works such as (Jun et al., 2019; Lu et al., 2021; Kang et al., 2022), gives suboptimal regret guarantees compared to our reduction to misspecified linear bandits. Additionally, we present an algorithm based on the reduction to almost-low-dimensional linear bandits that leverages our entrywise guarantees, with regret $R^\pi(T) = \widetilde{O}((m+n)\sqrt{T})$ in the homogeneous case. Our results complement observations made in (Kang et al., 2024), where reducing to misspecified linear bandits yields $\Theta(r^{1/4}(m+n)^{1/4})$ higher regret than reducing to almost-low-dimensional linear bandits.

As mentioned in Section 4.2, using spectral methods introduces in the subspace recovery error bound a dependence on $L = \|M\|_{\max} \vee \sigma$ instead of σ (see also (Keshavan et al., 2009)). This may affect the regret bounds one may obtain in settings where $\sigma \ll \|M\|_{\max}$. Indeed, should this improvement be possible, one may expect a regret upper bound of order $\widetilde{O}(\sigma(n+m)^{3/4}\sqrt{T} + \|M\|_{\max}(n+m))$ when M and ω are homogeneous (see Appendix I.4 for more details) which is better than ours for $\sigma \ll \|M\|_{\max}$. We believe that the dependence on σ can be improved using either the method proposed in Appendix I.3 or methods based on nuclear norm regularization (Chen et al., 2020) as discussed in Appendix I.4.

Proof sketch of Theorem 8.1. For the sake of simplicity, we assume that $\mu, \kappa = \Theta(1)$ and $m = \Theta(n)$ in the proof sketch. The regret accumulated during the exploration phase (Phase 1) is upper bounded by $2T_1\|M\|_{\max}$. To upper bound the regret of the second phase, we observe that we face a misspecified linear bandit with misspecification $\epsilon_{\max} = \widetilde{O}\left((L^2r)/(T_1\sigma_r\omega_{\min})\right)$ for T_1 sufficiently large according to Corollary 5.1, and we apply the regret bound for misspecified linear bandits from Theorem 1 in (Takemura et al., 2021) (see Theorem G.1). Putting all of this together,

we get:

$$R^\pi(T) = \widetilde{O}\left(T_1\|M\|_{\max} + \frac{L^2}{T_1}\frac{r^{3/2}\sqrt{n}T}{\sigma_r\omega_{\min}} + L\sqrt{rnT}\right).$$

Choosing T_1 that minimizes this expression while using that $\|M\|_{\max}/\sigma_r \lesssim r/\sqrt{mn}$ (see Lemma B.8) yields the claimed regret.

$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad \square$

9. Conclusion

We devised new algorithms achieving state-of-the-art guarantees in several learning tasks for contextual low-rank bandits. A key observation behind our results is that subspace recovery guarantees in the two-to-infinity norm can be leveraged to perform a succinct reduction from a low-rank bandit to a *misspecified* linear bandit problem. This motivates a two-stage approach as an algorithm design principle, and constitutes a core message of our work.

For PE and BPI, we followed such a two-stage approach and proposed RS-PE and RS-BPI whose minimax sample complexity typically scales as $\widetilde{O}(r(m+n)/\varepsilon^2)$. We also considered a single-stage approach to propose SIPS and SBPI with a similar minimax sample complexity. However, experimentally, RS-PE and RS-BPI exhibit far superior performance in comparison with SIPS and SBPI. Shedding some light on this discrepancy is of future interest.

For regret minimization, the benefits of the two-stage approach are even more pronounced. Indeed, the proposed RS-RMIN achieves a non-trivial regret upper bound typically scaling as $\widetilde{O}(r^{5/4}(m+n)^{3/4}\sqrt{T})$. It is unclear if achieving $\widetilde{O}(\sqrt{\text{poly}(r)(m+n)T})$ is possible for general low-rank reward matrices, but we believe this to be an exciting research question.

Acknowledgements

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *Advances in neural information processing systems*, volume 24, 2011.

Abbe, E., Fan, J., Wang, K., and Zhong, Y. Entrywise eigenvector analysis of random matrices with low expected rank. *Annals of statistics*, 48(3):1452, 2020.

Agarwal, A., Foster, D. P., Hsu, D. J., Kakade, S. M., and Rakhlin, A. Stochastic convex optimization with bandit feedback. In *Advances in Neural Information Processing Systems*, volume 24, pp. 1035–1043, 2011.

Auer, P. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

Barata, J. C. A. and Hussein, M. S. The moore–penrose pseudoinverse: A tutorial review of the theory. *Brazilian Journal of Physics*, 42(1–2):146–165, December 2011. ISSN 1678-4448.

Bayati, M., Cao, J., and Chen, W. Speed up the cold-start learning in two-sided bandits with many arms. *arXiv preprint arXiv:2210.00340*, 2022.

Cai, J., Chen, R., Wainwright, M. J., and Zhao, L. Doubly high-dimensional contextual bandits: An interpretable model for joint assortment-pricing. *arXiv preprint arXiv:2309.08634*, 2023.

Cai, T. T. and Zhang, A. Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *The Annals of Statistics*, 46(1): 60–89, 2018.

Candes, E. and Recht, B. Exact matrix completion via convex optimization. *Communications of the ACM*, 55 (6):111–119, 2012.

Candès, E. J. and Tao, T. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

Cape, J., Tang, M., and Priebe, C. E. The two-to-infinity norm and singular subspace geometry with applications to high-dimensional statistics. *The Annals of Statistics*, 47(5):2405–2439, 2019.

Chen, Y., Chi, Y., Fan, J., Ma, C., and Yan, Y. Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. *SIAM journal on optimization*, 30(4):3098–3121, 2020.

Chen, Y., Chi, Y., Fan, J., Ma, C., et al. Spectral methods for data science: A statistical perspective. *Foundations and Trends® in Machine Learning*, 14(5):566–806, 2021.

Chu, W., Li, L., Reyzin, L., and Schapire, R. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 208–214. JMLR Workshop and Conference Proceedings, 2011.

Combes, R. and Proutiere, A. Unimodal bandits: Regret lower bounds and optimal algorithms. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32, pp. 521–529. JMLR Workshop and Conference Proceedings, 2014.

Dani, V., Hayes, T., and Kakade, S. Stochastic linear optimization under bandit feedback. In *COLT*, 2008.

Duan, Y., Jia, Z., and Wang, M. Minimax-optimal off-policy evaluation with linear function approximation. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2701–2709. PMLR, 13–18 Jul 2020.

Eldridge, J., Belkin, M., and Wang, Y. Unperturbed: spectral analysis beyond davis-kahan. In *Algorithmic Learning Theory*, pp. 321–358. PMLR, 2018.

Fan, J., Wang, W., and Zhong, Y. An ℓ_∞ eigenvector perturbation bound and its application to robust covariance estimation. *Journal of Machine Learning Research*, 18 (207):1–42, 2018.

Fan, X., Grama, I., and Liu, Q. Exponential inequalities for martingales with applications. *Electronic Journal of Probability*, 20:1 – 22, 2015.

Foster, D. J., Gentile, C., Mohri, M., and Zimmert, J. Adapting to misspecification in contextual bandits. *Advances in Neural Information Processing Systems*, 33:11478–11489, 2020.

Gentile, C., Li, S., and Zappella, G. Online clustering of bandits. In *International conference on machine learning*, pp. 757–765. PMLR, 2014.

Ghosh, A., Chowdhury, S. R., and Gopalan, A. Misspecified linear bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Gopalan, A., Maillard, O.-A., and Zaki, M. Low-rank bandits with latent mixtures. *arXiv preprint arXiv:1609.01508*, 2016.

Hopkins, S. B., Schramm, T., Shi, J., and Steurer, D. 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*, pp. 178–191, 2016.

Jang, K., Jun, K.-S., Yun, S.-Y., and Kang, W. Improved regret bounds of bilinear bandits using action space analysis. In *International Conference on Machine Learning*, pp. 4744–4754. PMLR, 2021.

Jang, K., Zhang, C., and Jun, K.-S. Efficient low-rank matrix estimation, experimental design, and arm-set-dependent low-rank bandits. *arXiv preprint arXiv:2402.11156*, 2024.

Jun, K.-S., Willett, R., Wright, S., and Nowak, R. Bilinear bandits with low-rank structure. In *International Conference on Machine Learning*, pp. 3163–3172. PMLR, 2019.

Kang, Y., Hsieh, C.-J., and Lee, T. C. M. Efficient frameworks for generalized low-rank matrix bandit problems. *Advances in Neural Information Processing Systems*, 35: 19971–19983, 2022.

Kang, Y., Hsieh, C.-J., and Lee, T. Efficient frameworks for generalized low-rank matrix bandit problems. *arXiv preprint arXiv:2401.07298*, 2024.

Katariya, S., Kveton, B., Szepesvari, C., Vernade, C., and Wen, Z. Stochastic rank-1 bandits. In *Artificial Intelligence and Statistics*, pp. 392–401. PMLR, 2017.

Kaufmann, E., Cappé, O., and Garivier, A. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1): 1–42, 2016.

Keshavan, R., Montanari, A., and Oh, S. Matrix completion from noisy entries. In *Advances in Neural Information Processing Systems*, volume 22, pp. 952–960, 2009.

Kocák, T., Munos, R., Kveton, B., Agrawal, S., and Valko, M. Spectral bandits. *The Journal of Machine Learning Research*, 21(1):9003–9046, 2020.

Kveton, B., Szepesvári, C., Rao, A., Wen, Z., Abbasi-Yadkori, Y., and Muthukrishnan, S. Stochastic low-rank bandits. *arXiv preprint arXiv:1712.04644*, 2017.

Lai, T. L. and Robbins, H. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1): 4–22, 1985.

Lale, S., Azizzadenesheli, K., Anandkumar, A., and Hassibi, B. Stochastic linear bandits with hidden low rank structure. *arXiv preprint arXiv:1901.09490*, 2019.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Lattimore, T., Szepesvari, C., and Weisz, G. Learning with good feature representations in bandits and in rl with a generative model. In *International Conference on Machine Learning*, pp. 5662–5670. PMLR, 2020.

Lee, C.-W., Liu, Q., Abbasi-Yadkori, Y., Jin, C., Lattimore, T., and Szepesvari, C. Context-lumpable stochastic bandits. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Lu, Y., Meisami, A., and Tewari, A. Low-rank generalized linear bandit problems. In *International Conference on Artificial Intelligence and Statistics*, pp. 460–468. PMLR, 2021.

Mackey, L. W., Talwalkar, A., and Jordan, M. I. Distributed matrix completion and robust factorization. *J. Mach. Learn. Res.*, 16(1):913–960, 2015.

Mitzenmacher, M. and Upfal, E. *Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis*. Cambridge University Press, 2017.

Negahban, S. and Wainwright, M. J. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research*, 13(1):1665–1697, 2012.

Pal, S. and Jain, P. Online low rank matrix completion. In *The Eleventh International Conference on Learning Representations*, 2022.

Pal, S., Suggala, A. S., Shanmugam, K., and Jain, P. Optimal algorithms for latent bandits with cluster structure. In *International Conference on Artificial Intelligence and Statistics*, pp. 7540–7577. PMLR, 2023.

Piziak, R. and Odell, P. L. Full rank factorization of matrices. *Mathematics Magazine*, 72(3):193–201, 1999. ISSN 0025570X, 19300980.

Recht, B. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12(12):3413–3430, 2011.

Rudelson, M. and Vershynin, R. Delocalization of eigenvectors of random matrices with independent entries. *Duke Mathematical Journal*, 164(13):2507 – 2538, 2015.

Stojanovic, S., Jedra, Y., and Proutiere, A. Spectral entry-wise matrix estimation for low-rank reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Takemura, K., Ito, S., Hatano, D., Sumita, H., Fukunaga, T., Kakimura, N., and Kawarabayashi, K.-i. A parameter-free algorithm for misspecified linear contextual bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 3367–3375. PMLR, 2021.

Taupin, J., Jedra, Y., and Proutiere, A. Best policy identification in linear mdps. In *2023 59th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 1–8. IEEE, 2023.

Trinh, C., Kaufmann, E., Vernade, C., and Combes, R. Solving bernoulli rank-one bandits with unimodal thompson sampling. In *Algorithmic Learning Theory*, pp. 862–889. PMLR, 2020.

Tropp, J. A. Freedman's inequality for matrix martingales. *Electron. Commun. Probab.*, 16:262–270, 2011a. doi: 10.1214/ECP.v16-1624.

Tropp, J. A. User-friendly tail bounds for matrix martingales. ACM Report 2011-01, Caltech, Pasadena, 2011b.

Valko, M., Munos, R., Kveton, B., and Kocák, T. Spectral bandits for smooth graph functions. In *International Conference on Machine Learning*, pp. 46–54. PMLR, 2014.

Wang, Y.-X., Agarwal, A., and Dudík, M. Optimal and adaptive off-policy evaluation in contextual bandits. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3589–3597. PMLR, 06–11 Aug 2017.

Xi, X., Yu, C. L., and Chen, Y. Matrix estimation for offline reinforcement learning with low-rank structure. *3rd Offline Reinforcement Learning Workshop at Neural Information Processing Systems*, 2022.

Yin, M. and Wang, Y.-X. Asymptotically efficient off-policy evaluation for tabular reinforcement learning. In Chiappa, S. and Calandra, R. (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3948–3958. PMLR, 26–28 Aug 2020.

Zanette, A., Lazaric, A., Kochenderfer, M., and Brunskill, E. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pp. 10978–10989. PMLR, 2020.

A. Additional Related Work

In this appendix, we provide additional related work. Most specifically, we survey some of the recent algorithms for *misspecified* contextual linear bandits. Next, we also discuss bilinear bandits from which the idea of reducing a low-rank bandit problem to an *almost low-dimensional linear bandit* originated.

Misspecified contextual linear bandits. Misspecified contextual linear bandits have recently received a lot of attention (Gopalan et al., 2016; Ghosh et al., 2017; Lattimore et al., 2020; Foster et al., 2020; Zanette et al., 2020; Takemura et al., 2021). The best achievable minimax regret bounds for this setting are $\widetilde{O}(d\sqrt{T} + \epsilon\sqrt{d}T)$ where d is the ambient dimension, ϵ is the misspecification, and T is the time horizon. This guarantee is achievable even if ϵ is not known as proven by (Foster et al., 2020). Interestingly, (Lattimore et al., 2020) show that one cannot hope to improve the dependence on d in the term $\epsilon\sqrt{d}$ due to misspecification for most interesting regimes. However, the first term, that is of order $d\sqrt{T}$, is due to the assumption that the set of arms per context is continuous or infinite. In fact, (Lattimore et al., 2020) also shows that a minimax regret of order $\widetilde{O}(\sqrt{d\log(K)T} + \epsilon\sqrt{d}T)$ is achievable when the misspecified linear bandit has a finite number of arms K. Furthermore, in the case of misspecified contextual linear bandits with a finite number of arms per context, say K, (Takemura et al., 2021) show that SupLin-UCB, initially proposed by (Chu et al., 2011), achieves a regret guarantee of order $\widetilde{O}(\sqrt{d\log^2(K)T} + \epsilon\sqrt{d\log(K)}T)$ without knowledge of ϵ. They further propose a variant of SupLin-UCB that has an improved regret upper bound of order $\widetilde{O}(\sqrt{d\log(K)T} + \epsilon\sqrt{d}T)$.

Bilinear bandits with low-rank structure. In this setting, originally introduced by (Jun et al., 2019), in each round, the learner selects a pair of features (x, y) in $\mathcal{X} \times \mathcal{Y} \subseteq \mathbb{R}^m \times \mathbb{R}^n$, and then observes a reward with expected value $x^\top M y$, where M is assumed to be of rank $r \ll m + n$. This setting is closely related to ours, because if one assumes that $\mathcal{X}$ and $\mathcal{Y}$ correspond to the canonical basis in $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively, then one recovers our reward model. To solve this problem, (Jun et al., 2019) proposed ESTR, a two-stage algorithm that first estimates the singular subspaces of M using spectral methods, and then solves an almost low-dimensional linear bandit problem, using an adaptation of OFUL (Abbasi-Yadkori et al., 2011) referred to as LowOFUL. The key insight is that subspace recovery enables a reformulation of bilinear bandit problems as low-dimensional linear bandits. This lead to the first regret guarantees of order $\widetilde{O}((m + n)^{3/2}\sqrt{T})$. Since then, there has been many subsequent work with new algorithms, improvements, and various generalizations (Jang et al., 2021; Lu et al., 2021; Kang et al., 2022; Cai et al., 2023). Notably, (Kang et al., 2022) proposes an algorithm, G-ESTT, to solve generalized bilinear bandit problems. Their algorithm also achieves a regret guarantee of order $\widetilde{O}((m + n)^{3/2}\sqrt{T})$, and relies on the same two-stage algorithm design idea as (Jun et al., 2019), but uses instead a novel subspace estimator that is based on Stein's method. A major limitation of existing state-of-the-art algorithms for bilinear bandits is that when $|\mathcal{X} \times \mathcal{Y}| = \widetilde{O}(mn)$, as is the case in our setting, the obtained regret guarantees are no better than the ones achievable by minimax optimal algorithms for unstructured bandits. Indeed, an algorithm such as UCB that does not use the low-rank structure achieves a regret rate of order $\widetilde{O}(\sqrt{mnT})$ (Lattimore & Szepesvári, 2020)).

It is worth noting that in a concurrent work, (Jang et al., 2024) derive a regret bound that scales with $\tilde{O}((m + n)^{1/4}\sqrt{BT})$ where B is a constant that depends on the action set geometry. This constant appears in their bounds because they use an experimental design approach. It is unclear how this constant would scale with our action set, but for the case of the unit sphere, which would also include our action set, they obtain a regret bound of order $\widetilde{O}((m + n)^{5/4}\sqrt{T})$. This still does not attain a sub-linear dependence on $m + n$ as we do in our work.

B. Singular Subspace Recovery

In this appendix, we give the proofs of our results on the singular subspace recovery in the two-to-infinity norm. In Appendix B.1, we provide the proof of Theorem 4.1 which in fact follows from Theorem B.1. The latter theorem is also a guarantee on the two-to-infinity norm but where the estimation error is evaluated in an alternative way. The proof of Theorem B.1 is provided in Appendix B.2 and features two key ideas, namely the leave-one-out analysis (Abbe et al., 2020; Chen et al., 2021) and the Poisson approximation argument (Stojanovic et al., 2024).

B.1. Proof of Theorem 4.1 – Recovery in $\| \cdot \|_{2\to\infty}$

We now present the two key results from which Theorem 4.1 follows. We defer the proofs of these two claims to Appendices B.2 and B.3, respectively.

Theorem B.1. *Let $\delta \in (0,1)$. Then, the following event*

$$\max\left(\|U - \widehat{U}(\widehat{U}^\top U)\|_{2\to\infty}, \|V - \widehat{V}(\widehat{V}^\top V)\|_{2\to\infty}\right) \leq C\frac{L}{\sigma_r(M)}\sqrt{\frac{\mu^2\kappa^2 r(m+n)}{T\omega_{\min}(m\wedge n)}\log^3\left(\frac{(m+n)T}{\delta}\right)},$$

holds with probability at least $1 - \delta$ as long as

$$T \geq c\frac{L^2(m+n)}{\sigma_r^2(M)\omega_{\min}}\log^3\left(\frac{(m+n)T}{\delta}\right)$$

for some universal constants $C, c > 0$.

Lemma B.2. *Under the event $\|\widetilde{M} - M\|_{\mathrm{op}} \leq \sigma_r(M)/2$, we have*

$$d_{2\to\infty}(U, \widehat{U}) = \|UU^\top - \widehat{U}\widehat{U}^\top\|_{2\to\infty} \leq 2\|U - \widehat{U}\widehat{U}^\top U\|_{2\to\infty} + \frac{4\|U\|_{2\to\infty}\|\widetilde{M} - M\|_{\mathrm{op}}}{\sigma_r(M)},$$

$$d_{2\to\infty}(V, \widehat{V}) = \|VV^\top - \widehat{V}\widehat{V}^\top\|_{2\to\infty} \leq 2\|V - \widehat{V}\widehat{V}^\top V\|_{2\to\infty} + \frac{4\|V\|_{2\to\infty}\|\widetilde{M} - M\|_{\mathrm{op}}}{\sigma_r(M)}.$$

Now the proof of Theorem 4.1 follows straightforwardly. First, we note that under the event that $\|\widetilde{M} - M\|_{\mathrm{op}} \leq \sigma_r(M)/2$, by Lemma B.2 we only need to upper bound the terms $\|U - \widehat{U}\widehat{U}^\top U\|_{2\to\infty}, \|V - \widehat{V}\widehat{V}^\top V\|_{2\to\infty}, \|\widetilde{M} - M\|_{\mathrm{op}}$. Now, observe that we may use Theorem B.1 to upper bound $\|U - \widehat{U}\widehat{U}^\top U\|_{2\to\infty}$ and $\|V - \widehat{V}\widehat{V}^\top V\|_{2\to\infty}$ with high probability, and use Proposition B.5 to upper bound the term $\|\widetilde{M} - M\|_{\mathrm{op}}$ with high probability. The obtained upper bounds combined with the fact that $\|U\|_{2\to\infty} \vee \|V\|_{2\to\infty} \leq \mu\sqrt{r/(m\wedge n)}$ yield the final result.

B.2. Proof of Theorem B.1

The proof of Theorem B.1 is rather involved but follows the same steps as the proof of Lemma 30 in (Stojanovic et al., 2024), with the slight difference that in (Stojanovic et al., 2024), it is assumed that $\omega_{i,j} = 1/(mn)$ for all i, j. For the sake of completeness, we highlight the key steps of the proof next.

Step 1: Poisson approximation. For the sake of the analysis, it is rather convenient to describe the estimate $\widetilde{M}$ given in (1) with an alternative random matrix which has the same distribution. For all $(i, j) \in [m] \times [n]$, define $Z_{i,j} = \sum_{t=1}^T \mathbb{1}_{\{(i_t, j_t) = (i,j)\}}$ and let $(\xi'_{i,j,t})$ be a sequence of i.i.d. random variables that have the same distribution as ξ_1. Now, observe that our matrix $\widetilde{M}$ has the same distribution as the random matrix $\widetilde{M}'$ defined as:

$$\forall(i,j) \in [m] \times [n], \qquad \widetilde{M}'_{i,j} = \frac{1}{T\omega_{i,j}}\sum_{t=1}^{Z_{i,j}}(M_{i,j} + \xi'_{i,j,t}). \tag{13}$$

Let $\mathbb{P}$ denote the joint probability distribution of $(Z_{i,j})_{(i,j)\in[m]\times[n]}$ and $(\xi'_{i,j,t})_{(i,j)\in[m]\times[n],t\geq 1}$.

Next, we describe a compound Poisson random matrix model that will serve as an approximation of the random model (13). Let $Y \in \mathbb{R}^{m\times n}$ be a random matrix with independent entries, such that $Y_{i,j} \sim \mathrm{Poisson}(T\omega_{i,j})$, for all $(i,j) \in [m] \times [n]$.

Let us further define the random matrix X taking values $\mathbb{R}^{m \times n}$ as follows:

$$\forall (i,j) \in [m] \times [n], \qquad X_{i,j} = \frac{1}{T\omega_{i,j}} \sum_{t=1}^{Y_{i,j}} (M_{i,j} + \xi'_{i,j,t}). \tag{14}$$

We observe that the entries of the matrix X are independent and distributed according to compound Poisson distributions.

Let $\mathbb{P}'$ denote the joint probability distribution of $(Y_{i,j})_{(i,j) \in [m] \times [n]}$ and $(\xi'_{i,j,t})_{(i,j) \in [m] \times [n], t \geq 1}$. The Poisson approximation argument is formalized in the next lemma.

Lemma B.3 (Poisson Approximation)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ (resp. $(\Omega, \mathcal{F}, \mathbb{P}')$) be the probability space under the random matrix model* (13) *(resp.* (14)*). Then for any event $\mathcal{E} \in \mathcal{F}$, we have*

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

The proof of Lemma B.3 follows from Lemma B.4 and is similar to the proof of Lemma 20 in (Stojanovic et al., 2024).

Proof. Let $\mathcal{E} \in \mathcal{F}$ be an event and for convenience, introduce the notation $X = (\xi_{i,j,t})_{(i,j) \in [m] \times [n], t \geq 1}$. First, we note that $(Y_{i,j})_{(i,j) \in [m] \times [n]} \sim \text{Multinomial}(T, (\omega_{i,j})_{(i,j) \in m \times n})$. Let $f(Y, X) = \mathbb{1}_{\{\mathcal{E}\}}$. Then, by Lemma B.4, we have

$$\mathbb{E}[f(X,Y)|X] \leq e\sqrt{T}\mathbb{E}'[f(X,Z)|X].$$

By taking the expectation with respect to X, we get $\mathbb{E}[f(X,Y)] \leq \mathbb{E}[f(X,Z)]$, which means that $\mathbb{P}(\mathcal{E}) \leq e\sqrt{T}\mathbb{P}'(\mathcal{E})$. $\square$

We borrow the following Lemma from (Stojanovic et al., 2024) (see their Lemma 14) which itself is a mild generalization of Theorem 5.7 in (Mitzenmacher & Upfal, 2017)

Lemma B.4. *Let $Y_i^{(t)} \sim \text{Poisson}(tp_i)$, $i = 1, \ldots, n$, be independent random variables with $\sum_{i=1}^n p_i = 1$. Moreover, let $(Z_1^{(t)}, Z_2^{(t)}, \ldots, Z_n^{(t)}) \sim \text{Multinomial}(t, (p_1, \ldots, p_n))$. Let $f : \mathbb{R}^p \to \mathbb{R}_+$ be any non-negative function. Then*

$$\mathbb{E}[f(Z_1^{(t)}, \ldots, Z_p^{(t)})] \leq e\sqrt{t}\mathbb{E}[f(Y_1^{(t)}, \ldots, Y_p^{(t)})].$$

Step 2: Key concentration inequalities. The following propositions are direct consequences of a truncated matrix Bernstein inequality (see e.g., Proposition A.7 in (Hopkins et al., 2016)).

Proposition B.5. *Under the random matrix model* (14) *with compound Poisson entries, for all $\delta \in (0,1)$, for all $T \geq 13/(\omega_{\min}(m \wedge n)) \log^3((m+n)/\delta)$, we have:*

$$\mathbb{P}\left(\|\widetilde{M} - M\|_{\text{op}} \leq 36\sqrt{2}L\sqrt{\frac{1}{T\omega_{\min}}} \left(\sqrt{(m+n)\log\left(\frac{m+n}{\delta}\right)} + \log^{3/2}\left(\frac{m+n}{\delta}\right) \right) \right) \geq 1 - \delta$$

with $L = \|M\|_{\max} \vee \sigma$.

Proposition B.6. *Let A be a $m \times 2r$ deterministic matrix, and B be a $n \times 2r$ deterministic matrix. Then, under the random matrix model* (14) *with compound Poisson entries, and denoting $L = \|M\|_{\max} \vee \sigma$, we have:*

(i) for all $\ell \in [m]$, for all $\delta \in (0,1)$, for all $T \geq (1/(n\omega_{\min})) \log^3(en/\delta)$, the event

$$\|(\widetilde{M}_{\ell,:} - M_{\ell,:})A\|_2 \leq 73\sqrt{2}L\|A\|_{2\to\infty}\sqrt{\frac{1}{T\omega_{\min}}} \left(\sqrt{n\log\left(\frac{en}{\delta}\right)} + \log^{3/2}\left(\frac{en}{\delta}\right) \right) \tag{15}$$

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

(ii) for all $k \in [n]$, for all $\delta \in (0,1)$, for all $T \geq (1/(m\omega_{\min})) \log^3(em/\delta)$, the event

$$\|(\widetilde{M}_{:,k} - M_{:,k})^\top B\|_2 \leq 73\sqrt{2}L\|B\|_{2\to\infty}\sqrt{\frac{1}{T\omega_{\min}}} \left(\sqrt{m\log\left(\frac{em}{\delta}\right)} + \log^{3/2}\left(\frac{em}{\delta}\right) \right) \tag{16}$$

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

The proofs of Proposition B.5 and Proposition B.6 are essentially the same as those of Proposition 26 and Proposition 27 in (Stojanovic et al., 2024), and are omitted. Note that both Propositions in (Stojanovic et al., 2024) were derived under an i.i.d. sub-gaussian noise assumption.

Step 3: Leave-one-out analysis.   This part of the analysis follows the steps of the proof of Lemma 30 in (Stojanovic et al., 2024), but we repeat the main results for the sake of completeness.

(i) First, a dilation trick is needed to reduce the analysis to that of symmetric matrices. We introduce:

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

and recalling the SVD of M, $M = U\Sigma V^\top$, we can express the SVD of S as follows:

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

We can define $\widetilde{S}$ in a similar way, and denote $\widehat{Q} \in \mathbb{R}^{(m+n)\times 2r}$, the matrix containing the $2r$ eigenvectors of the best $2r$-rank approximation of $\widetilde{S}$. Here, we observe that:

$$\|Q - \widehat{Q}(\widehat{Q}^\top Q)\|_{2\to\infty} = \max\left\{ \|U - \widehat{U}(\widehat{U}^\top U)\|_{2\to\infty}, \|V - \widehat{V}(\widehat{V}^\top V)\|_{2\to\infty} \right\}. \tag{17}$$

Additionally, we note that $\sigma_{2r}(S) = \sigma_r(M)$. To simplify the notation, we denote $E = \widetilde{S} - S$ and note that $\|E\|_{\mathrm{op}} = \|\widetilde{M} - M\|_{\mathrm{op}}$.

(ii) Second, it can be shown, under the event

$$\mathcal{E}_1 = \{\|E\|_{\mathrm{op}} \le C_1 \sigma_r(M)\}$$

where C_1 is a universal and sufficiently small positive constant, that:

$$\|Q - \widehat{Q}(\widehat{Q}^\top Q)\|_{2\to\infty} \le \frac{1}{\sigma_{2r}(S)} \left(\frac{5\|SQ\|_{2\to\infty}\|E\|_{\mathrm{op}}}{\sigma_{2r}(S)} + 3\|EQ\|_{2\to\infty} + 2\|E(Q - \widehat{Q}(\widehat{Q}^\top Q))\|_{2\to\infty} \right). \tag{18}$$

We can easily control $\|E\|_{\mathrm{op}}$ using Proposition B.5 and $\|EQ\|_{2\to\infty}$ using Proposition B.6. On the other hand, the term $\|E(Q - \widehat{Q}(\widehat{Q}^\top Q))\|_{2\to\infty}$ is the bottleneck of the analysis because E and $(Q - \widehat{Q}(\widehat{Q}^\top Q))$ are dependent on each other in a non-trivial way. The leave-one-out analysis is used to deal with this term.

(iii) We make the leave-one-out analysis more precise. We introduce for all $\ell \in [m + n]$, the matrix $\widetilde{S}^{(\ell)}$: for all $(i, j) \in [m + n] \times [m + n]$,

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

We further define $\widehat{Q}^{(\ell)} \in \mathbb{R}^{(m+n)\times 2r}$ as the matrix that contains the $2r$ eigenvector of the best $2r$-rank approximation of $\widetilde{S}^{(\ell)}$. It can be shown that, again under $\mathcal{E}_1$ where C_1 is sufficiently small, that

$$\|E(Q - \widehat{Q}W_{\widehat{Q}})\|_{2\to\infty} \le \max_{\ell \in [m+n]} \|E_{\ell,:}(Q - \widehat{Q}^{(\ell)}W_{\widehat{Q}^{(\ell)}})\|_2 + \|E\|_{\mathrm{op}}\|\widehat{Q}W_{\widehat{Q}} - \widehat{Q}^{(\ell)}W_{\widehat{Q}^{(\ell)}}\|_F, \tag{19}$$

$$\|\widehat{Q}W_{\widehat{Q}} - \widehat{Q}^{(\ell)}W_{\widehat{Q}^{(\ell)}}\|_F \le \frac{16}{\sigma_r(M)} (\|E_{\ell,:}(Q - \widehat{Q}^{(\ell)}W_{\widehat{Q}^{(\ell)}})\|_2 + \|E_{\ell,:}Q\|_2 + \|E\|_{\mathrm{op}}\|Q - \widehat{Q}W_{\widehat{Q}}\|_{2\to\infty} + \|E\|_{\mathrm{op}}\|Q\|_{2\to\infty}), \tag{20}$$

where we denote $W_{\widehat{Q}} = \widehat{Q}^\top Q$ and $W_{\widehat{Q}^{(\ell)}} = \widehat{Q}^{(\ell)\top} Q$.

Now, the key observation that underpins the leave-one-out analysis is that $E_{\ell,:}$ and $Q - \widehat{Q}^{(\ell)}W_{\widehat{Q}^{(\ell)}}$ are independent provided that the entries of the error matrix E are independent. This will be ensured thanks to the Poisson approximation (using

Lemma B.3) by controlling the quantities of interest under the random matrix model (14). Define

$$\mathcal{B}(\delta) = \sqrt{\frac{m+n}{T\omega_{\min}} \log^3\left(\frac{m+n}{\delta}\right)}$$

and the events

$$\mathcal{E}_2 = \{\|E\|_{\mathrm{op}} \leq C_2 L \mathcal{B}(\delta)\}$$

$$\forall \ell \in [m+n], \qquad \mathcal{E}_3^{(\ell)} = \left\{\|E_{\ell,:}(Q - \widehat{Q}^{(\ell)} W_{\widehat{Q}^{(\ell)}})\|_2 \leq C_3 L \|Q - \widehat{Q}^{(\ell)} W_{\widehat{Q}^{(\ell)}}\|_{2\to\infty} \mathcal{B}(\delta)\right\}$$

where C_2 and C_3 are sufficiently large positive constants. Then, applying Proposition B.5 and Proposition B.6, we have that $\mathbb{P}(\mathcal{E}_2) \geq 1 - \delta$ and $\mathbb{P}(\mathcal{E}_3^{(\ell)}) \geq 1 - \delta$ for all $\ell \in [m+n]$, provided that

$$T \geq \frac{c}{\omega_{\min}(m \wedge n)} \log^3\left(\frac{m+n}{\delta}\right)$$

for $c > 0$ a sufficiently large universal constant. Additionally, using the fact that $L/\sigma_r(M) \geq 1/\sqrt{mn}$ from Lemma B.8, we can deduce that $\mathcal{E}_1 \supseteq \mathcal{E}_2$ and $\mathbb{P}(\mathcal{E}_2) \geq 1 - \delta$ as well as $\mathbb{P}(\mathcal{E}_3^{(\ell)}) \geq 1 - \delta$ for all $\ell \in [m+n]$ when

$$T \geq \frac{c'L^2(m+n)}{\sigma_r^2(M)\omega_{\min}} \log^3\left(\frac{m+n}{\delta}\right) \tag{21}$$

for a positive universal constant c' sufficiently large. At this point, what remains to be done is to deal with the inequalities (18), (19) and (20) under the intersection of the events $\mathcal{E}_2, \mathcal{E}_3^{(1)}, \ldots, \mathcal{E}_3^{(m+n)}$. We spare the readers the details of this tedious task and write directly the resulting bound

$$\|Q - \widehat{Q}(\widehat{Q}^\top Q)\|_{2\to\infty} \leq C' \frac{L}{\sigma_r(M)} \frac{\mu\kappa r^{1/2}\mathcal{B}(\delta)}{\sqrt{m \wedge n}}$$

for some universal constant $C' > 0$. Thus, by a union bound, the above bound holds with probability at least $1 - (1+m+n)\delta$, provided that condition (21) holds, under the random matrix model (14). Using Lemma B.3, we conclude that the same event holds with probability $1 - (1+m+n)e\sqrt{T}\delta$, provided that condition (21) holds, under the random matrix model (13). Re-parametrizing by $\delta' = e(m+n+1)\sqrt{T}\delta$ concludes the proof.

B.3. Proof of Lemma B.2

We start with the first inequality. Let $W \in \mathbb{R}^{r \times r}$ be an arbitrary orthogonal matrix and note that

$$\|UU^\top - \widehat{U}\widehat{U}^\top\|_{2\to\infty} \leq \|UU^\top - U(\widehat{U}W)^\top\|_{2\to\infty} + \|U(\widehat{U}W)^\top - \widehat{U}W(\widehat{U}W)^\top\|_{2\to\infty}$$

$$\leq \|U\|_{2\to\infty}\|U - \widehat{U}W\|_{\mathrm{op}} + \|U - \widehat{U}W\|_{2\to\infty} \tag{22}$$

where the first inequality follows from the triangular inequality and the fact that $WW^\top = I_r$, and the second inequality follows by the inequality $\|AB\|_{2\to\infty} \leq \|A\|_{2\to\infty}\|B\|_{\mathrm{op}}$ and the fact that $\|(\widehat{U}W)^\top\|_{\mathrm{op}} = 1$.

Next, we choose $W = \mathrm{sgn}(\widehat{U}^\top U)$ where $\mathrm{sgn}(A)$ denotes the sign matrix of A (see (Chen et al., 2021)). First, according to Lemma 4.15 in (Chen et al., 2021) and Davis-Kahan inequality, we know that

$$\|\widehat{U}^\top U - \mathrm{sgn}(\widehat{U}^\top U)\|_{\mathrm{op}} = \|\sin(\widehat{U}, U)\|_{\mathrm{op}}^2 \leq \frac{2\|\widetilde{M} - M\|_{\mathrm{op}}^2}{\sigma_r(M)^2} \leq \frac{\|\widetilde{M} - M\|_{\mathrm{op}}}{\sigma_r(M)} \leq \frac{1}{2}$$

where the last two inequalities hold under the event $\|\widetilde{M} - M\|_{\mathrm{op}} \leq \sigma_r(M)/2$. Thus, under this event, we have

$$\|U - \widehat{U}W\|_{2\to\infty} \leq \|U - \widehat{U}\widehat{U}^\top U\|_{2\to\infty} + \|\widehat{U}\|_{2\to\infty}\|\widehat{U}^\top U - \mathrm{sgn}(\widehat{U}^\top U)\|_{\mathrm{op}}$$

$$\leq \|U - \widehat{U}\widehat{U}^\top U\|_{2\to\infty} + (\|U - \widehat{U}W\|_{2\to\infty} + \|U\|_{2\to\infty})\|\widehat{U}^\top U - \mathrm{sgn}(\widehat{U}^\top U)\|_{\mathrm{op}}$$

$$\leq \|U - \widehat{U}\widehat{U}^\top U\|_{2\to\infty} + \frac{\|U - \widehat{U}W\|_{2\to\infty}}{2} + \frac{\|U\|_{2\to\infty}\|\widetilde{M} - M\|_{\mathrm{op}}}{\sigma_r(M)}$$

which implies

$$\|U - \widehat{U}W\|_{2\to\infty} \leq 2\|U - \widehat{U}\widehat{U}^\top U\|_{2\to\infty} + \frac{2\|U\|_{2\to\infty}\|\widetilde{M} - M\|_{\mathrm{op}}}{\sigma_r(M)}. \tag{23}$$

We also have by Davis-Kahan inequality and relations between the subspace distances (see Lemma 2.5 and Lemma 2.6 in (Chen et al., 2021))

$$\|U - \widehat{U}W\|_{\mathrm{op}} \leq \sqrt{2}\|UU^\top - \widehat{U}\widehat{U}^\top\|_{\mathrm{op}} = \sqrt{2}\|\sin(\widehat{U}, U)\|_{\mathrm{op}} \leq \frac{\sqrt{2}\|\widetilde{M} - M\|_{\mathrm{op}}}{\sigma_r(M)}. \tag{24}$$

Finally, combining the inequalities (22), (23) and (24), we conclude that under the event $\|\widetilde{M} - M\|_{\mathrm{op}} \leq \sigma_r(M)/2$, we have

$$\|UU^\top - \widehat{U}\widehat{U}^\top\|_{2\to\infty} \leq 2\|U - \widehat{U}\widehat{U}^\top U\|_{2\to\infty} + \frac{4\|U\|_{2\to\infty}\|\widetilde{M} - M\|_{\mathrm{op}}}{\sigma_r(M)}.$$

This concludes the proof of the first inequality. The second one follows similarly.

B.4. Subspace recovery in Frobenius norm

Deriving such guarantees is by now well understood under general noise assumptions (see e.g., (Keshavan et al., 2009; Chen et al., 2021)). Indeed, the classical proof follows by first deriving upper bounds on $\|\widetilde{M} - M\|_{\mathrm{op}}$ with high probability, then using Davis-Kahan's or Wedin's theorem to obtain upper bounds on the subspace recovery error in the Frobenius norm. The following lemma provides such upper bounds:

Lemma B.7 (Subspace recovery in $\|\cdot\|_{\mathrm{F}}$). *Let $\delta \in (0, 1)$ and consider the setting of Theorem 4.1. Then, with probability at least $1 - \delta$, it holds that*

$$\max\left(\|UU^\top - \widehat{U}\widehat{U}^\top\|_{\mathrm{F}}, \|VV^\top - \widehat{V}\widehat{V}^\top\|_{\mathrm{F}}\right) \lesssim \frac{L\sqrt{r}}{\sigma_r(M)}\sqrt{\frac{m+n}{T\omega_{\min}}\log^3\left(\frac{(m+n)T}{\delta}\right)},$$

provided that $T \gtrsim \frac{1}{\omega_{\min}(m\wedge n)}\log^3\left(\frac{(m+n)T}{\delta}\right)$.

We remark that after $\widetilde{\Omega}(m+n)$ observations, when $m+n = O(m\wedge n)$, we can recover the singular subspaces with an error in $\|\cdot\|_{\mathrm{F}}$ of order $\widetilde{O}(\sqrt{(m+n)/T})$. This contrasts with the error obtained from Theorem 4.1 that is of order $\widetilde{O}(\sqrt{1/T})$ (in the homogeneous case) with the same amount of observations. This suggests that the subspace recovery error (seen as a matrix) is *delocalized*, i.e., spread out along $m+n$ directions.

Proof of Lemma B.7. Define $\sin\Theta(U, \widehat{U})$ as a diagonal matrix with entries $\{\sin\theta_i\}_{i=1}^r$ where $\theta_i = \arccos\sigma_i(\widehat{U}^\top U)$. Applying first Lemma 1 in (Cai & Zhang, 2018), and then Davis-Kahan's theorem (Corollary 2.8 in (Chen et al., 2021)), we obtain:

$$\|UU^\top - \widehat{U}\widehat{U}^\top\|_F \leq \sqrt{2r}\|UU^\top - \widehat{U}\widehat{U}^\top\|_{\mathrm{op}} \leq 2\sqrt{r}\|\sin\Theta(U, \widehat{U})\|_{\mathrm{op}} \leq \frac{2\sqrt{2r}\|M - \widetilde{M}\|_{\mathrm{op}}}{\sigma_r(M)}.$$

An analogous bound holds for $\|VV^\top - \widehat{V}\widehat{V}^\top\|_F$. Lastly, using Proposition B.5 to bound $\|M - \widetilde{M}\|_{\mathrm{op}}$ and Lemma B.8 to bound $\|M\|_{\max}/\sigma_r(M)$ concludes the proof. □

B.5. Additional Lemmas

The following result is immediate (see Lemma 17 in (Stojanovic et al., 2024)).

Lemma B.8. *Let M be an $m \times n$ matrix, with rank r, incoherence parameter $\mu > 0$, and condition number $\kappa > 0$. Then,*

$$\frac{1}{\sqrt{mn}} \leq \frac{\|M\|_{\max}}{\sigma_1(M)} \leq \frac{\mu^2 r}{\sqrt{mn}},$$

Consequently, we also have

$$\frac{1}{\sqrt{mn}} \leq \frac{\|M\|_{\max}}{\sigma_r(M)} \leq \frac{\mu^2\kappa r}{\sqrt{mn}}.$$

C. Policy Evaluation for Contextual Linear Bandits

In this appendix, we study the PE task for the simpler contextual linear bandit problem, whose error bound analysis will be useful to handle the PE task for the contextual low-rank bandit problem. We also demonstrate that the linear bandit error bound is instance-optimal by deriving a sample complexity lower bound of independent interest.

We observe T i.i.d. samples of the form $(i_t, j_t, r_t)_{1 \le t \le T}$ where the context $i_t \in [m]$ is drawn independently from a known distribution ρ and the arm $j_t \in [n]$ is sampled from $\pi^b(.|i_t)$ for a known behavioral policy π^b. The reward distribution is assumed to have a linear structure with respect to a known feature map $\psi : [m] \times [n] \to \mathbb{R}^d$: there exists an unknown parameter θ in $\mathbb{R}^d$ such that $r_t = \psi_{i_t,j_t}^\top \theta + \xi_t$ where $(\xi_t)_{t \le 1}$ is an i.i.d. sequence of σ-subgaussian noise, independent from all the other random variables. The overall bandit model described above will be denoted as $\mathcal{M}_\theta$. For a given target policy π, we use the previously defined notations $\omega_{i,j} = \rho(x)\pi^b(j|i)$, $\omega_{i,j}^\pi = \rho(x)\pi(j|i)$ and we define

$$\Lambda = \sum_{(i,j) \in [m] \times [n]} \omega_{i,j} \psi_{i,j} \psi_{i,j}^\top \quad \text{and} \quad v_\theta^\pi = \sum_{i,j} \omega_{i,j}^\pi \psi_{i,j}^\top \theta.$$

For any $\delta \in (0, 1)$ and any $\varepsilon > 0$, we say that $\hat{v}^\pi$ is (ε, δ)-PAC[10] if $\mathbb{P}_\theta(|v_\theta^\pi - \hat{v}^\pi| \le \varepsilon) \ge 1 - \delta$ for every $\theta \in \mathbb{R}^d$. We will write v^π for v_θ^π when there is no ambiguity about the underlying parameter. We also write $N \succeq M$ (resp. $N \preceq M$) when $N - M$ (resp. $M - N$) is positive semi-definite, and denote the Moore-Penrose pseudoinverse (Barata & Hussein, 2011) of N by $N^\dagger$.

C.1. Instance-dependent sample complexity lower bound

In this section only, we assume that the noise follows a centered Gaussian distribution: $\xi_t \sim \mathcal{N}(0, \sigma^2)$.

Lemma C.1. *Denote by $q_\theta(.|i, j)$ the distribution of the reward when choosing action j under context i, and by $N_{i,j}$ the number of times (i, j) has been observed in the sample set. Let $L_T = L_T(i_1, j_1, r_1, \ldots, i_T, j_T, r_T)$ be the log-likelihood ratio between two models $\mathcal{M}_\theta$ and $\mathcal{M}_{\theta'}$. Then,*

$$\mathbb{E}_\theta[L_T] = \sum_{(i,j) \in [m] \times [n]} \mathbb{E}_\theta[N_{i,j}] \, \mathrm{KL}(q_\theta(.|i,j), q_{\theta'}(.|i,j)).$$

Proof of Lemma C.1. We can write

$$L_T = \sum_{t=1}^T \sum_{i,j} \mathbb{1}_{i_t=i, j_t=j} \log\left(\frac{q_\theta(r_t|i,j)}{q_{\theta'}(r_t|i,j)}\right)$$

$$= \sum_{i,j} \sum_{t=1}^{N_{i,j}} \log\left(\frac{q_\theta(r_t^{i,j}|i,j)}{q_{\theta'}(r_t^{i,j}|i,j)}\right)$$

where $r_t^{i,j}$ is the reward observed after the t-th time that arm j has been selected under context i. Since the $r_t^{i,j}$ are i.i.d. of distribution $q_\theta(.|i, j)$ (resp. $q_{\theta'}(.|i, j)$) under $\mathcal{M}_\theta$ (resp. $\mathcal{M}_{\theta'}$), Wald's Lemma yields the result. $\square$

Proposition C.2. *Let $\mathrm{Alt}_\varepsilon(\theta) = \{\theta' \in \mathbb{R}^d, |v_{\theta'}^\pi - v_\theta^\pi| \ge 2\varepsilon\}$. If $\hat{v}$ is (ε, δ)-PAC with $\delta \in (0, 1/2]$, then for every $\theta \in \mathbb{R}^d$,*

$$T \ge \frac{\mathrm{kl}(\delta, 1 - \delta)}{\displaystyle\inf_{\theta' \in \mathrm{Alt}_\varepsilon(\theta)} \frac{1}{2\sigma^2} \|\theta' - \theta\|_\Lambda^2}.$$

Proof of Proposition C.2. The information-theoretic arguments used in the proof of Lemma 19 from (Kaufmann et al., 2016) ensure that for any θ and θ' and any measurable set E with respect to $\sigma(i_1, j_1, r_1, \ldots, i_T, j_T, r_T)$,

$$\mathbb{E}_\theta[L_T] \ge \mathrm{kl}(\mathbb{P}_\theta(E), \mathbb{P}_{\theta'}(E)).$$

---

[10]We will often omit the superscript π to ease the notations.

The choices $E = \{|v_\theta^\pi - \hat{v}| \geq \varepsilon\}$ and $\theta' \in \text{Alt}_\varepsilon(\theta)$ directly yield $\mathbb{P}_\theta(E) \leq \delta$. Furthermore, by the triangular inequality, $|v_\theta^\pi - \hat{v}| \geq |v_\theta^\pi - v_{\theta'}^\pi| - |v_{\theta'}^\pi - \hat{v}|$. Hence, $\mathbb{P}_{\theta'}(E) \geq \mathbb{P}_{\theta'}(|v_{\theta'}^\pi - \hat{v}| \geq \varepsilon) \geq 1 - \delta$. Since $0 < \delta \leq 1/2$, we have in particular $\delta \leq 1 - \delta$ and $\mathbb{P}_\theta(E) \leq \mathbb{P}_{\theta'}(E)$. When $x \leq y$, $x \mapsto \text{kl}(x, y)$ is decreasing and $y \mapsto \text{kl}(x, y)$ is increasing. Thus, $\text{kl}(\mathbb{P}_\theta(E), \mathbb{P}_{\theta'}(E)) \geq \text{kl}(\delta, 1 - \delta)$, so that

$$\mathbb{E}_\theta[L_T] \geq \text{kl}(\delta, 1 - \delta).$$

Since the reward follows a Gaussian distribution, Lemma C.1 yields

$$\mathbb{E}_\theta[L_T] = \frac{T}{2\sigma^2} \sum_{(i,j) \in [m] \times [n]} \omega_{i,j}((\theta - \theta')^\top \psi_{i,j})^2,$$

where we have used that in our setting, $\mathbb{E}_\theta[N_{i,j}] = T\omega_{i,j}$. Combining this result with the inequality above gives

$$T \geq \frac{\text{kl}(\delta, 1 - \delta)}{\frac{1}{2\sigma^2}(\theta' - \theta)^\top \Lambda(\theta' - \theta)},$$

and taking the infimum over every $\theta' \in \text{Alt}_\varepsilon(\theta)$ completes the proof.

$\square$

This infimum can be computed explicitly with the use of Lemma C.8. It yields the following result.

Theorem C.3. *Let $\delta \in (0, 1/2]$ and assume that $\psi_\pi := \sum_{i,j} \omega_{i,j}^\pi \psi_{i,j} \in \text{Im}(\Lambda)$. The sample complexity of any (ϵ, δ)-PAC estimator of v^π must satisfy*

$$T \geq \frac{\sigma^2 \text{kl}(\delta, 1 - \delta)}{2\varepsilon^2} \|\psi_\pi\|_{\Lambda^\dagger}^2.$$

Proof of Theorem C.3. From Proposition C.2, we know that

$$T \geq \frac{\sigma^2 \text{kl}(\delta, 1 - \delta)}{\inf_{\theta' \in \text{Alt}_\varepsilon(\theta)} \frac{1}{2} \|\theta' - \theta\|_\Lambda^2}.$$

Let us show that the infimum can be rewritten as the infimum in Lemma C.8 for $n = 1$, $\Lambda_1 = \Lambda$ and $\nu_1 = \psi_\pi$. First, note that since $v_\theta^\pi = \theta^\top \left(\sum_{i,j} \omega_{i,j}^\pi \psi_{i,j}\right)$, the constraint $\theta' \in \text{Alt}_\varepsilon(\theta)$ can be rewritten as $|(\theta' - \theta)^\top \psi_\pi| \geq 2\varepsilon$. Since replacing $\theta' - \theta$ by $\theta - \theta'$ yields the same objective, the absolute value can be removed. Moreover, the inequality constraint can be simplified to an equality constraint: if we had $(\theta' - \theta)^\top \psi_\pi > 2\varepsilon$, we could normalize $\theta' - \theta$ to make the objective strictly smaller. Finally, note that the constrained problem only depends on θ' through $\mu := \theta' - \theta$. Computing the infimum consequently amounts to solving the quadratic program

$$\inf_{\mu \in \mathbb{R}^d} \frac{1}{2} \|\mu\|_\Lambda^2 \tag{25}$$

$$\text{s.t. } \mu^\top \psi_\pi = 2\varepsilon. \tag{26}$$

Since $\psi_\pi \in \text{Im}(\Lambda)$, either $\|\psi_\pi\|_{\Lambda^\dagger} = 0$ and the bound is trivial, or $\|\psi_\pi\|_{\Lambda^\dagger} \neq 0$ and Lemma C.8 yields the bound.

$\square$

Remark C.4. The proof shows that interestingly, the bound of Proposition C.2 does not depend on θ.

Remark C.5. If $\psi_\pi \notin \text{Im}(\Lambda)$, Lemma C.8 entails $T \geq \infty$, and there is no (ϵ, δ)-PAC estimator of v^π.

This sample complexity lower bound can be rewritten as a bound on the optimal estimation accuracy for T samples. When $\psi_\pi \in \text{Im}(\Lambda)$ and $\hat{v}$ is an (ϵ, δ)-PAC estimator of v^π, then

$$\varepsilon \geq \|\psi_\pi\|_{\Lambda^\dagger} \frac{\sigma\sqrt{\text{kl}(\delta, 1 - \delta)}}{\sqrt{2T}}.$$

In other words, for any $\delta \in (0, 1/2]$, there is no (ε, δ)−PAC estimator with

$$\varepsilon < \|\psi_\pi\|_{\Lambda^\dagger} \frac{\sigma\sqrt{\text{kl}(\delta, 1 - \delta)}}{\sqrt{2T}}.$$

Remark C.6. Contrary to the lower bound of Theorem 5 in (Duan et al., 2020) for linear Markov decision processes, our lower bound does not require additional assumptions on the set of contexts.

C.2. Matching finite-sample error bound

We assume in this section that Λ is invertible, and the noise is now only assumed σ-subgaussian. Recall that $v^\pi = \theta^\top \psi_\pi$ for $\psi_\pi = \sum_{i,j} \omega_{i,j}^\pi \psi_{i,j}$. Let $\tau > 0$. Since ψ_π is known, a natural estimator of v^π is $\hat{v} = \hat{\theta}^\top \psi_\pi$ where

$$\hat{\theta} = \left(\sum_{t=1}^T \psi_{i_t,j_t} \psi_{i_t,j_t}^\top + \tau I_d \right)^{-1} \left(\sum_{t=1}^T r_t \psi_{i_t,j_t} \right)$$

is a regularized least squares estimator of θ. Note that $\sum_{t=1}^T \psi_{i_t,j_t} \psi_{i_t,j_t}^\top + \tau I_d$ is invertible for any $\tau > 0$ since $\sum_{t=1}^T \psi_{i_t,j_t} \psi_{i_t,j_t}^\top$ is a positive semi-definite matrix with nonnegative eigenvalues. Let $\widehat{\Lambda}_\tau := \frac{1}{T} \left(\sum_{t=1}^T \psi_{i_t,j_t} \psi_{i_t,j_t}^\top + \tau I_d \right)$, $L_\psi := \max_{i,j} \psi_{i,j}^\top \Lambda^{-1} \psi_{i,j}$ and $L_\theta := \max_{i,j} |\psi_{i,j}^\top \theta|$.

Theorem C.7. *Assume that $T \geq 18 L_\psi \log(8d/\delta)$ and $\tau \leq L_\psi \log(8d/\delta) \lambda_{\min}(\Lambda)$. With probability at least $1 - \delta$,*

$$|v^\pi - \hat{v}| \leq \frac{1}{\sqrt{T}} E_1 + \frac{1}{T} E_2 + \frac{1}{T^{3/2}} E_3 + \frac{1}{T^2} E_4$$

with $E_1 := \|\psi_\pi\|_{\Lambda^{-1}} \sqrt{2\sigma^2 \log(8/\delta)}$, $E_2 := \|\psi_\pi\|_{\Lambda^{-1}} \left(L_\psi L_\theta + \sigma(1 + 4\sqrt{2d})\sqrt{L_\psi} \right) \log(16(d+1)/\delta)$, $E_3 := \|\psi_\pi\|_{\Lambda^{-1}} \left(\frac{\sigma\sqrt{2}}{3} + 4\sigma\sqrt{d} + 4\sqrt{L_\psi}L_\theta \right) L_\psi \log^{3/2}(16(d+1)/\delta)$, $E_4 := \|\psi_\pi\|_{\Lambda^{-1}} \frac{4\sigma\sqrt{2}}{3}\sqrt{d}L_\psi^{3/2} \log^2(16(d+1)/\delta)$. In particular,

$$|v^\pi - \hat{v}| \leq \|\psi_\pi\|_{\Lambda^{-1}} \sqrt{\frac{2\sigma^2 \log(8/\delta)}{T}} + O(1/T)$$

with probability at least $1 - \delta$.

This error bound matches the sampling lower bound of Theorem C.3 up to an universal constant. This result is similar to the error bound from (Duan et al., 2020), but it requires the use of variants of the classical Freedman's inequality adapted to subgaussian random variables, which are stated in Appendix C.3. This is due to the fact that we assume in our work that the noise is subgaussian instead of bounded.

Proof of Theorem C.7. Note that

$$\hat{\theta} - \theta = \left(\sum_{t=1}^T \psi_{i_t,j_t} \psi_{i_t,j_t}^\top + \tau I_d \right)^{-1} \left(\sum_{t=1}^T (r_t - \psi_{i_t,j_t}^\top \theta)\psi_{i_t,j_t} - \tau\theta \right)$$

$$= \frac{1}{T} \widehat{\Lambda}_\tau^{-1} \left(\sum_{t=1}^T \xi_t \psi_{i_t,j_t} - \tau\theta \right).$$

Thus, we have

$$\hat{v} - v^\pi = \psi_\pi^\top \left(\hat{\theta} - \theta \right)$$

$$= \frac{1}{T} \psi_\pi^\top \widehat{\Lambda}_\tau^{-1} \left(\sum_{t=1}^T \xi_t \psi_{i_t,j_t} - \tau\theta \right)$$

$$= L_1 + L_2$$

for $L_1 := \frac{1}{T} \sum_{t=1}^T \xi_t \psi_\pi^\top \Lambda^{-1} \psi_{i_t,j_t}$ and $L_2 := \frac{1}{T} \sum_{t=1}^T \xi_t \psi_\pi^\top (\Lambda^{-1} - \widehat{\Lambda}_\tau^{-1}) \psi_{i_t,j_t} + \frac{\tau}{T} \psi_\pi^\top \widehat{\Lambda}_\tau^{-1} \theta$.

Control of L_1. To control this term, which turns out to be the dominant term in the bound, we will use Proposition C.9. Let $\mathcal{F}_{t-1} = \sigma(i_1, j_1, r_1, \ldots, i_{t-1}, j_{t-1}, r_{t-1}, i_t, j_t)$, $U_t = \psi_\pi^\top \Lambda^{-1} \psi_{i_t,j_t}$, $e_t = \xi_t U_t$ and $V_t = \text{Var}(e_t | \mathcal{F}_{t-1})$. We first note

that for any choice of $\gamma > 0$,

$$\mathbb{P}\left(|L_1| \geq \varepsilon\right) \leq \mathbb{P}\left(\left|\sum_{t=1}^{T} e_t\right| \geq T\varepsilon \text{ and } \sum_{t=1}^{T} V_t \leq \gamma^2\right) + \mathbb{P}\left(\sum_{t=1}^{T} V_t \geq \gamma^2\right).$$

Since ξ_t is independent of $\mathcal{F}_{t-1}$, we have

$$\mathbb{E}\left[e_t|\mathcal{F}_{t-1}\right] = U_t\mathbb{E}\left[\xi_t\right] = 0,$$

so that (e_t) is a martingale difference sequence with respect to $(\mathcal{F}_t)$. Consequently,

$$\begin{aligned} V_t &= \mathbb{E}\left[e_t^2|\mathcal{F}_{t-1}\right] \\ &= U_t^2\mathbb{E}\left[\xi_t^2\right] \\ &= \sigma^2 U_t^2, \end{aligned}$$

and

$$\begin{aligned} \mathbb{E}\left[e^{ce_t}|\mathcal{F}_{t-1}\right] &= \mathbb{E}\left[e^{c\xi_t U_t}|\mathcal{F}_{t-1}\right] \\ &= \mathbb{E}\left[e^{c\xi_t U_t}|U_t\right] \\ &\leq e^{c^2 V_t/2}, \end{aligned}$$

where the second to last inequality follows from the $\mathcal{F}_{t-1}$-measurability of U_t and the independence of ξ_t from $\mathcal{F}_{t-1}$, and the last inequality follows from the fact that ξ_t is σ-subgaussian and independent of U_t, and from $U_t^2 = V_t$. Hence, by Proposition C.9,

$$\mathbb{P}\left(\left|\sum_{t=1}^{T} e_t\right| \geq T\varepsilon \text{ and } \sum_{t=1}^{T} V_t \leq \gamma^2\right) \leq 2\exp\left(-\frac{T^2\varepsilon^2}{2\gamma^2}\right)$$

for any $\varepsilon, \gamma > 0$. Furthermore, $\sum_{t=1}^{T} V_t = \sigma^2 T \psi_\pi^\top \Lambda^{-1} \widehat{\Lambda}_0 \Lambda^{-1} \psi_\pi \leq \sigma^2 T \|\psi_\pi\|_{\Lambda^{-1}}^2 \|\Lambda^{-1/2}\widehat{\Lambda}_0\Lambda^{-1/2}\|_{\mathrm{op}}$. Indeed, since $\Lambda^{-1/2}\widehat{\Lambda}_0\Lambda^{-1/2}$ is a symmetric positive semi-definite matrix, the min-max theorem entails $\|\Lambda^{-1/2}\widehat{\Lambda}_0\Lambda^{-1/2}\|_{\mathrm{op}} = \lambda_{\max}\left(\Lambda^{-1/2}\widehat{\Lambda}_0\Lambda^{-1/2}\right) \geq \frac{z^\top \Lambda^{-1/2}\widehat{\Lambda}_0\Lambda^{-1/2}z}{z^\top z}$ for $z = \Lambda^{-1/2}\psi_\pi$. Lemma B.5 from (Duan et al., 2020) consequently ensures that $\sum_{t=1}^{T} V_t \leq \gamma^2$ with probability larger than $1 - \delta/4$ when γ is chosen such that

$$\gamma^2 = \sigma^2 T \|\psi_\pi\|_{\Lambda^{-1}}^2 (1 + K_{\delta/4})$$

where $K_{\delta/4} := \sqrt{\frac{2\log(8d/\delta)L_\psi}{T}} + \frac{2\log(8d/\delta)L_\psi}{3T}$.

Finally, we choose ε adequately with $\varepsilon := \frac{\gamma}{T}\sqrt{2\log(8/\delta)} = \|\psi_\pi\|_{\Lambda^{-1}}\sqrt{\frac{2\sigma^2\log(8d/\delta)}{T}}\sqrt{1 + K_{\delta/4}}$ so that

$$\mathbb{P}\left(|L_1| \geq \varepsilon\right) \leq \delta/4 + \delta/4 = \delta/2.$$

Finally, $\sqrt{1+x} \leq 1 + x/2$ for all $x \geq 0$ implies that $|L_1| \leq \|\psi_\pi\|_{\Lambda^{-1}}\sqrt{\frac{2\sigma^2\log(8d/\delta)}{T}}\left(1 + \frac{K_{\delta/4}}{2}\right)$ with probability larger than $1 - \delta/2$.

Control of L_2. Let us see that $L_2 = O(1/T)$ with probability larger than $1 - \delta/2$ by adapting the proof of Lemma B.6 from (Duan et al., 2020). Let $\Delta = \Lambda^{-1} - \widehat{\Lambda}_\tau^{-1}$ and $W = \frac{1}{T}\sum_{t=1}^{T} \xi_t \psi_{i_t,j_t}$, so that $\frac{1}{T}\sum_{t=1}^{T} \xi_t \psi_\pi^\top (\Lambda^{-1} - \widehat{\Lambda}_\tau^{-1})\psi_{i_t,j_t} = \psi_\pi^\top \Delta W = \left(\psi_\pi^\top \Lambda^{-1/2}\right)\left(\Lambda^{1/2}\Delta\Lambda^{1/2}\right)\left(\Lambda^{-1/2}W\right)$. Note also that $\frac{\tau}{T}\psi_\pi^\top \widehat{\Lambda}_\tau^{-1}\theta = \frac{\tau}{T}\left(\psi_\pi^\top \Lambda^{-1/2}\right)\left(\Lambda^{1/2}\widehat{\Lambda}_\tau^{-1}B^{1/2}\Lambda^{-1}\right)\left(\Lambda^{1/2}\theta\right)$. By the Cauchy-Schwartz inequality and the definition of the operator norm,

$$\begin{aligned} |L_2| &\leq \|\psi_\pi^\top \Lambda^{-1/2}\|_2 \|\Lambda^{1/2}\Delta\Lambda^{1/2}\Lambda^{-1/2}W\|_2 + \frac{\tau}{T}\|\psi_\pi^\top \Lambda^{-1/2}\|_2 \|\Lambda^{1/2}\widehat{\Lambda}_\tau^{-1}\Lambda^{1/2}\|_{\mathrm{op}} \|\Lambda^{-1}\|_{\mathrm{op}} \|\Lambda^{1/2}\theta\|_2 \\ &\leq \|\psi_\pi\|_{\Lambda^{-1}} \|\Lambda^{1/2}\Delta\Lambda^{1/2}\|_{\mathrm{op}} \|\Lambda^{-1/2}W\|_2 + \frac{\tau}{\lambda_{\min}(\Lambda)T}\|\psi_\pi\|_{\Lambda^{-1}} \|\Lambda^{1/2}\widehat{\Lambda}_\tau^{-1}\Lambda^{1/2}\|_{\mathrm{op}} \|\Lambda^{1/2}\theta\|_2 \\ &\leq \|\psi_\pi\|_{\Lambda^{-1}}\left(\|\Lambda^{1/2}\Delta\Lambda^{1/2}\|_{\mathrm{op}} \|\Lambda^{-1/2}W\|_2 + \frac{\tau L_\theta}{\lambda_{\min}(\Lambda)T}\left(1 + \|\Lambda^{1/2}\Delta\Lambda^{1/2}\|_{\mathrm{op}}\right)\right), \end{aligned}$$

where we have used $\|\Lambda^{1/2}\theta\|_2^2 = \theta^\top \Lambda \theta = \sum_{i,j} \omega_{i,j} \left(\psi_{i,j}^\top \theta\right)^2 \leq L_\theta^2$ for the last inequality.

The proof of Lemma B.6. from (Duan et al., 2020) ensures that as long as $T \geq 18L_\psi \log (8d/\delta)$ and $\tau \leq L_\psi \log(8d/\delta)\lambda_{\min}(\Lambda)$, we have $\|\Lambda^{1/2}\Delta\Lambda^{1/2}\|_{\text{op}} \leq 4\sqrt{\frac{\log(8d/\delta)L_\psi}{T}}$ with probability larger than $1 - \delta/4$. To control $\|\Lambda^{-1/2}W\|_2$, we first write $\Lambda^{-1/2}W = \frac{1}{T}\sum_{t=1}^T W_t$ where $W_t := \Lambda^{-1/2}\xi_t \psi_{i_t,j_t}$ defines a martingale difference sequence in $\mathbb{R}^d$ with respect to the filtration $(\mathcal{F}_t)_{t\geq 0}$ that was defined earlier. We will use the concentration inequality of Proposition C.11 to perform the analysis.

Let $P_t := \Lambda^{-1/2}\psi_{i_t,j_t}$. Note that P_t is $\mathcal{F}_{t-1}$-measurable, which implies that $W_t = \xi_t P_t$ defines a conditionally symmetric martingale difference sequence with respect to $(\mathcal{F}_t)$. In our case, we have $V_{1,t} = \mathbb{E}\left[W_t W_t^\top | \mathcal{F}_{t-1}\right] = \sigma^2 P_t P_t^\top \in \mathbb{R}^{d\times d}$ and $V_{2,t} = \mathbb{E}\left[W_t^\top W_t | \mathcal{F}_{t-1}\right] = \sigma^2 P_t^\top P_t \in \mathbb{R}$. Let us choose $Z_{1,t} := \xi_t \sqrt{P_t P_t^\top}$ and $Z_{2,t} := \xi_t \sqrt{P_t^\top P_t}$. By Lemma 4.3 of (Tropp, 2011b),

$$\mathbb{E}\left[e^{cZ_{1,t}}|\mathcal{F}_{t-1}\right] = \mathbb{E}\left[e^{c\xi_t \sqrt{P_t P_t^\top}}|P_t\right] \preceq e^{c^2 V_{1,t}/2}$$

and

$$\mathbb{E}\left[e^{cZ_{2,t}}|\mathcal{F}_{t-1}\right] = \mathbb{E}\left[e^{c\xi_t \sqrt{P_t^\top P_t}}|P_t\right] \preceq e^{c^2 V_{2,t}/2}.$$

Since $P_t P_t^\top$ and $P_t^\top P_t$ have the same non-zero eigenvalues,

$$\left\|\sum_{t=1}^T P_t P_t^\top\right\|_{\text{op}} \leq \sum_{t=1}^T \|P_t P_t^\top\|_{\text{op}} = \sum_{t=1}^T P_t^\top P_t,$$

so that

$$\max\left(\left\|\sum_{t=1}^T V_{1,t}\right\|_{\text{op}}, \left\|\sum_{t=1}^T V_{2,t}\right\|_{\text{op}}\right) = \sigma^2 \sum_{t=1}^T P_t^\top P_t.$$

Applying Corollary C.11 yields

$$\mathbb{P}\left(\|\Lambda^{-1/2}W\|_2 \geq \varepsilon\right) \leq \mathbb{P}\left(\left\|\sum_{t=1}^T W_t\right\|_2 \geq T\varepsilon \text{ and } \sum_{t=1}^T P_t^\top P_t \leq \gamma^2\right) + \mathbb{P}\left(\sigma^2 \sum_{t=1}^T P_t^\top P_t \geq \gamma^2\right)$$

$$\leq 2(d+1)\exp\left(-\frac{T^2\varepsilon^2}{2\gamma^2}\right) + \mathbb{P}\left(\sigma^2 \sum_{t=1}^T P_t^\top P_t \geq \gamma^2\right)$$

for any $\varepsilon, \gamma > 0$. We can control $\sum_{t=1}^T P_t^\top P_t$ by adapting the proof of Lemma B.9 from (Duan et al., 2020). We have

$$\sum_{t=1}^T P_t^\top P_t = \sum_{t=1}^T \psi_{i_t,j_t} \Lambda^{-1} \psi_{i_t,j_t}$$

$$= Td + T \operatorname{Tr}\left(\Lambda^{-1/2}\widehat{\Lambda}_\tau \Lambda^{-1/2} - I_d\right)$$

$$\leq Td\left(1 + \|\Lambda^{-1/2}\hat{\Lambda}_\tau \Lambda^{-1/2} - I_d\|_{\text{op}}\right).$$

By Lemma B.5. of (Duan et al., 2020), $\sigma^2 \sum_{t=1}^T P_t^\top P_t \leq \gamma^2$ with probability larger than $1 - \delta/8$ when γ is such that

$$\gamma^2 = \sigma^2 Td(1 + K_{\delta/8}).$$

Then, we let $\varepsilon := \frac{\gamma}{T}\sqrt{2\log(16(d+1)/\delta)}$ so that $\mathbb{P}\left(\|\Lambda^{-1/2}W\|_2 \geq \varepsilon\right) \leq \delta/8 + \delta/8 = \delta/4$. In particular, $\|\Lambda^{-1/2}W\|_2 \geq \sqrt{\frac{2d\sigma^2 \log(16(d+1)/\delta)}{T}}\left(1 + \frac{K_{\delta/8}}{2}\right)$ with probability smaller than $\delta/4$. To ease the notation, let $C_1 := \log(8d/\delta)$ and $C_2 := \log(16(d+1)/\delta)$. Combining the previous result with the bound on $\|\Lambda^{1/2}\Delta\Lambda^{1/2}\|_{\text{op}}$ yields that with probability

larger than $1 - \delta/2$,

$$
\begin{aligned}
|L_2| &\leq \|\psi_\pi\|_{\Lambda^{-1}} \left(\|\Lambda^{1/2}\Delta\Lambda^{1/2}\|_{\mathrm{op}}\|\Lambda^{-1/2}W\|_2 + \frac{\tau L_\theta}{\lambda_{\min}(\Lambda) T} \left(1 + \|\Lambda^{1/2}\Delta\Lambda^{1/2}\|_{\mathrm{op}}\right) \right) \\
&\leq \frac{\|\psi_\pi\|_{\Lambda^{-1}}}{T} \left(4\sqrt{2d\sigma^2 C_2 C_1 L_\psi} \left(1 + \frac{K_{\delta/8}}{2}\right) + \frac{\tau L_\theta}{\lambda_{\min}(\Lambda)} \left(1 + 4\sqrt{\frac{C_1 L_\psi}{T}}\right) \right) \\
&\leq \frac{\|\psi_\pi\|_{\Lambda^{-1}}}{T} \left(4\sqrt{2d\sigma^2 C_2 C_1 L_\psi} \left(1 + \frac{K_{\delta/8}}{2}\right) + L_\psi L_\theta C_1 \left(1 + 4\sqrt{\frac{C_1 L_\psi}{T}}\right) \right),
\end{aligned}
$$

where we used $\tau \leq L_\psi C_1 \lambda_{\min}(\Lambda)$ for the last inequality. In the end,

$$
\begin{aligned}
|v^\pi - \hat{v}| &\leq |L_1| + |L_2| \\
&\leq \|\psi_\pi\|_{\Lambda^{-1}} \sqrt{\frac{2\sigma^2 C_1}{T}} \left(1 + \frac{K_{\delta/4}}{2}\right) + \frac{\|\psi_\pi\|_{\Lambda^{-1}}}{T} 4\sqrt{2d\sigma^2 L_\psi} C_2 \left(1 + \frac{K_{\delta/8}}{2}\right) \\
&\quad + \frac{\|\psi_\pi\|_{\Lambda^{-1}} L_\psi L_\theta C_1}{T} \left(1 + 4\sqrt{\frac{C_1 L_\psi}{T}}\right) \\
&\leq \|\psi_\pi\|_{\Lambda^{-1}} \left(\sqrt{\frac{2\sigma^2 C_1}{T}} + \frac{4\sqrt{2d\sigma^2 L_\psi}}{T} C_2 \right) \left(1 + \sqrt{\frac{C_2 L_\psi}{2T}} + \frac{C_2 L_\psi}{3T}\right) \\
&\quad + \frac{\|\psi_\pi\|_{\Lambda^{-1}} L_\psi L_\theta C_1}{T} \left(1 + 4\sqrt{\frac{C_1 L_\psi}{T}}\right) \\
&\leq \frac{1}{\sqrt{T}} E_1 + \|\psi_\pi\|_{\Lambda^{-1}} \frac{4\sqrt{2d\sigma^2 L_\psi}}{T} C_2 + \frac{L_\psi L_\theta C_1}{T} \|\psi_\pi\|_{\Lambda^{-1}} \left(1 + 4\sqrt{\frac{C_1 L_\psi}{T}}\right) \\
&\quad + \|\psi_\pi\|_{\Lambda^{-1}} \left(\sqrt{\frac{2\sigma^2 C_2}{T}} + \frac{4\sqrt{2d\sigma^2 L_\psi}}{T} C_2 \right) \left(\sqrt{\frac{C_2 L_\psi}{2T}} + \frac{C_2 L_\psi}{3T}\right) \\
&\leq \frac{1}{\sqrt{T}} E_1 + \frac{1}{T} E_2 + \frac{1}{T^{3/2}} E_3 + \frac{1}{T^2} E_4
\end{aligned}
$$

with probability larger than $1 - \delta$.

$\square$

C.3. Additional Lemmas

The following result is an extension of Lemma 9 from (Taupin et al., 2023) to positive semi-definite matrices.

Lemma C.8. *For any $\nu_1, \ldots, \nu_n \in \mathbb{R}^d$ and any symmetric positive semi-definite matrices $\Lambda_1, \ldots, \Lambda_n \in \mathbb{R}^{d \times d}$, we have*

$$
\inf_{\substack{\mu \in \mathbb{R}^{n \times d}, \\ \sum_{i=1}^n \mu_i^\top \nu_i = 2\varepsilon}} \frac{1}{2} \sum_{i=1}^n \|\mu_i\|_{\Lambda_i}^2 = \begin{cases} \frac{2\varepsilon^2}{\sum_{i=1}^n \|\nu_i\|_{\Lambda_i^\dagger}^2} & \text{if } \forall i \in [n], \nu_i \in \mathrm{Im}(\Lambda_i) \text{ and } \sum_{i=1}^n \|\nu_i\|_{\Lambda_i^\dagger}^2 \neq 0 \\ 0 & \text{if } \exists j \in [n], \nu_j \notin \mathrm{Im}(\Lambda_j). \end{cases}
$$

Proof of Lemma C.8. Let $L(\mu, \lambda) := \frac{1}{2} \sum_{i=1}^n \mu_i^\top \Lambda \mu_i - \lambda \left(\sum_{i=1}^n \mu_i^\top \nu_i - 2\varepsilon \right)$ be the Lagrangian associated to the quadratic program. The KKT conditions yield $\Lambda_i \mu_i = \lambda \nu_i$ and $\sum_{i=1}^n \mu_i^\top \nu_i = 2\varepsilon$.

- If $\nu_i \in \mathrm{Im}(\Lambda_i)$ for every $i \in [n]$, $\lambda \Lambda_i^\dagger \nu_i$ is a solution of $\Lambda_i \mu_i = \lambda \nu_i$ for any λ. Thus, for $\lambda^* = \frac{2\varepsilon}{\sum_{i=1}^n \nu_i^\top \Lambda_i^\dagger \nu_i}$ and $\mu_i^* = \lambda_* \Lambda_i^\dagger \nu_i$, we have $\Lambda_i \mu_i^* = \lambda^* \nu_i$ and

$$
\begin{aligned}
\sum_{i=1}^n \mu_i^{*\top} \nu_i &= \frac{2\varepsilon}{\sum_{i=1}^n \nu_i^\top \Lambda_i^\dagger \nu_i} \sum_{i=1}^n \left(\Lambda_i^\dagger \nu_i \right)^\top \nu_i \\
&= 2\varepsilon,
\end{aligned}
$$

where the last equality follows from the fact that $\Lambda_i^\dagger$ is symmetric (since Λ_i is symmetric). It follows that μ^* is optimal, and the infimum equals $\frac{1}{2}\sum_{i=1}^n \mu_i^{*\top}\Lambda_i\mu_i^* = \frac{1}{2}\lambda^*\sum_{i=1}^n \mu_i^{*\top}\nu_i = \frac{2\varepsilon^2}{\sum_{i=1}^n \nu_i^\top \Lambda_i^\dagger \nu_i}$. Note that $\Lambda_i^\dagger$ is positive semi-definite since Λ_i is, so we have $\nu_i^\top \Lambda_i^\dagger \nu_i = \|\nu_i\|_{\Lambda_i^\dagger}^2$.

- If $\nu_j \notin \text{Im}(\Lambda_j)$ for some $j \in [n]$, $\Lambda_j\mu_j = \lambda\nu_j$ implies $\lambda = 0$. Thus, any optimal point μ^* must verify $\Lambda_i\mu_i^* = 0$ for every $i \in [n]$ which implies that $\frac{1}{2}\sum_{i=1}^n \mu_i^{*\top}\Lambda_i\mu_i^* = 0$.

$\square$

The following concentration inequalities are useful in the analysis of the PE error bounds.

Proposition C.9 (Corollary of Theorem 2.6 from (Fan et al., 2015))**.** *Let* $(e_t)_{t=1,\dots,T}$ *be a martingale difference sequence with respect to a filtration* $(\mathcal{F}_t)_{t=0,\dots,T-1}$ *(i.e.,* $\mathbb{E}[e_t|\mathcal{F}_{t-1}] = 0$ *for every t), and let* $V_t = \text{Var}(e_t|\mathcal{F}_{t-1})$*. If for all* $c \geq 0$ *and all* $t = 1,\dots,T$*, we have* $\mathbb{E}[e^{ce_t}|\mathcal{F}_{t-1}] \leq e^{c^2 V_t/2}$*, then for all* $\varepsilon, \gamma > 0$*,*

$$\mathbb{P}\left(\left|\sum_{t=1}^T e_t\right| \geq \varepsilon \text{ and } \sum_{t=1}^T V_t \leq \gamma^2\right) \leq 2\exp\left(-\frac{\varepsilon^2}{2\gamma^2}\right).$$

Proof of Proposition C.9. Applying Theorem 2.6 from (Fan et al., 2015) with $V_t = \text{Var}(e_t|\mathcal{F}_{t-1})$, $f(c) = c^2/2$ yields, after optimizing the bound in c,

$$\mathbb{P}\left(\sum_{t=1}^T e_t \geq \varepsilon \text{ and } \sum_{t=1}^T V_t \leq \gamma^2\right) \leq \exp\left(-\frac{\varepsilon^2}{2\gamma^2}\right).$$

The same bound applies to the martingale difference sequence $(-e_t)$, and a union bound concludes the proof. $\square$

The following concentration inequality is an extension of Proposition C.9 to martingale difference sequences of $\mathbb{R}^{d\times d'}$ for any $d' \geq 1$. It can be obtained as a corollary of Theorem 2.3 in (Tropp, 2011a).

Proposition C.10. *Let* $(W_t)_{t=1,\dots,T}$ *be a conditionally symmetric[11] martingale difference sequence of* $\mathbb{R}^{d\times d'}$ *with respect to a filtration* $(\mathcal{F}_t)_{t=0,\dots,T-1}$*, and let* $V_{1,t} \in \mathbb{R}^{d\times d}$*,* $V_{2,t} \in \mathbb{R}^{d'\times d'}$ *be* $\mathcal{F}_{t-1}$*-measurable symmetric matrices. Denote by* $Z_{1,t} \in \mathbb{R}^{d\times d}$ *and* $Z_{2,t} \in \mathbb{R}^{d'\times d'}$ *any symmetric matrices such that* $Z_{1,t}^2 = W_t W_t^\top$ *and* $Z_{2,t}^2 = W_t^\top W_t$*. If for some nonnegative function g, for all* $c \in \mathbb{R}$ *and all* $t = 1,\dots,T$*,* $\mathbb{E}[e^{cZ_{1,t}}|\mathcal{F}_{t-1}] \preceq e^{g(c)V_{1,t}}$ [12] *and* $\mathbb{E}[e^{cZ_{2,t}}|\mathcal{F}_{t-1}] \preceq e^{g(c)V_{2,t}}$*, then for all* $\varepsilon, \gamma > 0$*,*

$$\mathbb{P}\left(\lambda_{\max}\left(\sum_{t=1}^T W_t\right) \geq \varepsilon \text{ and } \max\left(\lambda_{\max}\left(\sum_{t=1}^T V_{1,t}\right), \lambda_{\max}\left(\sum_{t=1}^T V_{2,t}\right)\right) \leq \gamma^2\right) \leq (d+d')\inf_{c\in\mathbb{R}}\exp\left(-c\varepsilon + g(c)\gamma^2\right).$$

Proof of Proposition C.10. Let $Y_t := \begin{pmatrix} 0 & W_t \\ W_t^\top & 0 \end{pmatrix}$ and $V_t := \begin{pmatrix} V_{1,t} & 0 \\ 0 & V_{2,t} \end{pmatrix}$. We will use Theorem 2.3 of (Tropp, 2011a) on the martingale difference sequence $(Y_t)_{t\geq 0}$ of $\mathbb{R}^{(d+d')^2}$. To do so, we must prove that for all $c \in \mathbb{R}$, $\mathbb{E}[e^{cY_t}|\mathbb{F}_{t-1}] \preceq e^{g(c)V_t}$. It is straightforward to check that for all $k \geq 1$,

$$Y_t^{2k} = \begin{pmatrix} (W_t W_t^\top)^k & 0 \\ 0 & (W_t^\top W_t)^k \end{pmatrix}$$

and

$$Y_t^{2k+1} = \begin{pmatrix} 0 & (W_t W_t^\top)^k W_t \\ (W_t W_t^\top)^k W_t^\top & 0 \end{pmatrix}.$$

---

[11]This means that W_t and $-W_t$ have the same conditional distribution with respect to $\mathcal{F}_{t-1}$ for every t.
[12]Recall that $N \preceq M$ means that $M - N$ is positive semi-definite.

Note also that the conditional symmetry of W_t implies that $\mathbb{E}\left[Y_t^{2k+1}|\mathcal{F}_{t-1}\right] = 0$. Consequently,

$$
\begin{aligned}
\mathbb{E}\left[e^{cY_t}|\mathcal{F}_{t-1}\right] &= \mathbb{E}\left[\sum_{k=0}^{+\infty}\frac{c^{2k}}{(2k)!}Y_t^{2k} + \sum_{k=0}^{+\infty}\frac{c^{2k+1}}{(2k+1)!}Y_t^{2k+1}|\mathcal{F}_{t-1}\right] \\
&= \mathbb{E}\left[\sum_{k=0}^{+\infty}\frac{c^{2k}}{(2k)!}\begin{pmatrix}\left(W_tW_t^\top\right)^k & 0 \\ 0 & \left(W_t^\top W_t\right)^k\end{pmatrix}|\mathcal{F}_{t-1}\right] \\
&= \mathbb{E}\left[\begin{pmatrix}\cosh\left(cZ_{1,t}\right) & 0 \\ 0 & \cosh\left(cZ_{2,t}\right)\end{pmatrix}|\mathcal{F}_{t-1}\right].
\end{aligned}
$$

From the assumptions of Proposition C.10, we have

$$
\begin{aligned}
\mathbb{E}\left[\cosh\left(cZ_{1,t}\right)|\mathcal{F}_{t-1}\right] &= \frac{1}{2}\left(\mathbb{E}\left[\exp\left(cZ_{1,t}\right)|\mathcal{F}_{t-1}\right] + \mathbb{E}\left[\exp\left(-cZ_{1,t}\right)|\mathcal{F}_{t-1}\right]\right) \\
&\preceq e^{g(c)V_{1,t}}
\end{aligned}
$$

and similarly,

$$
\mathbb{E}\left[\cosh\left(cZ_{1,t}\right)|\mathcal{F}_{t-1}\right] \preceq e^{g(c)V_{2,t}}.
$$

This finally guarantees that

$$
\begin{aligned}
\mathbb{E}\left[e^{cY_t}|\mathcal{F}_{t-1}\right] &\preceq \begin{pmatrix}e^{g(c)V_{1,t}} & 0 \\ 0 & e^{g(c)V_{2,t}}\end{pmatrix} \\
&= e^{g(c)V_t}.
\end{aligned}
$$

We can now apply Theorem 2.3 of (Tropp, 2011a). It ensures that

$$
\mathbb{P}\left(\lambda_{\max}\left(\sum_{t=1}^T Y_t\right) \geq \varepsilon \text{ and } \lambda_{\max}\left(\sum_{t=1}^T V_t\right) \leq \gamma^2\right) \leq (d+d')\inf_{c\in\mathbb{R}}\exp\left(-c\varepsilon + g(c)\gamma^2\right).
$$

Note that

$$
\lambda_{\max}\left(\sum_{t=1}^T Y_t\right) = \lambda_{\max}\left(\sum_{t=1}^T W_t\right) \quad \text{and} \quad \lambda_{\max}\left(\sum_{t=1}^T V_t\right) = \max\left(\lambda_{\max}\left(\sum_{t=1}^T V_{1,t}\right), \lambda_{\max}\left(\sum_{t=1}^T V_{2,t}\right)\right),
$$

which concludes the proof. $\qquad\square$

The previous general result directly implies the following concentration inequality.

Corollary C.11. *Let $(W_t)_{t=1,\ldots,T}$ be a conditionally symmetric martingale difference sequence of $\mathbb{R}^{d\times d'}$ with respect to a filtration $(\mathcal{F}_t)_{t=0,\ldots,T-1}$, and let $V_{1,t} := \mathbb{E}\left[W_tW_t^\top|\mathcal{F}_{t-1}\right]$, $V_{2,t} := \mathbb{E}\left[W_t^\top W_t|\mathcal{F}_{t-1}\right]$. Denote by $Z_{1,t}\in\mathbb{R}^{d\times d}$ and $Z_{2,t}\in\mathbb{R}^{d'\times d'}$ any symmetric matrices such that $Z_{1,t}^2 = W_tW_t^\top$ and $Z_{2,t}^2 = W_t^\top W_t$. If for all $c\in\mathbb{R}$ and all $t = 1,\ldots,T$, $\mathbb{E}\left[e^{cZ_{1,t}}|\mathcal{F}_{t-1}\right] \preceq e^{c^2V_{1,t}/2}$ and $\mathbb{E}\left[e^{cZ_{2,t}}|\mathcal{F}_{t-1}\right] \preceq e^{c^2V_{2,t}/2}$, then for all $\varepsilon,\gamma > 0$,*

$$
\mathbb{P}\left(\left\|\sum_{t=1}^T W_t\right\|_{\mathrm{op}} \geq \varepsilon \text{ and } \max\left(\left\|\sum_{t=1}^T V_{1,t}\right\|_{\mathrm{op}}, \left\|\sum_{t=1}^T V_{2,t}\right\|_{\mathrm{op}}\right) \leq \gamma^2\right) \leq 2(d+d')\exp\left(-\frac{\varepsilon^2}{2\gamma^2}\right).
$$

Proof of Corollary C.11. We choose $g(c) = c^2/2$ in the statement of Proposition C.10. Note that $\inf_{c\in\mathbb{R}}\exp\left(-c\varepsilon + g(c)\gamma^2\right) = \exp\left(-\frac{\varepsilon^2}{2\gamma^2}\right)$, and that $V_{1,t}$ and $V_{2,t}$ are positive semi-definite matrices, which implies that $\lambda_{\max}\left(\sum_{t=1}^T V_{1,t}\right) = \|\sum_{t=1}^T V_{1,t}\|_{\mathrm{op}}$ and $\lambda_{\max}\left(\sum_{t=1}^T V_{2,t}\right) = \|\sum_{t=1}^T V_{2,t}\|_{\mathrm{op}}$. The same bound applies to the conditionally symmetric martingale difference sequence $(-W_t)$, and a union bound concludes the proof. $\qquad\square$

D. Proofs of results from Section 6.1 and 6.2 : Policy Evaluation Upper Bounds

D.1. Proof of Theorem 6.1

We first note that the max-norm guarantee[13] of Theorem 1 of (Stojanovic et al., 2024) can be generalized to the case where the samples are not necessarily selected uniformly at random by leveraging the two-to-infinity norm guarantees of Theorem 4.1. We state the general result below, where we recall that $L = \|M\|_{\max} \vee \sigma$ (note that the dependency in μ is also improved compared to Theorem 1 of (Stojanovic et al., 2024)).

Proposition D.1. *Let $\delta \in (0,1)$. If $T \geq \frac{cL^2\kappa^2(m+n)}{\sigma_r^2\omega_{\min}} \log^3\left(\frac{(m+n)T}{\delta}\right)$ for some universal constant $c > 0$, then with probability larger than $1 - \delta$,*

$$\|M - \widehat{M}\|_{\max} \lesssim \sqrt{\frac{L^2\mu^6\kappa^4 r^3(m+n)}{T\omega_{\min}(m \wedge n)^2} \log^3\left(\frac{(m+n)T}{\delta}\right)}.$$

Proof of Proposition D.1. The proof is divided in two steps: first, we leverage the bound on $\epsilon_{\text{Sub-Rec}}$ of Theorem 4.1 to bound $\|M - \widehat{M}\|_{2\to\infty}$, and then we bound $\|M - \widehat{M}\|_{\max}$ by relating it to $\epsilon_{\text{Sub-Rec}}$ and $\|M - \widehat{M}\|_{2\to\infty}$.

Bounding $\|M - \widehat{M}\|_{2\to\infty}$. Let $E := M - \widetilde{M}$. From (45) in (Stojanovic et al., 2024),

$$\|M - \widehat{M}\|_{2\to\infty} \leq \epsilon_{\text{Sub-Rec}}\left(\|E\|_{\text{op}} + \|M\|_{\text{op}}\right) + \|U\|_{2\to\infty}\|E\|_{\text{op}}.$$

Hence[14], with probability larger than $1 - \delta$, from Proposition B.5, and Theorem 4.1,

$$\|M - \widehat{M}\|_{2\to\infty} \lesssim \sigma_1\epsilon_{\text{Sub-Rec}} + \mu\sqrt{\frac{r}{m}}\|E\|_{\text{op}}$$

$$\lesssim \kappa\sqrt{\frac{L^2\mu^2\kappa^2 r(m+n)}{T\omega_{\min}(m \wedge n)} \log^3\left(\frac{(m+n)T}{\delta}\right)}$$

$$+ \mu\sqrt{\frac{r}{m}}L\sqrt{\frac{m+n}{T\omega_{\min}} \log^3\left(\frac{(m+n)T}{\delta}\right)}$$

$$\lesssim \sqrt{\frac{L^2\mu^2\kappa^4 r(m+n)}{T\omega_{\min}(m \wedge n)} \log^3\left(\frac{(m+n)T}{\delta}\right)}$$

under $T \gtrsim \frac{L^2(m+n)}{\sigma_r^2\omega_{\min}} \log^3\left(\frac{(m+n)T}{\delta}\right)$ (note that by Lemma B.8, this condition is stronger than the condition of Proposition B.5).

Bounding $\|M - \widehat{M}\|_{\max}$. We similarly write (see (Stojanovic et al., 2024)),

$$\|M - \widehat{M}\|_{\max} \leq \epsilon_{\text{Sub-Rec}}(\|M - \widehat{M}\|_{2\to\infty} + \|M\|_{2\to\infty}) + \|V\|_{2\to\infty}\|M - \widehat{M}\|_{2\to\infty}.$$

Note that by Lemma B.8, $\|M\|_{2\to\infty} \leq \sqrt{m}\|M\|_{\max} \leq \sigma_1\frac{\mu^2 r}{\sqrt{n}}$. Furthermore, under

$$T \gtrsim \frac{L^2\kappa^2(m+n)}{\sigma_r^2\omega_{\min}} \log^3\left(\frac{(m+n)T}{\delta}\right), \tag{27}$$

we know that by Theorem 4.1, $\epsilon_{\text{Sub-Rec}} \lesssim \frac{\mu\sqrt{r}}{\sqrt{m\wedge n}}$ with probability larger than $1 - \delta$. Thus, with probability larger than $1 - \delta$, under (27),

---

[13] The max-norm is denoted by $\|.\|_\infty$ in (Stojanovic et al., 2024).

[14] Union bounds are performed throughout the proof. Since the relevant inequalities are stated up to a constant, they are not impacted.

$$\|\widehat{M} - M\|_{\max} \lesssim \left(\epsilon_{\text{Sub-Rec}} + \mu\sqrt{\frac{r}{n}} \right) \|M - \widehat{M}\|_{2\to\infty} + \sigma_1 \epsilon_{\text{Sub-Rec}} \frac{\mu^2 r}{\sqrt{n}}$$

$$\lesssim \mu\sqrt{\frac{r}{m \wedge n}} \sqrt{\frac{L^2 \mu^2 \kappa^4 r(m+n)}{T\omega_{\min}(m \wedge n)} \log^3\left(\frac{(m+n)T}{\delta}\right)}$$

$$+ \frac{\sigma_1}{\sigma_r} \frac{\mu^2 r}{\sqrt{n}} \sqrt{\frac{L^2 \mu^2 \kappa^2 r(m+n)}{T\omega_{\min}(m \wedge n)} \log^3\left(\frac{(m+n)T}{\delta}\right)}$$

$$\lesssim \sqrt{\frac{L^2 \mu^6 \kappa^4 r^3(m+n)}{T\omega_{\min}(m \wedge n)^2} \log^3\left(\frac{(m+n)T}{\delta}\right)}.$$

$\square$

Proof of Theorem 6.1. Since $\sum_{i,j} \omega_{i,j}^\pi = 1$, we have $|v^\pi - \hat{v}_{\text{SIPS}}| = \left| \sum_{i,j} \omega_{i,j}^\pi \left(M_{i,j} - \widehat{M}_{i,j} \right) \right| \leq \|M - \widehat{M}\|_{\max}$. Proposition D.1 entails that $|v^\pi - \hat{v}_{\text{SIPS}}| \leq \varepsilon$ with probability larger than $1 - \delta$ as soon as

$$T \gtrsim \frac{L^2 \mu^6 \kappa^4 r^3(m+n)}{\omega_{\min}(m \wedge n)^2 \varepsilon^2} \log^3\left(\frac{(m+n)T}{\delta}\right)$$

and

$$T \gtrsim \frac{L^2 \kappa^2(m+n)}{\sigma_r^2 \omega_{\min}} \log^3\left(\frac{(m+n)T}{\delta}\right).$$

By Lemma B.8, $\frac{\kappa^2}{\sigma_r^2} \leq \frac{\mu^4 \kappa^2 r^2}{mn\|M\|_{\max}^2} \leq \frac{\mu^6 \kappa^4 r^3}{(m \wedge n)^2 \|M\|_{\max}^2}$. Since $\varepsilon \leq \|M\|_{\max}$, the second condition is weaker and can be dropped. $\square$

D.2. Proof of Theorem 6.2

Recall that $M_{i,j} = \phi_{i,j}^\top \theta + \epsilon_{i,j}$. In this appendix, the following additional notations will be used:

$$\Lambda = \sum_{i,j} \omega_{i,j}\psi_{i,j}\psi_{i,j}^\top, \Lambda_\phi = \sum_{i,j} \omega_{i,j}\phi_{i,j}\phi_{i,j}^\top, \widehat{\Lambda}_\tau = \frac{1}{T - T_1}\left(\sum_{t=T_1+1}^T \phi_{i_t,j_t}\phi_{i_t,j_t}^\top + \tau I_d \right) \text{ and } \phi_\pi = \sum_{i,j}\omega_{i,j}^\pi\phi_{i,j}.$$

We also recall that

$$\epsilon_{\max} = \max_{i\in[m],j\in[n]} |\epsilon_{i,j}|$$

is the misspecification of the misspecified linear bandit, and that the latter is of dimension $d = r(m+n) - r^2$. Finally, to ease the notation, we do not write the dependency in π of the estimator $\hat{v}_{\text{RS-PE}}^\pi$. We summarize RS-PE in Algorithm 3.

Algorithm 3 RECOVER SUBSPACE FOR POLICY EVALUATION (RS-PE)

Input: Samples $(i_t, j_t, r_t)_{t\in[T]}$, context distribution ρ, behavior policy π^b, target policy π, sample size T_1, regularization parameter τ, confidence level δ;
Phase 1: Subspace recovery
Use the first T_1 samples to construct $\widehat{U}, \widehat{V}$ as in (2)
Phase 2: Solving a misspecified linear bandit
Use $\widehat{U}, \widehat{V}$ to construct $\{\phi_{i,j} : (i,j) \in [m] \times [n]\}$ as in (7)
Use the remaining $T - T_1$ samples to construct the least squares estimator $\hat{\theta}$ as in (10)
Output: $\hat{v}_{\text{RS-PE}}^\pi = \sum_{i,j} w_{i,j}^\pi \phi_{i,j}^\top \hat{\theta}$.

The general idea of the proof of Theorem 6.2 is to decompose the PE error of RS-PE as a sum of a term that depends on the misspecification error and a term incurred by the error between $\hat{\theta}$ and θ. The analysis of the latter term, which is

in fact the largest contributor to the overall error, is heavily inspired by the analysis in Appendix C.2. We first derive a high probability guarantee with the lower-order term that one would expect from the contextual linear bandit analysis, i.e., roughly $\|\phi_\pi\|_{\Lambda_\phi^{-1}} \sqrt{\frac{\log(1/\delta)}{T}}$. This bound is however not satisfactory in itself since it is a function of the randomness of the first phase of the algorithm. This issue can be resolved by leveraging the subspace recovery guarantees of Theorem 4.1 to show that $\|\phi_\pi\|_2$ converges towards $\|\psi_\pi\|_2$.

The majority of this section is thus devoted to proving the following intermediary result.

Proposition D.2. *Assume that $T_1 = \lfloor \alpha T \rfloor$ for some $\alpha \in (0,1)$ and that the regularization parameter of $\widehat{\Lambda}_\tau$ verifies $\tau \leq \frac{r}{(m \wedge n)} \frac{\omega_{\min}}{\omega_{\max}} \log(16d/\delta)$. There exists a universal constant $C > 0$ such that as long as $T \geq \frac{CL^2\kappa^2(m+n)}{\sigma_r^2 \omega_{\min}} \log^3 \left(\frac{(m+n)T}{\delta} \right)$ and $T \geq \frac{C}{\omega_{\min}} \log(16d/\delta)$, then with probability larger than $1 - \delta$,*

$$|v^\pi - \hat{v}_{\text{RS-PE}}| \leq \|\phi_\pi\|_{\Lambda_\phi^{-1}} \sqrt{\frac{2\sigma^2 \log(16/\delta)}{(1-\alpha)T}} + K_0 \frac{\log^3((m+n)T/\delta)}{T}$$

where

$$K_0 \lesssim \frac{L^2}{\|M\|_{\max}} \frac{\mu^6 \kappa^4 r^3 (m+n)^2 \sqrt{mn}}{(\omega_{\min}mn)^2 (m \wedge n)}. \tag{28}$$

Proof of Proposition D.2. Note that $v^\pi = \sum_{i,j} \omega_{i,j}^\pi \left(\phi_{i,j}^\top \theta + \epsilon_{i,j} \right) = \phi_\pi^\top \theta + \sum_{i,j} \omega_{i,j}^\pi \epsilon_{i,j}$. One can further remark that

$$\hat{\theta} - \theta = \frac{1}{T_2} \widehat{\Lambda}_\tau^{-1} \left(\sum_{t=T_1+1}^{T} (\xi_t + \epsilon_{i_t,j_t}) \phi_{i_t,j_t} - \tau\theta \right)$$

for $T_2 := T - T_1$. Consequently,

$$|v^\pi - \hat{v}_{\text{RS-PE}}| \leq \left| \phi_\pi^\top \left(\hat{\theta} - \theta \right) \right| + \sum_{i,j} \left| \omega_{i,j}^\pi \epsilon_{i,j} \right|$$

$$\leq \left| \frac{1}{T_2} \sum_{t=T_1+1}^{T} \phi_\pi^\top (\xi_t + \epsilon_{i_t,j_t}) \widehat{\Lambda}_\tau^{-1} \phi_{i_t,j_t} \right| + \frac{\tau}{T_2} \phi_\pi^\top \widehat{\Lambda}_\tau^{-1} \theta + \epsilon_{\max}$$

$$\leq \left| \frac{1}{T_2} \sum_{t=T_1+1}^{T} \xi_t \phi_\pi^\top \widehat{\Lambda}_\tau^{-1} \phi_{i_t,j_t} \right| + \frac{\tau}{T_2} \phi_\pi^\top \widehat{\Lambda}_\tau^{-1} \theta \tag{29}$$

$$+ \epsilon_{\max} \left(1 + \left| \frac{1}{T_2} \sum_{t=T_1+1}^{T} \phi_\pi^\top \widehat{\Lambda}_\tau^{-1} \phi_{i_t,j_t} \right| \right). \tag{30}$$

Error decomposition. Let $\Delta := \Lambda_\phi^{-1} - \widehat{\Lambda}_\tau^{-1}$. We consider the following decomposition of the error:

$$|v^\pi - \hat{v}_{\text{RS-PE}}| \leq L_1 + L_2 + L_3$$

where

$$L_1 = \frac{1}{T_2} \sum_{t=T_1+1}^{T_2} \xi_t \phi_\pi^\top \Lambda_\phi^{-1} \phi_{i_t,j_t},$$

$$L_2 = \frac{1}{T_2} \sum_{t=T_1+1}^{T_2} \xi_t \phi_\pi^\top \Delta \phi_{i_t,j_t} + \frac{\tau}{T_2} \phi_\pi^\top \widehat{\Lambda}_\tau^{-1} \theta$$

29

and

$$L_3 = \epsilon_{\max}\left(1 + \left|\frac{1}{T_2}\sum_{t=T_1+1}^{T}\phi_\pi^\top\widehat{\Lambda}_\tau^{-1}\phi_{i_t,j_t}\right|\right).$$

As discussed before, we first perform the analysis conditionally on the first phase of the algorithm. We will then control the misspecification error induced by the first phase with Corollary 5.1. Let $L_\phi := \max_{i,j}\phi_{i,j}^\top\Lambda_\phi^{-1}\phi_{i,j}$, $L_\theta := \max_{i,j}|\phi_{i,j}^\top\theta|$. Adapting the proof of Theorem C.7 yields the following lemma.

Lemma D.3. *As long as $T_2 \geq \frac{18}{\omega_{\min}}\log(16d/\delta)$ and $\tau \leq \frac{r}{(m\wedge n)}\frac{\omega_{\min}}{\omega_{\max}}\log(16d/\delta)$, then with probability at least $1 - \delta/2$,*

$$|L_1| + |L_2| \leq \frac{E_1}{\sqrt{T_2}} + \frac{E_2}{T_2} + \frac{E_3}{T_2^{3/2}} + \frac{E_4}{T_2^2}$$

and

$$|L_3| \leq \epsilon_{\max}\left(1 + \sqrt{L_\phi}\|\phi_\pi\|_{\Lambda_\phi^{-1}}\left(1 + 4\sqrt{\frac{\log(16d/\delta)}{T_2}}\right)\right)$$

with $E_1 := \|\phi_\pi\|_{\Lambda_\phi^{-1}}\sqrt{2\sigma^2\log(16/\delta)}, E_2 := \|\phi_\pi\|_{\Lambda_\phi^{-1}}\left(L_\phi L_\theta + \sigma(1 + 4\sqrt{2d})\sqrt{L_\phi}\right)\log(32(d+1)/\delta), E_3 := \|\phi_\pi\|_{\Lambda_\phi^{-1}}\left(\frac{\sigma\sqrt{2}}{3} + 4\sigma\sqrt{d} + 4\sqrt{L_\phi}L_\theta\right)L_\phi\log^{3/2}(32(d+1)/\delta), E_4 := \|\phi_\pi\|_{\Lambda_\phi^{-1}}\frac{4\sigma\sqrt{2}}{3}\sqrt{d}L_\phi^{3/2}\log^2(32(d+1)/\delta).$

Remark D.4. To see why we must split the data for the two phases in the analysis, let $\mathcal{F}_t = \sigma(i_1, j_1, r_1, \ldots, i_{T_1}, j_{T_1}, r_{T_1}, \ldots, i_t, j_t, r_t, i_{t+1}, j_{t+1})$ for $t = T_1, \ldots, T - 1$. Because Λ_ϕ^{-1} is $\mathcal{F}_{T_1}$-measurable, $e_t := \phi_\pi^\top\xi_t\Lambda_\phi^{-1}\phi_{i_t,j_t}$ defines a martingale difference sequence with respect to $(\mathcal{F}_t)_{t=T_1,\ldots,T-1}$ and the contextual linear bandit analysis applies. If all the samples were used to build ϕ, e_t would not define a martingale difference sequence, which means that we could not use our concentration results.

Proof of Lemma D.3. The control of these terms closely follows the proof of theorem C.7; in particular, we have

$$|L_2| \leq \|\phi_\pi\|_{\Lambda_\phi^{-1}}\left(\|\Lambda_\phi^{1/2}\Delta\Lambda_\phi^{1/2}\|_{\mathrm{op}}\|\Lambda_\phi^{-1/2}W\|_2 + \frac{\tau L_\theta}{\lambda_{\min}(\Lambda_\phi)T_2}\left(1 + \|\Lambda^{1/2}\Delta\Lambda_\phi^{1/2}\|_{\mathrm{op}}\right)\right).$$

for $W = \frac{1}{T_2}\sum_{t=T_1+1}^{T_2}\xi_t\phi_{i_t,j_t}$. The only difference is the additional term L_3, which can be controlled similarly to L_2. Indeed, by Cauchy-Schwartz and the definition of the operator norm,

$$\frac{1}{T_2}\sum_{t=T_1+1}^{T}\phi_\pi^\top\widehat{\Lambda}_\tau^{-1}\phi_{i_t,j_t} = \frac{1}{T_2}\sum_{t=T_1+1}^{T}\phi_\pi^\top\Lambda_\phi^{-1}\phi_{i_t,j_t} + \frac{1}{T_2}\sum_{t=T_1+1}^{T}\phi_\pi^\top\left(\widehat{\Lambda}_\tau^{-1} - \Lambda_\phi^{-1}\right)\phi_{i_t,j_t}$$

$$\leq \frac{1}{T_2}\sum_{t=T_1+1}^{T}\|\phi_\pi\|_{\Lambda_\phi^{-1}}\sqrt{\phi_{i_t,j_t}^\top\Lambda_\phi^{-1}\phi_{i_t,j_t}}$$

$$+ \frac{1}{T_2}\sum_{t=T_1+1}^{T}\|\Lambda_\phi^{-1/2}\phi_\pi\|_2\|\Lambda_\phi^{1/2}\Delta\Lambda_\phi^{1/2}\|_{\mathrm{op}}\|\Lambda_\phi^{-1/2}\phi_{i_t,j_t}\|_2$$

$$\leq \sqrt{L_\phi}\|\phi_\pi\|_{\Lambda_\phi^{-1}}\left(1 + \|\Lambda_\phi^{1/2}\Delta\Lambda_\phi^{1/2}\|_{\mathrm{op}}\right).$$

Thus, $L_3 \leq \epsilon_{\max}\left(1 + \sqrt{L_\phi}\|\phi_\pi\|_{\Lambda_\phi^{-1}}\left(1 + \|\Lambda_\phi^{1/2}\Delta\Lambda_\phi^{1/2}\|_{\mathrm{op}}\right)\right).$

By the same arguments as in the proof of Lemma B.6. from (Duan et al., 2020), if $T_2 \geq 18L_\phi\log(16d/\delta)$ and $\tau \leq L_\phi\log(16d/\delta)\lambda_{\min}(\Lambda_\phi)$, then $\|\Lambda_\phi^{1/2}\Delta\Lambda_\phi^{1/2}\|_{\mathrm{op}} \leq 4\sqrt{\frac{\log(16d/\delta)L_\phi}{T_2}}$ with probability larger than $1 - \delta/8$.

We can now derive high probability bounds on L_1, L_2 and L_3 by adapting the contextual linear bandit analysis. The main difference is that here, the analysis is performed conditionally on the randomness of the first phase of the algorithm. One subtlety to note is that we need to use a version of Proposition C.9 that is conditional on the first T_1 samples because the

choices of γ and ε in the analysis depend on ϕ, which depends on the estimator $\widehat{M}$ built with the first T_1 samples. It is not an issue as it directly implies an unconditional high probability bound by taking the expectation.

Following the proof of Theorem C.7 thus yields the high probability bound in the statement as soon as $T_2 \geq 18 L_\phi \log(16d/\delta)$ and $\tau \leq L_\phi \log(16d/\delta)\lambda_{\min}(\Lambda_\phi)$. All that remains to be checked is that these inequalities are true as soon as $T_2 \geq \frac{18}{\omega_{\min}} \log(16d/\delta)$ and $\tau \leq \frac{r}{(m \wedge n)} \frac{\omega_{\min}}{\omega_{\max}} \log(16d/\delta)$. By Lemma D.11 and the fact that $\max_{i,j} \|\phi_{i,j}\|_2^2 \geq \max(\|U\|_{2\to\infty}^2, \|V\|_{2\to\infty}^2)$,

$$\frac{r}{(m \wedge n)\,\omega_{\max}} \leq \max_{i,j} \frac{\|\phi_{i,j}\|_2^2}{\lambda_{\max}(\Lambda_\phi)} \leq L_\phi \leq \max_{i,j} \frac{\|\phi_{i,j}\|_2^2}{\lambda_{\min}(\Lambda_\phi)} \leq \frac{1}{\omega_{\min}},$$

which entails the desired result. $\qquad\square$

Note that in addition to $L_\phi \leq \frac{1}{\omega_{\min}}$, we also have $L_\theta \leq \max_{i,j} \|\phi_{i,j}\|_2 \|\theta\|_2 \leq \sqrt{mn}\|M\|_{\max}$. This shows that the terms that depend on E_2, E_3 and E_4 are of lower order compared to the term that depends on E_1, and that they have a polynomial dependency in the model parameters (note that $\|\phi_\pi\|_{\Lambda_\phi^{-1}} \leq \sqrt{L_\phi}$). We will now see that we can obtain a sharper control of the dependency in the model parameters by leveraging high probability bounds on L_ϕ and L_θ instead of deterministic ones. These high probability bounds come freely with the control of the misspecification error.

Lemma D.5. *Let $\delta > 0$. There exists an universal constant $c > 0$ such that $T_1 \geq \frac{cL^2\kappa^2(m+n)}{\sigma_r^2\omega_{\min}} \log^3\left(\frac{(m+n)T_1}{\delta}\right)$ ensures that with probability larger than $1 - \delta$, the following inequalities happen at the same time :*

- $\epsilon_{\max} \lesssim \frac{L^2}{\sigma_r} \frac{\mu^2\kappa^3 r(m+n)}{T_1\omega_{\min}(m\wedge n)} \log^3\left(\frac{(m+n)T_1}{\delta}\right)$

- $L_\theta \lesssim \frac{\sigma_1\mu^2 r}{m\wedge n}$

- $L_\phi \lesssim \frac{\mu^2 r(m+n)}{\omega_{\min}mn}$.

Proof of Lemma D.5. Since $T_1 \gtrsim \frac{L^2(m+n)}{\sigma_r^2\omega_{\min}} \log^3\left(\frac{(m+n)T_1}{\delta}\right)$, by Theorem 4.1, there is a universal constant C such that the event

$$A_\delta = \left\{\epsilon_{\text{Sub-Rec}} \leq \frac{CL}{\sigma_r} \sqrt{\frac{\mu^2\kappa^2 r(m+n)}{T_1\omega_{\min}(m\wedge n)} \log^3\left(\frac{(m+n)T_1}{\delta}\right)}\right\}$$

holds with probability at least $1 - \delta$. Under A_δ, from (8),

$$\epsilon_{\max} \leq \sigma_1\epsilon_{\text{Sub-Rec}}^2 \lesssim \frac{L^2}{\sigma_r} \frac{\mu^2\kappa^3 r(m+n)}{T_1\omega_{\min}(m\wedge n)} \log^3\left(\frac{(m+n)T_1}{\delta}\right).$$

Let us now prove that both of the remaining inequalities are true under A_δ. Specifically, we will see that L_θ (resp. L_ϕ) can be deterministically bounded by a function of $\epsilon_{\max}$ (resp. $\epsilon_{\text{Sub-Rec}}$).

Control of L_θ. Recall that $M_{i,j} = \phi_{i,j}^\top\theta + \epsilon_{i,j}$. Consequently, $L_\theta = \max_{i,j}\left|\phi_{i,j}^\top\theta\right| \leq \|M\|_{\max} + \epsilon_{\max}$. Since $T_1 \gtrsim \frac{L^2\kappa^2(m+n)}{\sigma_r^2\omega_{\min}} \log^3\left(\frac{(m+n)T_1}{\delta}\right)$, under A_δ we have $\epsilon_{\max} \lesssim \sigma_1\frac{\mu^2 r}{m\wedge n}$, so that by Lemma B.8,

$$L_\theta \leq \|M\|_{\max} + \epsilon_{\max}$$
$$\lesssim \sigma_1\frac{\mu^2 r}{\sqrt{mn}} + \sigma_1\frac{\mu^2 r}{m\wedge n}$$
$$\lesssim \frac{\sigma_1\mu^2 r}{m\wedge n}.$$

Control of L_ϕ. Recall that $L_\phi \leq \max_{i,j} \dfrac{\|\phi_{i,j}\|_2^2}{\lambda_{\min}(\Lambda_\phi)} \leq \max_{i,j} \dfrac{\|\phi_{i,j}\|_2^2}{\omega_{\min}}$. Note that by Lemma D.12,
$\max_{i,j}\|\phi_{i,j}\|_2^2 \leq \mu^2 r \dfrac{m+n-\mu^2 r}{mn} + 3\epsilon_{\text{Sub-Rec}}$. Consequently, $L_\phi \leq \dfrac{1}{\omega_{\min}}\left(\mu^2 r \dfrac{m+n-\mu^2 r}{mn} + 3\epsilon_{\text{Sub-Rec}}\right)$. Since $T_1 \gtrsim$
$\dfrac{L^2\kappa^2(m+n)}{\sigma_r^2\omega_{\min}}\log^3\left(\dfrac{(m+n)T_1}{\delta}\right)$, under A_δ, we have $\epsilon_{\text{Sub-Rec}} \lesssim \dfrac{\mu\sqrt{r}}{\sqrt{m\wedge n}}$, so

$$L_\phi \lesssim \frac{1}{\omega_{\min}}\left(\mu^2 r\frac{m+n-\mu^2 r}{mn} + \frac{\mu\sqrt{r}}{\sqrt{m\wedge n}}\right) \lesssim \frac{\mu^2 r(m+n)}{\omega_{\min}mn}.$$

$\square$

Choice of T_1. Note that the high probability bounds of Lemma D.3 and D.5 together suggest that as long as $T_1 = O\left(T^\beta\right)$ for $1/2 < \beta \leq 1$, the OPE error that originates from the misspecification will be of higher order than the error induced by the least squares estimation of θ with high probability. Furthermore, to ensure that the least squares error term and the misspecification error term both have the best achievable scaling in T, we should have $T_1 = \Theta(T)$ and $T_2 = \Theta(T)$. Specifically, for some $\alpha \in (0,1)$, we can choose $T_1 = \lfloor \alpha T \rfloor$ and $T_2 = \lceil (1-\alpha)T \rceil$.

We can now finish the proof of Proposition D.2 by combining the error bounds of Lemma D.3 and D.5 with a union bound.

Note that the high probability bounds of Lemma D.12 and D.5 hold as soon as

$$\alpha T \geq 1 + \frac{cL^2\kappa^2(m+n)}{\sigma_r^2\omega_{\min}}\log^3\left(\frac{(m+n)\alpha T}{\delta}\right),$$

$$(1-\alpha)T \geq \frac{18}{\omega_{\min}}\log(16d/\delta)$$

$$\text{and } \tau \leq \frac{r}{m\wedge n}\frac{\omega_{\min}}{\omega_{\max}}\log(16d/\delta)$$

for some universal constant $c > 0$. Recall that we denoted by A_δ the event under which the bounds of Lemma D.5 holds. By Lemma D.3, under A_δ, with probability larger than $1 - \delta/2$, we have

$$|L_1| + |L_2| \leq \|\phi_\pi\|_{\Lambda_\phi^{-1}}\sqrt{\frac{2\sigma^2\log(16/\delta)}{(1-\alpha)T}} + K_0\frac{\log^2(e/\delta)}{T},$$

$$|L_3| \leq K_0\frac{\log^3((m+n)T/\delta)}{T}$$

for $K_0 = \text{poly}\left(\|M\|_{\max}, \sigma, \mu, \kappa, r, m, n, (\omega_{\min}mn)^{-1}\right)$ (we have used that $\frac{1}{\sqrt{T_2}} \leq \frac{1}{\sqrt{(1-\alpha)T}}$). A union bound concludes the proof of the first part of Proposition D.2.

Let us now upper bound the polynomial term K_0. Direct computations ensure that when the high probability bounds of Lemma D.3 and D.5 hold,

$$|L_3| \lesssim \epsilon_{\max}L_\phi \lesssim \frac{L^2}{\sigma_r}\frac{\mu^4\kappa^3 r^2(m+n)^2\log^3((m+n)T/\delta)}{T\omega_{\min}^2 mn(m\wedge n)}$$

$$\lesssim \frac{L^2}{\|M\|_{\max}}\frac{\mu^6\kappa^4 r^3(m+n)^2\sqrt{mn}\log^3((m+n)T/\delta)}{T(\omega_{\min}mn)^2(m\wedge n)},$$

$$\frac{E_2}{T} \lesssim \frac{\sqrt{L_\phi}\left(L_\phi L_\theta + \sigma\sqrt{dL_\phi}\right)\log(d/\delta)}{T} \lesssim \frac{(\|M\|_{\max}+\sigma)\mu^5 r^{5/2}(m+n)^{3/2}\sqrt{mn}\log(d/\delta)}{T(\omega_{\min}mn)^{3/2}(m\wedge n)},$$

$$\frac{E_3}{T^{3/2}} \lesssim \frac{L_\phi\left(\sigma\sqrt{d} + \sqrt{L_\phi}L_\theta\right)\log(d/\delta)}{T} \lesssim \frac{(\|M\|_{\max}+\sigma)\mu^5 r^{5/2}(m+n)^{3/2}\sqrt{mn}\log(d/\delta)}{T(\omega_{\min}mn)^{3/2}(m\wedge n)},$$

$$\text{and } \frac{E_4}{T^2} \lesssim \frac{\sigma \sqrt{d} L_\phi \log (d/\delta)}{T} \lesssim \frac{\sigma \mu^2 r^{3/2} (m+n)^{3/2} \log (d/\delta)}{T \omega_{\min} mn}$$

where we have used Lemma B.8 to bound σ_1 and σ_r, and $T \gtrsim L_\phi \log(d/\delta)$ (see the end of the proof of Lemma D.3) in the last inequalities. Since $\frac{4L^2}{\|M\|_{\max}} \geq \frac{(\|M\|_{\max}+\sigma)^2}{\|M\|_{\max}} \geq \|M\|_{\max} + \sigma \geq \sigma$, the first error term dominates all the others, and we can take

$$K_0 := C' \frac{L^2}{\|M\|_{\max}} \frac{\mu^6 \kappa^4 r^3 (m+n)^2 \sqrt{mn}}{(\omega_{\min} mn)^2 (m \wedge n)}$$

for some universal constant $C' > 0$. $\qquad \square$

Remark D.6. To derive an instance-dependent upper bound from Proposition D.2, one may be tempted to prove that with high probability,

$$\|\phi_\pi\|_{\Lambda_\phi^{-1}} = \|\psi_\pi\|_{\Lambda^{-1}} + o(1).$$

This would yield an error bound whose dominant term scales with $\|\psi_\pi\|_\Lambda^{-1} \sqrt{\frac{\log(1/\delta)}{T}}$ and would echo the linear bandit upper bound of Theorem C.7, that we have demonstrated is instance-optimal. Although numerical experiments suggest this result holds, we were unsuccessful in our proof attempts. Notably, one could write $\|\phi_\pi\|_{\Lambda_\phi^{-1}}^2 = \|\psi_\pi\|_{\Lambda^{-1}}^2 + (\|\phi_\pi\|_{\Lambda_\phi^{-1}}^2 - \|\phi_\pi\|_{\Lambda^{-1}}^2) + (\|\phi_\pi\|_{\Lambda^{-1}}^2 - \|\psi_\pi\|_{\Lambda^{-1}}^2)$ and attempt to prove that the two rightmost terms converge to 0 with high probability. However, numerical experiments suggest that these terms do not both vanish in general, but rather cancel out, and thus a more careful analysis is required.

By leveraging Lemma D.12, we can however obtain the following instance-dependent bound, with a dependency on $\omega_{\min}$ instead of the entire behavior policy. Theorem 6.2 will follow from this result.

Theorem D.7. *Under the assumptions of Proposition D.2, with probability larger than $1 - \delta$,*

$$|v^\pi - \hat{v}_{\text{RS-PE}}| \leq \frac{\|\psi_\pi\|_2}{\sqrt{\omega_{\min}}} \sqrt{\frac{2\sigma^2 \log(16/\delta)}{(1-\alpha) T}} + K_0 \left(\frac{\log^3 ((m+n) T/\delta)}{T} + \frac{\log^{5/4} ((m+n) T/\delta)}{T^{3/4}} \right),$$

where

$$K_0 \lesssim \frac{L^2}{\|M\|_{\max}} \frac{\mu^6 \kappa^4 r^3 (m+n)^2 \sqrt{mn}}{(\omega_{\min} mn)^2 (m \wedge n)}.$$

Proof of Theorem D.7. Note that by Lemmas D.11 and D.12,

$$\|\phi_\pi\|_{\Lambda_\phi^{-1}}^2 \leq \frac{\|\phi_\pi\|_2^2}{\omega_{\min}} \leq \frac{1}{\omega_{\min}} \left(\|\psi_\pi\|_2^2 + 3\epsilon_{\text{Sub-Rec}} \right).$$

Under the event A_δ defined in the proof of Lemma D.5, we consequently also have (using $\frac{1}{\sigma_r} \leq \frac{\mu^2 \kappa r}{\sqrt{mn}\|M\|_{\max}}$)

$$\|\psi_\pi\|_{\Lambda_\phi^{-1}}^2 \leq \frac{1}{\omega_{\min}} \left(\|\psi_\pi\|_2^2 + \frac{C}{\sqrt{T}} \frac{L}{\|M\|_{\max}} \mu^3 \kappa^2 r^{3/2} \sqrt{\frac{(m+n)}{\omega_{\min} mn (m \wedge n)}} \log^{3/2} \left(\frac{(m+n) T}{\delta} \right) \right) \qquad (31)$$

for some universal constant $C > 0$. Since $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ for every $x, y \geq 0$, this implies

$$\|\phi_\pi\|_{\Lambda_\phi^{-1}} \leq \frac{\|\psi_\pi\|_2}{\sqrt{\omega_{\min}}} + C^{1/2} \sqrt{\frac{L}{\|M\|_{\max}}} \mu^{3/2} \kappa r^{3/4} \omega_{\min}^{-3/4} \left(\frac{m+n}{mn (m \wedge n)} \right)^{1/4} \frac{\log^{3/4} ((m+n) T/\delta)}{T^{1/4}}. \qquad (32)$$

Combining this result with Proposition D.2[15] and noting that $\sigma \sqrt{\frac{L}{\|M\|_{\max}}} \leq \frac{\sigma L}{\|M\|_{\max}} \leq \frac{L^2}{\|M\|_{\max}} := A$, we get that with

---

[15]Note that there is no need for a union bound as the high probability bound (32) is true under A_δ.

probability larger than $1 - \delta$,

$$|v^\pi - \hat{v}_{\text{RS-PE}}| \leq \frac{\|\psi_\pi\|_2}{\sqrt{\omega_{\min}}} \sqrt{\frac{2\sigma^2 \log(16/\delta)}{(1-\alpha)T}} + C_1 A \sqrt{\log(e/\delta)} \mu^{3/2} \kappa r^{3/4} \omega_{\min}^{-3/4} \left(\frac{m+n}{mn(m \wedge n)}\right)^{1/4} \frac{\log^{3/4}((m+n)T/\delta)}{T^{3/4}}$$

$$+ C_2 A \mu^6 \kappa^4 r^3 \frac{(m+n)^2 \sqrt{mn}}{(\omega_{\min}mn)^2 (m \wedge n)} \frac{\log^3((m+n)T/\delta)}{T}$$

for some universal constants $C_1, C_2 > 0$. Since $\sqrt{\log(e/\delta)}\mu^{3/2}\kappa r^{3/4}\omega_{\min}^{-3/4}\left(\frac{m+n}{mn(m \wedge n)}\right)^{1/4}\log^{3/4}((m+n)T/\delta) \lesssim \mu^6\kappa^4 r^3 \frac{(m+n)^2\sqrt{mn}}{(\omega_{\min}mn)^2(m \wedge n)}\log^{5/4}((m+n)T/\delta)$, this concludes the proof.

$\square$

Remark D.8. When M and ω are homogeneous, $K_0 \lesssim \frac{(\|M\|_{\max} \vee \sigma)^2}{\|M\|_{\max}}mn$, and if we additionally have $\|M\|_{\max} = \Theta(\sigma)$, $K_0 \lesssim \sigma mn$.

When both the context distribution and the behavior policy are uniform, the upper bound of Theorem D.7 corresponds to the conjectured instance-dependent upper bound mentioned in Remark D.6. Indeed, Lemma D.11 ensures that in this case, $\Lambda = \frac{1}{mn}I_d$, and thus $\|\psi_\pi\|_{\Lambda^{-1}} = \sqrt{mn}\|\psi_\pi\|_2 = \frac{\|\psi_\pi\|_2}{\sqrt{\omega_{\min}}}$.

We now explain how to reformulate the statement of Theorem D.7 as a sample complexity result and retrieve the statement of Theorem 6.2.

Proof of Theorem 6.2. Let us assume that we are under the assumptions of Proposition D.2. By Theorem D.7,

$$|v^\pi - \hat{v}_{\text{RS-PE}}| \leq \frac{\|\psi_\pi\|_2}{\sqrt{\omega_{\min}}} \sqrt{\frac{2\sigma^2 \log(16/\delta)}{(1-\alpha)T}} + K_0 \frac{\log^3((m+n)T/\delta)}{T} + K_0 \frac{\log^{5/4}((m+n)T/\delta)}{T^{3/4}}$$

with probability larger than $1 - \delta$. In particular, $|v^\pi - \hat{v}_{\text{RS-PE}}| \leq \varepsilon$ with probability larger than $1 - \delta$ as soon as

$$\frac{\varepsilon}{3} \geq \frac{\|\psi_\pi\|_2}{\sqrt{\omega_{\min}}} \sqrt{\frac{2\sigma^2 \log(16/\delta)}{(1-\alpha)T}}, \quad \frac{\varepsilon}{3} \geq K_0 \frac{\log^3((m+n)T/\delta)}{T}, \quad \text{and} \quad \frac{\varepsilon}{3} \geq K_0 \frac{\log^{5/4}((m+n)T/\delta)}{T^{3/4}}.$$

It is straightforward to check that

$$T \gtrsim \frac{\sigma^2 \|\psi_\pi\|_2^2}{\omega_{\min}\, \varepsilon^2} \log\left(\frac{e}{\delta}\right) + \frac{K_0^{4/3}}{\varepsilon^{4/3}} \log^3\left(\frac{(m+n)T}{\delta}\right) \tag{33}$$

is enough to satisfy all three inequalities, where we recall that K_0 has been defined in (28).

Furthermore, the conditions on the number of samples of Theorem D.7, namely $T \gtrsim \frac{L^2\kappa^2(m+n)}{\sigma_r^2\omega_{\min}}\log^3\left(\frac{(m+n)T}{\delta}\right)$ and $T \gtrsim \frac{1}{\omega_{\min}}\log(16d/\delta)$, are both implied by (33) when $\varepsilon \leq \|M\|_{\max}$ because this choice implies $T \gtrsim \frac{K_0}{\|M\|_{\max}}\log^3\left(\frac{(m+n)T}{\delta}\right)$ (to see that the second condition is verified, note that $\frac{K_0}{\|M\|_{\max}} \gtrsim \frac{(m+n)^2}{\omega_{\min}\sqrt{mn}(m \wedge n)} \gtrsim \frac{1}{\omega_{\min}}$).

The statement of the theorem thus holds with the multiplicative factor

$$K_1 := K_0^{4/3} \lesssim \left(\frac{L^2}{\|M\|_{\max}} \frac{\mu^6\kappa^4 r^3(m+n)^2\sqrt{mn}}{(\omega_{\min}mn)^2(m \wedge n)}\right)^{4/3}. \tag{34}$$

We finally note that by Remark D.8, when M and ω are homogeneous and $\|M\|_{\max} = \Theta(\sigma)$, $K_1 \lesssim (\sigma mn)^{4/3}$. $\square$

Minimax sample complexity guarantee. Let us note that by (40), $\|\psi_\pi\|_2^2 \leq \mu^2 r \frac{m+n-\mu^2 r}{mn}$, which implies that $|v^\pi - \hat{v}_{\text{RS-PE}}^\pi| \leq \varepsilon$ with probability larger than $1 - \delta$ as soon as

$$T \gtrsim \frac{\sigma^2\mu^2 r(m+n-\mu^2 r)}{\omega_{\min}mn\, \varepsilon^2} \log\left(\frac{e}{\delta}\right) + \frac{K_1}{\varepsilon^{4/3}} \log^3\left(\frac{(m+n)T}{\delta}\right).$$

By adapting the previous analysis with a more refined argument, the dependency in ε and the model parameters of the second term can be improved. This is what we demonstrate now.

Theorem D.9. *Let $\varepsilon, \delta \in (0,1)$. With the choices $T_1 = \lfloor \alpha T \rfloor$ and $\tau \leq r(m \wedge n)^{-1}(\omega_{\min}/\omega_{\max}) \log(16d/\delta)$, $|v^{\pi} - \hat{v}^{\pi}_{\mathrm{RS-PE}}| \leq \varepsilon$ with probability larger than $1 - \delta$ as soon as*

$$T \gtrsim \frac{\sigma^2 \mu^2 r (m+n)}{\omega_{\min} mn \, \varepsilon^2} \log \left(\frac{e}{\delta} \right) + \frac{K_0}{\varepsilon} \log^3 \left(\frac{(m+n)T}{\delta} \right)$$

where

$$K_0 \lesssim \frac{L^2}{\|M\|_{\max}} \frac{\mu^6 \kappa^4 r^3 (m+n)^2 \sqrt{mn}}{(\omega_{\min} mn)^2 (m \wedge n)}. \tag{35}$$

Remark D.10. When M and ω are homogeneous and $\|M\|_{\max} = \Theta(\sigma)$, $K_0 \lesssim \sigma mn$.

Proof of Theorem D.9. Combining (31) with the bound on $\|\psi_\pi\|_2^2$ entails

$$\begin{aligned}
\|\psi_\pi\|^2_{\Lambda_\phi^{-1}} &\leq \frac{1}{\omega_{\min}} \left(\frac{\mu^2 r (m+n)}{mn} + \frac{C}{\sqrt{T}} \frac{L}{\|M\|_{\max}} \mu^3 \kappa^2 r^{3/2} \sqrt{\frac{m+n}{mn (m \wedge n) \omega_{\min}}} \log^{3/2} \left(\frac{(m+n)T}{\delta} \right) \right) \\
&= \frac{\mu^2 r (m+n)}{\omega_{\min} mn} \left(1 + \frac{C}{\sqrt{T}} \frac{L}{\|M\|_{\max}} \mu \kappa^2 r^{1/2} \sqrt{\frac{mn}{(m+n)(m \wedge n) \omega_{\min}}} \log^{3/2} \left(\frac{(m+n)T}{\delta} \right) \right)
\end{aligned} \tag{36}$$

for some universal constant $C > 0$. Now, using the inequality $\sqrt{1+x} \leq 1 + \frac{x}{2}$ (instead of $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ which was used previously), (36) implies

$$\|\psi_\pi\|_{\Lambda_\phi^{-1}} \leq \mu \sqrt{\frac{r(m+n)}{\omega_{\min} mn}} \left(1 + \frac{C}{2\sqrt{T}} \frac{L}{\|M\|_{\max}} \mu \kappa^2 r^{1/2} \sqrt{\frac{mn}{(m+n)(m \wedge n) \omega_{\min}}} \log^{3/2} \left(\frac{(m+n)T}{\delta} \right) \right)$$

Combining this result with Proposition D.2 and noting that $\frac{\sigma L}{\|M\|_{\max}} \leq \frac{L^2}{\|M\|_{\max}} = A$, we get that with probability larger than $1 - \delta$,

$$\begin{aligned}
|v^{\pi} - \hat{v}^{\pi}_{\mathrm{RS-PE}}| &\leq \mu \sqrt{\frac{r(m+n)}{\omega_{\min} mn}} \sqrt{\frac{2\sigma^2 \log(16/\delta)}{(1-\alpha)T}} + C_1 A \sqrt{\log(e/\delta)} \mu^2 \kappa^2 r \omega_{\min}^{-1} (m \wedge n)^{-1/2} \frac{\log^{3/2}((m+n)T/\delta)}{T} \\
&\quad + C_2 A \mu^6 \kappa^4 r^3 \frac{(m+n)^2 \sqrt{mn}}{(\omega_{\min} mn)^2 (m \wedge n)} \frac{\log^3((m+n)T/\delta)}{T}
\end{aligned}$$

for some universal constants $C_1, C_2 > 0$. Since $\sqrt{\log(e/\delta)} \mu^2 \kappa^2 r \omega_{\min}^{-1} (m \wedge n)^{-1/2} \log^{3/2}((m+n)T/\delta) \lesssim \mu^6 \kappa^4 r^3 \frac{(m+n)^2 \sqrt{mn}}{(\omega_{\min} mn)^2 (m \wedge n)} \log^3((m+n)T/\delta)$, we obtain that under the assumptions of Proposition D.2,

$$|v^{\pi} - \hat{v}^{\pi}_{\mathrm{RS-PE}}| \leq \mu \sqrt{\frac{r(m+n)}{\omega_{\min} mn}} \sqrt{\frac{2\sigma^2 \log(16/\delta)}{(1-\alpha)T}} + K_0 \frac{\log^3((m+n)T/\delta)}{T} \tag{37}$$

with probability larger than $1 - \delta$, and transforming this result into a sample complexity guarantee as in the proof of Theorem 6.2 yields that $|v^{\pi} - \hat{v}^{\pi}_{\mathrm{RS-PE}}| \leq \varepsilon$ with probability larger than $1 - \delta$ as soon as

$$T \gtrsim \frac{\sigma^2 \mu^2 r (m+n)}{\omega_{\min} mn \, \varepsilon^2} \log \left(\frac{e}{\delta} \right) + \frac{K_0}{\varepsilon} \log^3 \left(\frac{(m+n)T}{\delta} \right).$$

$\square$

The above sample complexity guarantee can be reformulated as

$$
T \gtrsim \frac{\sigma^2 \mu^2 r \, (m+n)}{\omega_{\min} mn\varepsilon^2} \log\left(\frac{e}{\delta}\right)
$$

when $\frac{\sigma^2 \mu^2 r(m+n)}{\omega_{\min} mn\,\varepsilon^2} \log\left(\frac{e}{\delta}\right) \gtrsim \frac{K_0}{\varepsilon} \log^3\left(\frac{(m+n)T}{\delta}\right)$, i.e.,

$$
\varepsilon \lesssim \frac{\sigma^2 \|M\|_{\max}}{L^2} \frac{\omega_{\min} mn\,(m \wedge n)}{\mu^4 \kappa^2 r^2\,(m+n)\,\sqrt{mn}} \frac{\log(e/\delta)}{\log^3\left((m+n)\,T/\delta\right)}. \tag{38}
$$

When M and ω are homogeneous and $\sigma \gtrsim \|M\|_{\max}$, this simplifies to

$$
T \gtrsim \frac{\sigma^2\,(m+n)}{\varepsilon^2} \log\left(\frac{e}{\delta}\right)
$$

when

$$
\varepsilon \lesssim \frac{\|M\|_{\max}}{m+n} \frac{\log(e/\delta)}{\log^3\left((m+n)\,T/\delta\right)}. \tag{39}
$$

D.3. Additional Lemmas

The following lemma entails that Λ and Λ_ϕ are always invertible under our assumption $\omega_{\min} > 0$.

Lemma D.11. $\sum_{i,j} \psi_{i,j}\psi_{i,j}^\top = \sum_{i,j} \phi_{i,j}\phi_{i,j}^\top = I_d$. *Consequently,*

$$
\omega_{\max} \geq \lambda_{\max}(\Lambda) \geq \lambda_{\min}(\Lambda) \geq \omega_{\min}
$$

and

$$
\omega_{\max} \geq \lambda_{\max}(\Lambda_\phi) \geq \lambda_{\min}(\Lambda_\phi) \geq \omega_{\min}.
$$

Proof of Lemma D.11. A direct computation shows that the diagonal entries of $\sum_{i,j} \psi_{i,j}\psi_{i,j}^\top$ can all be written as $\sum_{i,j} U_{i,k}^2 V_{j,l}^2$, $\sum_{i,j} U_{\perp,i,k}^2 V_{j,l}^2$ or $\sum_{i,j} U_{i,k}^2 V_{\perp,j,l}^2$ for some k and l. All of these terms are equal to 1 by the orthogonality of $\begin{bmatrix} U & U_\perp \end{bmatrix}$ and $\begin{bmatrix} V & V_\perp \end{bmatrix}$. The entries outside of the diagonal can all be factorized by a scalar product between two columns of $\begin{bmatrix} U & U_\perp \end{bmatrix}$ or two columns of $\begin{bmatrix} V & V_\perp \end{bmatrix}$, which is equal to 0 by the same orthogonality argument. The same reasoning holds for $\sum_{i,j} \phi_{i,j}\phi_{i,j}^\top$, as $\begin{bmatrix} \widehat{U} & \widehat{U}_\perp \end{bmatrix}$ and $\begin{bmatrix} \widehat{V} & \widehat{V}_\perp \end{bmatrix}$ are also orthogonal. The other part of the statement directly follows from the identities $\omega_{\max} \sum_{i,j} \psi_{i,j}\psi_{i,j}^\top \succeq \Lambda \succeq \omega_{\min} \sum_{i,j} \psi_{i,j}\psi_{i,j}^\top$ and $\omega_{\max} \sum_{i,j} \phi_{i,j}\phi_{i,j}^\top \succeq \Lambda_\phi \succeq \omega_{\min} \sum_{i,j} \phi_{i,j}\phi_{i,j}^\top$. $\square$

Lemma D.12. *Let ν be a distribution on $[m] \times [n]$ so that $\sum_{i,j} \nu_{i,j} = 1$, and let $\psi_\nu = \sum_{i,j} \nu_{i,j}\psi_{i,j}$, $\phi_\nu = \sum_{i,j} \nu_{i,j}\phi_{i,j}$. Then*

- $\left| \|\phi_\nu\|_2^2 - \|\psi_\nu\|_2^2 \right| \leq 3\epsilon_{\text{Sub-Rec}}$

- $\max_{i,j} \|\psi_{i,j}\|_2^2 \leq \mu^2 r \dfrac{m+n-\mu^2 r}{mn}.$

Proof of Lemma D.12. We have $\|\phi_\nu\|_2^2 = \displaystyle\sum_{(i,j),(i',j')} \nu_{i,j}\nu_{i',j'}\phi_{i,j}^\top\phi_{i',j'}$, so it is enough to show that $\phi_{i,j}^\top\phi_{i',j'}$ approaches $\psi_{i,j}^\top\psi_{i',j'}$ for every $(i,j),(i',j')$. A direct computation yields

$$
\phi_{i,j}^\top\phi_{i',j'} = \left(\widehat{U}_i^\top\widehat{U}_{i'}\right)\left(\widehat{V}_j^\top\widehat{V}_{j'}\right) + \left(\widehat{U}_{\perp,i}^\top\widehat{U}_{\perp,i'}\right)\left(\widehat{V}_j^\top\widehat{V}_{j'}\right) + \left(\widehat{U}_i^\top\widehat{U}_{i'}\right)\left(\widehat{V}_{\perp,j}^\top\widehat{V}_{\perp,j'}\right)
$$

and

$$
\psi_{i,j}^\top\psi_{i',j'} = \left(U_i^\top U_{i'}\right)\left(V_j^\top V_{j'}\right) + \left(U_{\perp,i}^\top U_{\perp,i'}\right)\left(V_j^\top V_{j'}\right) + \left(U_i^\top U_{i'}\right)\left(V_{\perp,j}^\top V_{\perp,j'}\right).
$$

By orthogonality arguments, $\widehat{U}_i^\top \widehat{U}_{i'} + \widehat{U}_{\perp,i}^\top \widehat{U}_{\perp,i'} = U_i^\top U_{i'} + U_{\perp,i}^\top U_{\perp,i'} = 0$ when $i \neq i'$, and $\widehat{V}_j^\top \widehat{V}_{j'} + \widehat{V}_{\perp,j}^\top \widehat{V}_{\perp,j'} = V_j^\top V_{j'} + V_{\perp,j}^\top V_{\perp,j'} = 0$ when $j \neq j'$. Thus, two out of the three terms in the equalities above cancel out when $(i,j) \neq (i',j')$. When $(i,j) = (i',j')$, we can combine two terms by using the fact that $\|U_i\|_2^2 + \|U_{\perp,i}\|_2^2 = 1$ to write $\psi_{i,j}^\top \psi_{i,j} = \|V_j\|_2^2 + \|U_i\|_2^2 \|V_{\perp,j}\|_2^2$. In any case, each term in the difference can be controlled by the subspace recovery error. For example,

$$\left(\widehat{U}_i^\top \widehat{U}_{i'}\right)\left(\widehat{V}_{\perp,j}^\top \widehat{V}_{\perp,j'}\right) - \left(U_i^\top U_i'\right)\left(V_{\perp,j}^\top V_{\perp,j'}\right) = \left(\widehat{V}_{\perp,j}^\top \widehat{V}_{\perp,j'}\right)\left(\widehat{U}_i^\top \widehat{U}_{i'} - U_i^\top U_i'\right) + \left(U_i^\top U_i'\right)\left(\widehat{V}_{\perp,j}^\top \widehat{V}_{\perp,j'} - V_{\perp,j}^\top V_{\perp,j'}\right)$$
$$\leq \|\widehat{V}_{\perp,j}\|_2 \|\widehat{V}_{\perp,j'}\|_2 \|\widehat{U}\widehat{U}^\top - UU^\top\|_{2\to\infty}$$
$$+ \|U_i\|_2 \|U_{i'}\|_2 \|\widehat{V}_\perp \widehat{V}_\perp^\top - V_\perp V_\perp^\top\|_{2\to\infty}$$
$$\leq 2\epsilon_{\text{Sub-Rec}}.$$

Consequently,

$$\phi_{i,j}^\top \phi_{i',j'} - \psi_{i,j}^\top \psi_{i',j'} \leq \left(2 + \mathbb{1}_{(i,j)=(i',j')}\right)\epsilon_{\text{Sub-Rec}}.$$

In the end,

$$\left|\|\phi_\nu\|_2^2 - \|\psi_\nu\|_2^2\right| = \left|\sum_{(i,j),(i',j')} \nu_{i,j}\nu_{i',j'}\left(\phi_{i,j}^\top \phi_{i',j'} - \psi_{i,j}^\top \psi_{i',j'}\right)\right|$$
$$\leq \left(\sum_{(i,j),(i',j')} \nu_{i,j}\nu_{i',j'}\left(2 + \mathbb{1}_{(i,j)=(i',j')}\right)\right)\epsilon_{\text{Sub-Rec}}$$
$$= \left(2 + \sum_{i,j} \nu_{i,j}^2\right)\epsilon_{\text{Sub-Rec}},$$

which is slightly sharper than the desired result.

For the second bullet point, note that for any (context, arm) pair (i,j),

$$\|\psi_{i,j}\|_2^2 = \|U_i\|_2^2\|V_j\|_2^2 + \|U_{\perp,i}\|_2^2\|V_j\|_2^2 + \|U_i\|_2^2\|V_{\perp,j}\|_2^2$$
$$= \|V_j\|_2^2 + \|U_i\|_2^2\left(1 - \|V_j\|_2^2\right)$$
$$\leq \|V_j\|_2^2 + \|U\|_{2\to\infty}^2\left(1 - \|V_j\|_2^2\right)$$
$$= \|U\|_{2\to\infty}^2 + \|V_j\|_2^2\left(1 - \|U\|_{2\to\infty}^2\right)$$
$$\leq \|U\|_{2\to\infty}^2 + \|V\|_{2\to\infty}^2\left(1 - \|U\|_{2\to\infty}^2\right)$$
$$\leq \mu^2\frac{r}{n} + \mu^2\frac{r}{m}\left(1 - \mu^2\frac{r}{n}\right)$$
$$= \mu^2 r\frac{m + n - \mu^2 r}{mn},$$

where the last inequality is obtained by distinguishing the cases $\mu^2 = \frac{m}{r}\|U\|_{2\to\infty}^2$ and $\mu^2 = \frac{n}{r}\|V\|_{2\to\infty}^2$. $\qquad\square$

There are two interesting things to mention about the second inequality of Lemma D.12. First, it is tight in the sense that it is attained for $r = m \wedge n$. Indeed, assume then that $m \wedge n = n$ without loss of generality. On the one hand, $\|\psi_{i,j}\|_2^2 = \|V_j\|_2^2 + \|U_i\|_2^2\left(1 - \|V_j\|_2^2\right) = 1$ since V_j is a row of the orthogonal matrix V. On the other hand, by orthogonality of U, $\mu = \|U\|_{2\to\infty} = 1$ so that $\frac{\mu^2 r(m + n - \mu^2 r)}{mn} = 1$. Second, it directly entails an upper bound on the norm of ψ_π. Indeed,

$$\|\psi_\pi\|_2 = \|\sum_{i,j} w_{i,j}^\pi \psi_{i,j}\|_2 \leq \sum_{i,j} w_{i,j}^\pi \|\psi_{i,j}\|_2 \leq \max_{i,j}\|\psi_{i,j}\|_2,$$

where we have used $\sum_{i,j} w_{i,j}^\pi = 1$ in the last inequality. This ensures that

$$\|\psi_\pi\|_2^2 \leq \mu^2 r\frac{m + n - \mu^2 r}{mn}. \tag{40}$$

E. Proofs of Section 6.3 : Policy Evaluation Lower Bounds

E.1. Proof of Theorem 6.3

In this appendix, we assume that $\xi_t \sim \mathcal{N}(0, \sigma^2)$ and we choose the rank factorization $M = U\Sigma V^\top = PQ^\top$ with $P = U\sqrt{\Sigma}$ and $Q = V\sqrt{\Sigma}$. Furthermore, we denote by N_i the i-th row vector of a matrix N. As noted in (Jun et al., 2019), the bilinear nature of the reward in low-rank bandit problems makes the derivation of sampling lower bounds more challenging than for linear bandits. We overcome this challenge by considering a subset of perturbations of the reward matrix that behave linearly, yielding a relaxed sampling lower bound.

The quantities associated with a perturbation M' of the reward matrix will be denoted by P', Q' and so on. We will use the following straightforward characterization of matrices of rank at most r.

Lemma E.1. $M \in \mathbb{R}^{m \times n}$ is of rank at most r if and only if there exists $P \in \mathbb{R}^{m \times r}$ and $Q \in \mathbb{R}^{n \times r}$ such that $M = PQ^\top$.

Note that $M = PQ^\top$ yields $M_{i,j} = P_i^\top Q_j$ for every (i,j). By using the same arguments as in the proof of the sampling lower bound for linear bandits (see Proposition C.2) and the characterization of Lemma E.1, we obtain the following result.

Lemma E.2. The sample complexity of any (ε, δ)-PAC estimator of v^π must satisfy

$$T \geq \frac{\mathrm{kl}\,(\delta, 1 - \delta)}{I}$$

for

$$I = \inf_{\substack{P' \in \mathbb{R}^{m \times r}, Q' \in \mathbb{R}^{n \times r}, \\ |v^\pi - v'^\pi| \geq 2\varepsilon}} \frac{1}{2\sigma^2} \sum_{i,j} \omega_{i,j} (P_i^\top Q_j - P_i'^\top Q_j')^2.$$

Proof of Lemma E.2. Adapting the proof of Proposition C.2 yields $T \geq \frac{\mathrm{kl}(\delta, 1-\delta)}{J}$ for

$$J := \inf_{\substack{M', \mathrm{rank}(M')=r, \\ |v^\pi - v'^\pi| \geq 2\varepsilon}} \frac{1}{2\sigma^2} \sum_{i,j} \omega_{i,j} (M_{i,j} - M'_{i,j})^2.$$

Since the set of rank r matrices is dense in the set of matrices of rank smaller than r, a continuity argument ensures that

$$J = \inf_{\substack{M', \mathrm{rank}(M') \leq r, \\ |v^\pi - v'^\pi| \geq 2\varepsilon}} \frac{1}{2\sigma^2} \sum_{i,j} \omega_{i,j} (M_{i,j} - M'_{i,j})^2.$$

The characterization of Lemma E.1 concludes the proof. $\qquad\square$

Optimizing in P' and Q' at the same time is delicate, so it is difficult to compute I directly. To alleviate this problem, we provide an upper bound on I by optimizing in P' and Q' separately. We denote by J_1 (resp. J_2) the infimum obtained when adding the constraint $Q' = Q$ (resp. $P' = P$). We naturally have $I \leq \min(J_1, J_2)$, and we will see that J_1 and J_2 can both be computed explicitly. We recall the definitions $\Lambda_Q^i = \sum_{j=1}^n \omega_{i,j} Q_j Q_j^\top, Q_\pi^i = \sum_{j=1}^n \omega_{i,j}^\pi Q_j, \Lambda_P^j = \sum_{i=1}^m \omega_{i,j} P_i P_i^\top, P_\pi^j = \sum_{i=1}^m \omega_{i,j}^\pi P_i$. These quantities will allow us to simplify the objective and the constraints of the optimization problems.

Lemma E.3. For every (i,j), Λ_Q^i and Λ_P^j are invertible. In addition,

$$J_1 = \frac{2\varepsilon^2}{\sigma^2 \sum_i \|Q_\pi^i\|_{(\Lambda_Q^i)^{-1}}^2} \quad \text{and} \quad J_2 = \frac{2\varepsilon^2}{\sigma^2 \sum_j \|P_\pi^j\|_{(\Lambda_P^j)^{-1}}^2}.$$

Proof of Lemma E.3. Since $P = U\sqrt{\Sigma}$ and $Q = V\sqrt{\Sigma}$, the (k,l)-th entry of $\sum_i P_i P_i^\top$ (resp. $\sum_j Q_j Q_j^\top$) is $\sqrt{\sigma_k \sigma_l} \sum_i U_{i,k} U_{i,l}$ (resp. $\sqrt{\sigma_k \sigma_l} \sum_j V_{j,k} V_{j,l}$). By the semi-orthogonality of U and V, these entries are equal to σ_k when $k = l$ and 0 otherwise:

$$\sum_i P_i P_i^\top = \sum_j Q_j Q_j^\top = \Sigma.$$

Consequently, $\omega_{\min}\Sigma \preceq \Lambda_Q^i \preceq \omega_{\max}\Sigma$ and $\omega_{\min}\Sigma \preceq \Lambda_P^j \preceq \omega_{\max}\Sigma$. In particular, since $\omega_{\min} > 0$, Λ_Q^i and Λ_P^j are invertible.

Now, notice that when $Q' = Q$, we have $v^\pi - v'^\pi = \sum_i D_i^\top Q_\pi^i$ and $\sum_{i,j} \omega_{i,j}(P_i^\top Q_j - P_i'^\top Q_j')^2 = \sum_i D_i^\top \Lambda_Q^i D_i$ for $D_i = P_i - P_i'$. By the same arguments as in the proof of Theorem C.3, this observation allows us to write

$$J_1 = \inf_{\substack{D \in \mathbb{R}^{m \times r}, \\ \sum_i D_i^\top Q_\pi^i = 2\varepsilon}} \frac{1}{2\sigma^2} \sum_i \|D_i\|_{\Lambda_Q^i}^2,$$

and similarly,

$$J_2 = \inf_{\substack{D \in \mathbb{R}^{n \times r}, \\ \sum_j D_j^\top P_\pi^j = 2\varepsilon}} \frac{1}{2\sigma^2} \sum_j \|D_j\|_{\Lambda_P^j}^2.$$

An application of Lemma C.8 concludes the proof. $\qquad\square$

Proof of Theorem 6.3. By combining Lemmas E.2 and E.3, we directly get the main statement of Theorem 6.3. Finally, let us show that the lower bound is independent of the choice of rank factorization $M = PQ^\top$. Let $M = P'Q'^\top$ be another rank factorization of M. Then, there exists an invertible matrix $R \in \mathbb{R}^{r \times r}$ such that $Q' = QR$ and $P' = PR^{-\top}$ (Piziak & Odell, 1999). This implies that for every (i,j), $Q_j' = (Q_j^\top R)^\top = R^\top Q_j$, and $P_i' = (P_i^\top R^{-\top})^\top = R^{-1}P_i$, which in turn ensures that $\Lambda_{Q'}^i = R^\top \Lambda_Q^i R$ and $\Lambda_{P'}^j = R^\top \Lambda_P^j R$ are invertible. Furthermore,

$$\|Q_\pi'^i\|_{(\Lambda_{Q'}^i)^{-1}}^2 = (R^\top Q_\pi^i)^\top (R^\top \Lambda_Q^i R)^{-1} (R^\top Q_\pi^i)$$
$$= (Q_\pi^i)^\top RR^{-1} (\Lambda_Q^i)^{-1} R^{-\top} R^\top Q_\pi^i$$
$$= \|Q_\pi^i\|_{(\Lambda_Q^i)^{-1}}^2,$$

and similarly,

$$\|P_\pi'^j\|_{(\Lambda_{P'}^j)^{-1}}^2 = \|P_\pi^j\|_{(\Lambda_P^j)^{-1}}^2.$$

$\qquad\square$

E.2. Proof of Proposition 6.4

We first provide a relaxed sampling lower bound as a corollary of Theorem 6.3.

Corollary E.4. *The policy and model-dependent quantity $L_{M,\pi}$ in the sampling lower bound of Theorem 6.3 can be controlled as*

$$\frac{1}{\sigma_r \omega_{\min}} \max\left(\sum_i \|Q_\pi^i\|_2^2, \sum_j \|P_\pi^j\|_2^2 \right) \geq \max\left(\sum_i \|Q_\pi^i\|_{(\Lambda_Q^i)^{-1}}^2, \sum_j \|P_\pi^j\|_{(\Lambda_P^j)^{-1}}^2 \right)$$

$$\geq \frac{1}{\sigma_1 \omega_{\max}} \max\left(\sum_i \|Q_\pi^i\|_2^2, \sum_j \|P_\pi^j\|_2^2 \right).$$

In particular, the sample complexity of any (ε, δ)-PAC estimator of v^π must satisfy

$$T \geq \frac{\sigma^2 \, \mathrm{kl}\,(\delta, 1-\delta)}{2\sigma_1 \omega_{\max} \varepsilon^2} \max\left(\sum_i \|Q_\pi^i\|_2^2, \sum_j \|P_\pi^j\|_2^2 \right).$$

Proof of Corollary E.4. By the proof of Lemma E.3, $\omega_{\min}\Sigma \preceq \Lambda_Q^i \preceq \omega_{\max}\Sigma$ and $\omega_{\min}\Sigma \preceq \Lambda_P^j \preceq \omega_{\max}\Sigma$. In particular, $\sigma_r \omega_{\min} \leq \lambda_{\min}(\Lambda_Q^i) \leq \lambda_{\max}(\Lambda_Q^i) \leq \sigma_1 \omega_{\max}$, and the same results hold for Λ_P^j. Now, note that for any $x \in \mathbb{R}^r$ and any

invertible matrix $B \in \mathbb{R}^{r \times r}$, $\lambda_{\min}(B)^{-1} \|x\|_2^2 \geq \|x\|_{B^{-1}}^2 \geq \lambda_{\max}(B)^{-1} \|x\|_2^2$. Consequently, for any (i, j),

$$\frac{\|Q_\pi^i\|_2^2}{\sigma_r \omega_{\min}} \geq \|Q_\pi^i\|_{\left(\Lambda_Q^i\right)^{-1}}^2 \geq \frac{\|Q_\pi^i\|_2^2}{\sigma_1 \omega_{\max}} \text{ and } \frac{\|P_\pi^j\|_2^2}{\sigma_r \omega_{\min}} \geq \|P_\pi^j\|_{\left(\Lambda_P^j\right)^{-1}}^2 \geq \frac{\|P_\pi^j\|_2^2}{\sigma_1 \omega_{\max}},$$

and the first result follows. The sampling lower bound can then be directly deduced from Theorem 6.3. $\qquad\square$

The control of the lower bound identified above is tight when ω is homogeneous and $\kappa = \Theta(1)$, since in this case, the two sides of the inequality only differ by a constant factor. In particular, the relaxed lower bound then scales as the lower bound of Theorem 6.3.

Remark E.5. We could also upper bound the maximal eigenvalues of the covariance matrices by noticing that their positive semi-definiteness entails $\lambda_{\max}\left(\Lambda_Q^i\right) \leq \text{Tr}\left(\Lambda_Q^i\right) = \sum_j \omega_{i,j} \|Q_j\|^2$ and $\lambda_{\max}\left(\Lambda_P^j\right) \leq \text{Tr}\left(\Lambda_P^j\right) = \sum_i \omega_{i,j} \|P_i\|^2$. When $r = 1$, this would yield a bound that coincides with the bound of Theorem 6.3. However, when $r > 1$, these eigenvalue upper bounds would scale with $\sum_{i=1}^r \sigma_i$ instead of σ_1.

We are now ready to derive the worst-case lower bound of Proposition 6.4. To that end, we will show that

$$K_{M,\pi} := \frac{1}{\sigma_1 \omega_{\max}} \max\left(\sum_i \|Q_\pi^i\|_2^2, \sum_j \|P_\pi^j\|_2^2\right)$$

is of order $\frac{m+n}{\omega_{\max} mn}$ in the worst case (i.e., for a choice of homogeneous reward matrix and target policy that maximizes the lower bound). We use the notation $\pi_j := \sum_{i=1}^m \rho_i \pi(j|i)$. We note that the third bullet point of the following proposition corresponds to the statement of Proposition 6.4.

Proposition E.6. *Denote by $\mathcal{H}$ the set of homogeneous matrices of $\mathcal{M}(c, c')$ for some fixed c, c'.*

- *For any $M \in \mathbb{R}^{m \times n}$ of rank r, $K_{M,\pi} \leq \frac{\mu^2 r}{(m \wedge n) \omega_{\max}} \max\left(\sum_i \rho_i^2, \sum_j \pi_j^2\right)$*

- *When $m = \Theta(n)$ and $r = \Theta(1)$, $\displaystyle\sup_{M \in \mathcal{H}} K_{M,\pi} = \Theta\left(\frac{\max\left(\sum_i \rho_i^2, \sum_j \pi_j^2\right)}{(m \vee n) \omega_{\max}}\right)$*

- *When $m = \Theta(n)$ and $r = \Theta(1)$, there is a target policy π and a homogeneous reward matrix $M \in \mathcal{H}$ such that any (ε, δ)-PAC estimator of v^π must satisfy*

$$T \gtrsim \frac{\sigma^2 (m + n)}{\omega_{\max} mn\varepsilon^2} \text{ kl}(\delta, 1 - \delta).$$

Remark E.7. The upper bound of Theorem 6.2 matches the scaling in μ and r suggested by the first bullet point of Proposition E.6.

Proof of Proposition E.6. To prove the first bullet point, notice that

$$\begin{aligned}
K_{M,\pi} &= \frac{1}{\sigma_1 \omega_{\max}} \max\left(\sum_i \|\sum_j \omega_{i,j}^\pi Q_j\|_2^2, \sum_j \|\sum_i \omega_{i,j}^\pi P_i\|_2^2\right) \\
&\leq \frac{1}{\sigma_1 \omega_{\max}} \max\left(\sum_i \left(\sum_j \omega_{i,j}^\pi \|Q_j\|_2\right)^2, \sum_j \left(\sum_i \omega_{i,j}^\pi \|P_i\|_2\right)^2\right) \\
&\leq \frac{1}{\sigma_1 \omega_{\max}} \max\left(\max_j \|Q_j\|_2^2 \left(\sum_i \rho_i^2\right), \max_i \|P_i\|_2^2 \left(\sum_j \pi_j^2\right)\right).
\end{aligned}$$

Since $P = U\sqrt{\Sigma}$ and $Q = V\sqrt{\Sigma}$, we have $\max_i \|P_i\|_2^2 = \max_i \sum_{j=1}^r \sigma_j U_{x,j}^2 \leq \sigma_1 \|U\|_{2\to\infty}^2$ and similarly, $\max_j \|Q_j\|_2^2 \leq \sigma_1 \|V\|_{2\to\infty}^2$. This finally ensures that

$$K_{M,\pi} \leq \frac{1}{\omega_{\max}} \max(\|U\|_{2\to\infty}^2, \|V\|_{2\to\infty}^2) \max\left(\sum_x \rho_i^2, \sum_a \pi_j^2\right)$$

$$\leq \frac{\mu^2 r \max(1/m, 1/n)}{\omega_{\max}} \max\left(\sum_x \rho_i^2, \sum_a \pi_j^2\right)$$

$$= \frac{\mu^2 r}{(m \wedge n)\,\omega_{\max}} \max\left(\sum_x \rho_i^2, \sum_a \pi_j^2\right).$$

For the second bullet point, let $M \in \mathcal{H}$. The first bullet point and the homogeneity properties of M entail $K_{M,\pi} \lesssim \frac{\max\left(\sum_i \rho_i^2, \sum_j \pi_j^2\right)}{(m\wedge n)\omega_{\max}} \lesssim \frac{\max\left(\sum_i \rho_i^2, \sum_j \pi_j^2\right)}{(m\vee n)\omega_{\max}}$, where we have used $m = \Theta(n)$ in the last inequality. To achieve this upper bound, we choose a homogeneous matrix of rank r such that U can be chosen with a first column equal to $(1/\sqrt{m}, \ldots, 1/\sqrt{m})^\top$, and V can be chosen with a first column equal to $(1/\sqrt{n}, \ldots, 1/\sqrt{n})^\top$. An explicit example for $r = 1$ is the matrix that has each entry equal to 1. For such a choice, we have

$$\sum_j \|P_\pi^j\|_2^2 = \sum_j \left\|\sum_i \rho_i \pi(j|i) P_i\right\|_2^2$$

$$\geq \sum_j \left(\sum_i \rho_i \pi(j|i) \sqrt{\sigma_1} U_{x,1}\right)^2,$$

$$= \frac{\sigma_1}{m}\left(\sum_j \pi_j^2\right),$$

and

$$\sum_i \|Q_\pi^i\|_2^2 = \sum_i \left\|\sum_j \rho_i \pi(j|i) Q_j\right\|_2^2$$

$$\geq \sum_i \left(\sum_i \rho_i \pi(j|i) \sqrt{\sigma_1} V_{j,1}\right)^2,$$

$$= \frac{\sigma_1}{n}\left(\sum_i \rho_i^2\right).$$

In particular,

$$K_{M,\pi} \geq \frac{\max\left(\sum_i \rho_i^2, \sum_j \pi_j^2\right)}{(m \vee n)\,\omega_{\max}}.$$

Lastly, notice that the supremum is largest when π always selects the same arm, independently of the context[16]. In this case, there is a reward matrix $M \in \mathcal{H}$ such that we have the sampling lower bound

$$T \geq \frac{\sigma^2 \,\mathrm{kl}(\delta, 1-\delta)}{2(m \vee n)\,\omega_{\max}\varepsilon^2}.$$

Finally, notice that $m = \Theta(n)$ entails $m \vee n = \Theta\left(\frac{mn}{m+n}\right)$. $\qquad\square$

---

[16]Intuitively, when the behavior policy is uniform, the distribution shift is maximized when the target policy is constant.

F. Proofs of results from Section 7: Best Policy Identification

F.1. Proof of Theorem 7.1

In this appendix, we denote by Π the set of deterministic policies (in other words, the set of functions from $[m]$ to $[n]$). To ease the notation, we also denote by $\hat{v}_{\text{RS}}^{\pi}$ the $\texttt{RS-PE}$ estimator of v^{π} for a given policy π, and by $\hat{\pi}_{\text{RS}}$ the policy learned by $\texttt{RS-BPI}$.

We first derive an upper bound on the max-norm error of the reward matrix estimator $\overline{M}$ utilized in our PE and BPI algorithms, which is defined by

$$\overline{M}_{i,j} := \phi_{i,j}^T \hat{\theta}. \tag{41}$$

Proposition F.1. *Assume that the regularization parameter of $\hat{\Lambda}_\tau$ verifies $\tau \leq \frac{r}{(m \wedge n)} \frac{\omega_{\min}}{\omega_{\max}} \log(16d/\delta)$. With probability larger than $1 - \delta$,*

$$\|M - \overline{M}\|_{\max} \lesssim \varepsilon$$

as soon as

$$T \gtrsim \frac{\sigma^2 \mu^2 r (m+n)}{\omega_{\min} mn \, \varepsilon^2} \log\left(\frac{m+n}{\delta}\right) + \frac{K_0}{\varepsilon} \log^3\left(\frac{(m+n)T}{\delta}\right)$$

where K_0 has been defined in (28).

Proof of Proposition F.1. By replacing ϕ_π by $\phi_{i,j}$ in the error analysis performed in Appendix D, we directly obtain that for every (context, arm) pair (i, j), with probability larger than $1 - \frac{\delta}{mn}$,

$$|M_{i,j} - \phi_{i,j}^T \hat{\theta}| \leq \varepsilon$$

as soon as

$$T \gtrsim \frac{\sigma^2 \mu^2 r (m+n)}{\omega_{\min} mn \, \varepsilon^2} \log\left(\frac{emn}{\delta}\right) + \frac{K_0}{\varepsilon} \log^3\left(\frac{mn(m+n)T}{\delta}\right).$$

Note that $\log\left(\frac{emn}{\delta}\right) \lesssim \log\left(\frac{m+n}{\delta}\right)$ and $\log^3\left(\frac{mn(m+n)T}{\delta}\right) \lesssim \log^3\left(\frac{(m+n)T}{\delta}\right)$. A union bound ensures that with probability larger than $1 - \delta$,

$$\|M - \overline{M}\|_{\max} = \max_{i,j} |M_{i,j} - \phi_{i,j}^T \hat{\theta}| \leq \varepsilon$$

under the stated condition on the sample complexity. $\qquad\square$

We also wish to note that, although this will not be used in the proof of Theorem 7.1, the same argument can be utilized to derive a max-norm version of the PE error bound (37). Namely, under the assumptions of Proposition D.2,

$$\|M - \overline{M}\|_{\max} \leq \mu \sqrt{\frac{r(m+n)}{\omega_{\min} mn}} \sqrt{\frac{2\sigma^2 \log(16mn/\delta)}{(1-\alpha)T}} + K_0 \frac{\log^3(mn(m+n)T/\delta)}{T} \tag{42}$$

with probability larger than $1 - \delta$, where K_0 has been defined in (28).

We are now ready to prove Theorem 7.1.

Proof of Theorem 7.1. We provide a detailed proof of the second statement of the theorem first. Let us recall that

$$\max_{\pi \in \Pi} |v^\pi - \hat{v}_{\text{RS-PE}}^\pi| \leq \max_{\pi \in \Pi} \max_{i,j} |M_{i,j} - \overline{M}_{i,j}| \sum_{i,j} w_{i,j}^\pi \leq \|M - \overline{M}\|_{\max}. \tag{43}$$

Note that for a homogeneous context distribution, we have $\omega_{\min} = \Theta\left(\frac{1}{mn}\right)$ since the learner samples the arms uniformly. Consequently, when $T \gtrsim \frac{\sigma^2 \mu^2 r(m+n)}{\varepsilon^2} \log\left(\frac{m+n}{\delta}\right) + \frac{K_0}{\varepsilon} \log^3\left(\frac{(m+n)T}{\delta}\right)$, Proposition F.1 yields $v^* - v_{\text{RS-PE}}^{\hat{\pi}} \leq \varepsilon$ with

probability larger than $1 - \delta$. Indeed, $v^* - v^{\hat{\pi}_{\text{RS}}} \leq |v^{\hat{\pi}_{\text{RS}}} - \hat{v}^{\hat{\pi}_{\text{RS}}}_{\text{RS}}| + |\hat{v}^{\hat{\pi}_{\text{RS}}}_{\text{RS}} - v^*|$. Under the event $\|M - \overline{M}\|_{\max} \leq \varepsilon/2$, The first term can be directly controlled by $\varepsilon/2$ with (43). For the second term, note that $\hat{\pi}_{\text{RS}} = \arg\max_\pi \hat{v}^\pi_{\text{RS}}$. Consequently, under the same event, $\hat{v}^{\hat{\pi}_{\text{RS}}}_{\text{RS}} \geq \hat{v}^{\pi^*}_{\text{RS}} \geq v^* - \varepsilon/2$, and $v^* \geq v^{\hat{\pi}_{\text{RS}}} \geq \hat{v}^{\hat{\pi}_{\text{RS}}}_{\text{RS}} - \varepsilon/2$. By Proposition F.1, with probability larger than $1 - \delta$, we have

$$v^* - v^{\hat{\pi}_{\text{RS}}} \leq \varepsilon \text{ for } T \gtrsim \frac{\sigma^2 \mu^2 r (m+n)}{\varepsilon^2} \log\left(\frac{m+n}{\delta}\right) + \frac{K_0}{\varepsilon} \log^3\left(\frac{(m+n)T}{\delta}\right)$$

for

$$K_0 \lesssim \frac{L^2 \mu^6 \kappa^4 r^3 (m+n)^2 \sqrt{mn}}{\|M\|_{\max} (m \wedge n)}. \tag{44}$$

We finally recall that from Remark D.8, $K_0 \lesssim \sigma mn$ when M is also homogeneous and $\|M\|_{\max} = \Theta(\sigma)$.

We can of course derive the corresponding result for SBPI in the same way. More specifically, we obtain that $v^\star - v^{\hat{\pi}_{\text{SBPI}}} \leq \varepsilon$ with probability larger than $1 - \delta$ as soon as

$$T \gtrsim \frac{L^2 \mu^6 \kappa^4 r^3 (m+n) mn}{(m \wedge n)^2 \varepsilon^2} \log^3\left(\frac{(m+n)T}{\delta}\right)$$

by combining the max-norm guarantee on $\widehat{M}$ of Proposition D.1 with the previous arguments[17]. □

F.2. General context distribution case

We now explain how to drop the homogeneous context assumption in Theorem 7.1 while retaining the optimal scaling up to logarithmic factors with the technique highlighted in (Lee et al., 2023). The idea is to split the set of contexts into subsets on which the context distribution is homogeneous and to apply RS-BPI or SBPI on each subset. Let

$$\mathcal{L} = \lceil \log_2(m/\varepsilon) \rceil \tag{45}$$

and define $\mathcal{C}_\mathcal{L} = \left\{ i \in [m], 2^{-l-1} < \rho_i \leq 2^{-l} \right\}$ for $0 \leq l \leq \mathcal{L} - 1$, and $\mathcal{C}_\mathcal{L} = \left\{ i \in [m], \rho_i \leq 2^{-\mathcal{L}} \right\}$. On each $\mathcal{C}_l$ for $0 \leq l \leq \mathcal{L} - 1$, the distribution defined by the normalized context weights is homogeneous. On $\mathcal{C}_\mathcal{L}$, the context weights are small enough to not meaningfully contribute to the BPI error. We summarize below the algorithm structure for RS-BPI, and we recall that α is the proportion of samples used in the first phase of the algorithm. Of course, a SBPI version of this algorithm can be constructed in the same way.

Algorithm 4 GENERALIZED **RS**-BPI (G-RS-BPI)

Input: Budget of rounds T, data splitting parameter α, context distribution ρ, regularization parameter τ, confidence level δ
for $t = 1$ **to** T **do**
    Observe a context i_t according to ρ
    Select an arm j_t uniformly at random
end for
Split $[m]$ into the subsets $\mathcal{C}_l$ defined above
for $l = 0, \ldots, \mathcal{L} - 1$ **do**
    Execute Algorithm 1 on $\mathcal{C}_l$ to learn $\hat{\pi}^l_{\text{RS}}$
end for
Output: $\hat{\pi}_{\text{RS}}$ such that $\hat{\pi}_{\text{RS}} = \hat{\pi}^l_{\text{RS}}$ on $\mathcal{C}_l$ for $0 \leq l \leq \mathcal{L} - 1$ and arbitrary on $\mathcal{C}_\mathcal{L}$.

To clarify, all samples are gathered at the start. Then, for each l, we set T_l as the number of samples of the form (i_t, j_t, r_t) for $i_t \in \mathcal{C}_l$, and we construct a reward estimator $\overline{M}^l$ from the T_l samples with the definition (41). We then select $\hat{\pi}^l_{\text{RS}}$ with $\hat{\pi}^l_{\text{RS}}(i) = \arg\max_{1 \leq j \leq n} \overline{M}^l_{i,j}$. This generalized version of RS-BPI satisfies the following guarantee.

---

[17]to get the exact statement of the theorem, note that $(m+n) mn < (m+n)^3$.

Theorem F.2. *Let $\varepsilon > 0$, $\delta \in (0,1)$, and define $\rho\left(\mathcal{C}_l\right) := \sum_{i \in \mathcal{C}_l} \rho_i$. Assume that $\varepsilon < \left(\min_{0 \leq l \leq \mathcal{L}-1} \rho\left(\mathcal{C}_l\right)\right) \|M\|_{\max}$. With the choices $T_1 = \lfloor \alpha T \rfloor$ and $\tau \leq r(m \wedge n)^{-1}(\rho_{\min}/\rho_{\max}) \log(16d/\delta)$, $v^\star - v^{\hat{\pi}_{\mathrm{RS}}} \leq \varepsilon'$ with probability larger than $1 - \delta$ as soon as*

$$T \gtrsim \frac{\sigma^2 \mu^2 r\left(m+n\right)}{\varepsilon^2} \log\left(\frac{\mathcal{L}\left(m+n\right)}{\delta}\right) + \frac{K_0}{\varepsilon} \log^3\left(\frac{\mathcal{L}(m+n)T}{\delta}\right),$$

where K_0 has been defined in (44), $\mathcal{L}$ has been defined in (45) and $\varepsilon' = \varepsilon\left(\mathcal{L} + 2\|M\|_{\max}\right)$.

Proof of Theorem F.2. Let $0 \leq l \leq \mathcal{L} - 1$ and $\delta_l = \delta/\mathcal{L}$. From Lemma 9 in (Lee et al., 2023), we know that with high probability, at least

$$\frac{T\rho\left(\mathcal{C}_l\right)}{2} \gtrsim \frac{\sigma^2 \mu^2 r\left(m+n\right)}{\varepsilon_l^2} \log\left(\frac{m+n}{\delta_l}\right) + \frac{K_0}{\varepsilon_l} \log^3\left(\frac{(m+n)T}{\delta_l}\right)$$

out of the T samples correspond to a context in $\mathcal{C}_l$ for $\varepsilon_l := \varepsilon/\rho\left(\mathcal{C}_l\right)$. Since $\varepsilon_l \leq \|M\|_{\max}$, by Theorem 7.1, Algorithm 1 on $\mathcal{C}_l$ yields a policy $\hat{\pi}_{\mathrm{RS}}^l$ such that

$$\sum_{i \in \mathcal{C}_l} \frac{\rho_i}{\rho\left(\mathcal{C}_l\right)} \left(M_{i,\hat{\pi}_{\mathrm{RS}}^l(i)} - M_{i,\pi^\star(i)}\right) \leq \frac{\varepsilon}{\rho\left(\mathcal{C}_l\right)}$$

with probability larger than $1 - \delta_l$. A union bound ensures that with probability larger than $1 - \delta$, the overall BPI error is bounded as follows:

$$\sum_{i=1}^{m} \rho_i \left(M_{i,\hat{\pi}_{\mathrm{RS}}(i)} - M_{i,\pi^\star(i)}\right) = \sum_{l=0}^{\mathcal{L}-1} \sum_{i \in \mathcal{C}_l} \rho_i \left(M_{i,\hat{\pi}_{\mathrm{RS}}^l(i)} - M_{i,\pi^\star(i)}\right) + \sum_{i \in \mathcal{C}_{\mathcal{L}}} \rho_i \left(M_{i,\hat{\pi}_{\mathrm{RS}}^l(i)} - M_{i,\pi^\star(i)}\right)$$

$$\leq \varepsilon\mathcal{L} + 2m2^{-\mathcal{L}}\|M\|_{\max}$$

$$\leq \varepsilon\mathcal{L} + 2\varepsilon\|M\|_{\max}$$

$$= \varepsilon'.$$

$\square$

G. Proof of results from Section 8: Regret Minimization

In this appendix, we present the regret analysis of Algorithm 2. We postpone the proof of Theorem 8.1 to Section G.2, and first repeat the analysis of SupLinUCB algorithm given in (Takemura et al., 2021), culminating in Theorem G.1, which we then use to prove our upper bound on the regret of RS-RMIN in Section G.2.

G.1. Misspecified linear bandits

We assume that in each round t, the learner is provided with a context set of K actions $\mathcal{X}_t = \{x_{t,1}, \ldots, x_{t,K}\}$ with $x_{t,a} \in \mathbb{R}^d$ and $\|x_{t,a}\|_2 \leq 1$ for $a \in [K]$. The context sets are drawn according to some distribution ρ and are independent across rounds.

The reward of action a in round t is expressed as $r_{t,a} = x_{t,a}^\top \theta + \epsilon_{t,a} + \xi_t$. We assume that the misspecification is bounded, i.e., $\sup_{t,a} |\epsilon_{t,a}| \leq \epsilon_{\max}$ and that $\max_{t,a} |r_{t,a}| = r_{\max}$. The noise process $\{\xi_t\}_{t \geq 1}$ is i.i.d. with σ-subgaussian distribution. The regret up to round T of an algorithm π selecting a_t^π in round t is defined by:

$$\mathcal{R}^\pi(T) = \sum_{t=1}^{T} \left(r_{t,a_t^\star} - r_{t,a_t^\pi} \right),$$

where $a_t^\star = \arg\max_{a \in [A]} \mathbb{E}[r_{t,a}]$. To solve the misspecified linear bandit problem, we apply Algorithm 1 in (Takemura et al., 2021), a variant of SupLinUCB (Chu et al., 2011). The pseudo-code is presented in Algorithm 5.

Algorithm 5 SupLinUCB (Algorithm 1 in (Takemura et al., 2021))

Input: number of rounds T, threshold β and regularization matrix Λ
$J = \lceil \log_2(T/d)/2 \rceil + 1$
$\Psi_1^j = \emptyset, \ \forall j \in [J]$
for $t = 1$ **to** T **do**
  $j = 1, \mathcal{A}_1 = [K]$
  repeat
    $V_t = \Lambda + \sum_{\tau \in \Psi_t^j} x_{\tau,a_\tau} x_{\tau,a_\tau}^\top$
    $\widehat{\theta}_t = V_t^{-1} \sum_{\tau \in \Psi_t^j} x_{\tau,a_\tau} r_{\tau,a_\tau}$
    for $a \in \mathcal{A}_j$ **do**
      $\hat{r}_{t,a}^j = \langle \widehat{\theta}_t, x_{t,a} \rangle$
      $w_{t,a}^j = \|x_{t,a}\|_{V_t^{-1}} \beta$
    end for
    if $w_{t,a}^j \leq \beta \sqrt{d/T}$ for all $a \in \mathcal{A}_j$ **then**
      $a_t = \arg\max_{a \in \mathcal{A}_j}(\hat{r}_{t,a}^j + w_{t,a}^j)$
      $\Psi_{t+1}^{j'} \leftarrow \Psi_t^{j'}$ for all $j' \in [J]$
    else if $w_{t,a}^j \leq \beta 2^{-j}$ for all $a \in \mathcal{A}_j$ **then**
      $\mathcal{A}_{j+1} \leftarrow \{a \in \mathcal{A}_j : (\hat{r}_{t,a}^j + w_{t,a}^j) \geq \max_{a' \in \mathcal{A}_j}(\hat{r}_{t,a'}^j + w_{t,a'}^j) - 2^{1-j}\beta\}$
      $j \leftarrow j + 1$
    else
      Choose $a_t \in \mathcal{A}_j$ s.t. $w_{t,a_t}^j > \beta 2^{-j}$
      $\Psi_{t+1}^{j'} \leftarrow \begin{cases} \Psi_t^{j'} \cup \{t\}, & \text{if } j' = j \\ \Psi_t^{j'}, & \text{else} \end{cases}$
    end if
  until an action a_t is found.
end for

Theorem G.1 (Simplified version of Theorem 1 in (Takemura et al., 2021)). *Let $B > 0$ be such that $\|\theta\|_2 \leq B$, and let*

$\Lambda = (\sigma^2/B^2)I_d$. *Moreover, let* $\delta \in (0,1)$, $T > 0$ *and define the threshold*

$$\beta(\delta) = \sigma(1 + \sqrt{2\log\left(\frac{TK}{\delta}\left\lceil\frac{1}{2}\log\frac{T}{d}\right\rceil\right)}). \tag{46}$$

Then, there exists a universal constant $C > 0$ *such that the regret of the algorithm* $\pi = $ SUPLINUCB *satisfies:*

$$\mathcal{R}^\pi(T) \leq C\left(\sigma\sqrt{dT\log K} + \epsilon_{\max}\sqrt{dT} + dr_{\max}\right)\log(T+1)\log\left(1 + \frac{TB^2}{d\sigma^2}\right),$$

with probability at least $1 - \delta$.

The result of Theorem G.1 entails that Algorithm 5 has an expected regret[18] of order $\widetilde{O}(\sqrt{dT\log(K)} + \epsilon_{\max}\sqrt{dT})$, where $\widetilde{O}$ hides polylogarithmic factors in d and T. Therefore, SUPLINUCB enjoys tight dependence in the dimension in both the term $\widetilde{O}(\epsilon_{\max}\sqrt{dT})$ that is due to the misspecification (Lattimore et al., 2020) and in the term $\sqrt{dT\log(K)}$ that corresponds to the minimal regret without misspecification (Auer, 2002; Chu et al., 2011).

G.2. Proof of Theorem 8.1

Let us denote by $j_t^\star$ the optimal arm given context i_t, i.e., let $j_t^\star = \pi^\star(i_t)$ for all $t \in [T]$. We split the regret into the sum of two terms corresponding to the uniform exploration phase and a second phase where we apply Algorithm 2:

$$R^\pi(T) = \underbrace{\sum_{t=1}^{T_1} \mathbb{E}[M_{i_t,j_t^\star} - M_{i_t,j_t^\pi}]}_{\substack{R_1^\pi \\ \text{(Regret of phase 1)}}} + \underbrace{\sum_{t=T_1+1}^{T} \mathbb{E}[M_{i_t,j_t^\star} - M_{i_t,j_t^\pi}]}_{\substack{R_2^\pi \\ \text{(Regret of phase 2)}}}.$$

Step 1: (Regret of phase 1). Since $|M_{i,j}| \leq \|M\|_{\max}$ for all $(i,j) \in [m] \times [n]$, the first phase lasting for T_1 rounds accumulate a regret of at most:

$$R_1^\pi \leq 2T_1\|M\|_{\max}. \tag{47}$$

Step 2: (Regret of phase 2). In the second phase, which lasts for $T_2 = T - T_1$ rounds, we can identify our observations $M_{i_t,j_t} + \xi_t = \phi_{i_t,j_t}^\top\theta + \epsilon_{i_t,j_t} + \xi_t$ with rewards r_{t,j_t} from the previous section, where we define the set of arms for the context i_t by setting $(x_{t,a})_{a\in[K]} = (\phi(i_t,j))_{j\in[n]}$ with $K = n$, $x_{a_t^\star} = \phi(i_t,j_t^\star)$ and $d = r(m+n) - r^2$. Thus, the regret R_2^π generated during this phase can be equivalently written as:

$$R_2^\pi = \sum_{t=T_1+1}^{T} \mathbb{E}[r_{t,a_t^\star} - r_{t,a_t^\pi}] = \mathbb{E}[\mathcal{R}^\pi(T_2)].$$

Thus, we can apply Theorem G.1 to show that there exists a universal constant C_2 such that with probability at least $1 - \delta$:

$$\mathcal{R}^\pi(T_2) \leq C_2\sqrt{r(m+n)}\left(\sigma\sqrt{T\log\left(\frac{Tmn}{\delta}\right)} + \epsilon_{\max}T + \|M\|_{\max}\sqrt{r(m+n)}\right)\log^2\left(T(m+n)\frac{L^2}{\sigma^2}\right) \tag{48}$$

where we used that $T_2 \leq T$. To bound the second term in the parenthesis, we use the upper bound on the misspecification $\epsilon_{\max}$ of Corollary 5.1. Indeed, we have that with probability at least $1 - \delta$:

$$\epsilon_{\max} \leq \frac{CL^2\kappa^3\mu^2 r(m+n)}{\sigma_r T_1\omega_{\min}(m\wedge n)}\log^3\left(\frac{(m+n)T_1}{\delta}\right), \tag{49}$$

---

[18]The regret guarantee of Theorem G.1 is stated with high probability, but the result can be immediately used to bound the regret in expectation by choosing for example $\delta = 1/T$.

provided that $T_1 \geq c \frac{L^2(m+n)}{\sigma_r^2 \omega_{\min}} \log^3 \left(\frac{(m+n)T}{\delta} \right)$, for some universal constants $C, c > 0$. Define $\mathcal{E}$ as the event under which (49) holds. Next, we decompose R_2^π as follows:

$$R_2^\pi = \sum_{t=T_1+1}^{T} \mathbb{E} \left[r_{t,a_t^\star} - r_{t,a_t^\pi} \right] \leq 2 \|M\|_{\max} \mathbb{P}(\mathcal{E}^c)T + \sum_{t=T_1+1}^{T} \mathbb{E} \left[\left(r_{t,a_t^\star} - r_{t,a_t^\pi} \right) \mathbb{1}_{\{\mathcal{E}\}} \right]$$

$$\leq 2 \|M\|_{\max} \delta T + \sum_{t=T_1+1}^{T} \mathbb{E} \left[\left(r_{t,a_t^\star} - r_{t,a_t^\pi} \right) \mathbb{1}_{\{\mathcal{E}\}} \right].$$

Define $\mathcal{L}_{mnT} = \log(1 + mnT)$. Substituting $\epsilon_{\max}$ with its upper bound in (49) and selecting $\delta = 1/T$ in (48) gives:

$$R_2^\pi \leq C_3 \sqrt{r(m+n)} \left(\sigma \sqrt{T \mathcal{L}_{mnT}} + T \frac{L^2 \kappa^3 \mu^2 r(m+n)}{\sigma_r T_1 \omega_{\min}(m \wedge n)} \mathcal{L}_{mnT}^3 + \|M\|_{\max} \sqrt{r(m+n)} \right) \log^2 \left(T(m+n) \frac{L^2}{\sigma^2} \right) \quad (50)$$

provided that $T_1 \geq c \frac{L^2(m+n)}{\sigma_r^2 \omega_{\min}} \log^3 \left((m+n)T^2 \right)$, for some universal constant $C_3, c > 0$ and $T_1 \leq T$.

Step 3: (Optimizing T_1). Observe that the upper bound of R_1^π scales as T_1 (see (47)) while the second term in the upper bound of R_2^π scales as $1/T_1$ (see (50)). Therefore, to obtain a tight regret bound on $R^\pi(T)$, we need to balance these two terms that depend on T_1 with a proper choice of T_1. We can easily verify that an optimal choice is:

$$T_1^\star = \sqrt{\frac{C_3}{2}} \frac{\mu \kappa^{3/2} L}{\sqrt{\|M\|_{\max}}} \frac{r^{3/4}(m+n)^{3/4}}{\sqrt{\sigma_r \omega_{\min}(m \wedge n)}} \sqrt{T} \log^{3/2} (1 + mnT) \log \left(T(m+n) \frac{L^2}{\sigma^2} \right).$$

Next, note that we can use that $\|M\|_{\max}/\sigma_r(M) \leq \mu^2 \kappa r / \sqrt{mn}$ from Lemma B.8 and the fact that $\omega_{\min} = \rho_{\min}/n$ to simplify further the last quantity and define:

$$T_1 = \sqrt{\frac{C_3}{2}} \frac{\mu^2 \kappa^2 L}{\|M\|_{\max}} \frac{r^{5/4}(m+n)^{3/4}(mn)^{1/4}}{\sqrt{m\rho_{\min}(m \wedge n)}} \sqrt{T} \log^{3/2} (1 + mnT) \log \left(T(m+n) \frac{L^2}{\sigma^2} \right). \quad (51)$$

This choice of T_1 entails that the regret can be upper bounded as follows:

$$R^\pi(T) = \widetilde{O} \left(\mu^2 \kappa^2 r^{5/4} L \frac{(m+n)^{3/4}(mn)^{1/4}}{\sqrt{m\rho_{\min}(m \wedge n)}} \sqrt{T} \right).$$

Step 4: (Checking conditions on T_1). It remains to check whether our choice of T_1 verifies $T_1 \geq c \frac{L^2(m+n)}{\sigma_r^2 \omega_{\min}} \log^3 \left(\frac{(m+n)T}{\delta} \right)$ and $T_1 \leq T$. These conditions are satisfied if

$$T = \widetilde{\Omega} \left(\mu^4 \kappa^4 r^{5/2} L^2 \frac{(m+n)^{3/2}\sqrt{mn}}{\|M\|_{\max}^2 m\rho_{\min}(m \wedge n)} \right).$$

Now, note that when the above condition does not hold, i.e.,

$$T = \widetilde{O} \left(\mu^4 \kappa^4 r^{5/2} L^2 \frac{(m+n)^{3/2}\sqrt{mn}}{\|M\|_{\max}^2 m\rho_{\min}(m \wedge n)} \right),$$

then the trivial bound $R^\pi(T) \leq 2T\|M\|_{\max}$, gives us after a few simple computations the upper bound

$$R^\pi(T) \leq 2\|M\|_{\max}T = \widetilde{O} \left(\mu^2 \kappa^2 r^{5/4} L \frac{(m+n)^{3/4}(mn)^{1/4}}{\sqrt{m\rho_{\min}(m \wedge n)}} \sqrt{T} \right).$$

If, furthermore, M and ρ are homogeneous, we obtain, for any $T \geq 1$:

$$R^\pi(T) = \widetilde{O} \left(L(m+n)^{3/4}\sqrt{T} \right).$$

H. Reduction to almost low-dimensional linear bandits and its limitations

We start by observing that our low-rank bandit settings can be viewed as an instance of the bilinear bandit problem (Jun et al., 2019), where for any $i, j \in [m] \times [n]$, $M_{i,j} = e_i^\top M e_j$, and we may think of e_i and e_j as the left and right feature respectively. Similarly to Section 5, we define extended vectors $\phi_{i,j}^{\text{ext}}, \theta^{\text{ext}} \in \mathbb{R}^{mn}$ as follows:

$$
\phi_{i,j}^{\text{ext}} = \begin{bmatrix} \text{vec}(\widehat{U}^\top e_i e_j^\top \widehat{V}) \\ \text{vec}(\widehat{U}^\top e_i e_j^\top \widehat{V}_\perp) \\ \text{vec}(\widehat{U}_\perp^\top e_i e_j^\top \widehat{V}) \\ \text{vec}(\widehat{U}_\perp^\top e_i e_j^\top \widehat{V}_\perp) \end{bmatrix}, \quad \theta^{\text{ext}} = \begin{bmatrix} \text{vec}(\widehat{U}^\top M \widehat{V}) \\ \text{vec}(\widehat{U}^\top M \widehat{V}_\perp) \\ \text{vec}(\widehat{U}_\perp^\top M \widehat{V}) \\ \text{vec}(\widehat{U}_\perp^\top M \widehat{V}_\perp) \end{bmatrix}.
$$

We note that according to (5), $M_{i,j} = \langle \phi_{i,j}^{\text{ext}}, \theta^{\text{ext}} \rangle$. Moreover, one can readily show that, for any $(i, j) \in [m] \times [n]$.

$$
\|\phi_{i,j}^{\text{ext}}\|_2 \leq 1, \quad \|\theta^{\text{ext}}\|_2 = \|M\|_F \leq \sqrt{nm}\|M\|_{\max}. \tag{52}
$$

We can also write $\theta^{\text{ext}} = \begin{bmatrix} \theta^\top & \theta_\ell^\top \end{bmatrix}^\top$, where we denote $\theta_\ell = \text{vec}(\widehat{U}_\perp^\top M \widehat{V}_\perp) \in \mathbb{R}^{(m-r) \times (n-r)}$. Observe that

$$
\|\theta_\ell\|_2 = \|\widehat{U}_\perp^\top M \widehat{V}_\perp\|_F \leq \|\widehat{U}_\perp^\top U\|_F \|M\|_{\text{op}} \|\widehat{V}_\perp^\top V\|_F \leq \|UU^\top - \widehat{U}\widehat{U}^\top\|_F \|VV^\top - \widehat{V}\widehat{V}^\top\|_F \sqrt{nm}\|M\|_{\max}. \tag{53}
$$

Thus, one may use guarantees on the subspace recovery in the Frobenius norm such as Lemma B.7 to show that the impact of θ_ℓ in the reward is much smaller than that of θ. We have with probability at least $1 - \delta$:

$$
\|\theta_\ell\|_2 \lesssim \frac{L^2 \kappa r(m+n)}{T_1 \omega_{\min} \sigma_r} \log^3\left(\frac{(m+n)T_1}{\delta}\right)
$$

provided that $T_1 \gtrsim \frac{1}{\omega_{\min}(m \wedge n)} \log^3\left(\frac{(m+n)T_1}{\delta}\right)$.

From these observations, together with the reformulation $M_{i,j} = \langle \phi_{i,j}^{\text{ext}}, \theta^{\text{ext}} \rangle$, one can claim a reduction from a low-rank bandit model to an almost low-dimensional linear bandit (Valko et al., 2014; Jun et al., 2019; Kocák et al., 2020; Kang et al., 2022). More precisely, the sub-vector corresponding to the last $(m-r)(n-r)$ components of θ (i.e., those corresponding to θ_ℓ) do not have a significant impact on the rewards, hence reducing the effective dimension of this linear bandit to $d = r(m+n) - r^2$.

We can derive an analogue of Theorem 8.1 with a similar regret analysis:

Proposition H.1. *Let B_2 be an upper bound of $\|\theta^{\text{ext}}\|_2$ and let B_ℓ be an upper bound of $\|\theta_\ell\|_2$. Define $T_1 = \widetilde{\Theta}(\mu^2 \kappa r^{3/2} \sqrt{\frac{m+n}{\omega_{\min}\sqrt{mn}}})\sqrt{T})$, and let $\lambda = B_2^{-2}$, $\lambda_\perp = \frac{T}{d\log(1+\frac{T}{\lambda})}$ and:*

$$
\Lambda = \text{diag}(\overbrace{\lambda, \ldots, \lambda}^{d}, \overbrace{\lambda_\perp, \ldots, \lambda_\perp}^{mn-d}), \qquad \beta = \sigma\sqrt{2\log\left(\frac{10Tmn}{\delta}\log(1+T)\right)} + B_2\sqrt{\lambda} + B_\ell\sqrt{\lambda_\perp}.
$$

Then, Algorithm 2 with SUPLINUCB parameters $T - T_1, \beta(\delta), \Lambda$ has an expected regret that satisfies $R_T^\pi = \widetilde{O}\left(L\mu\kappa r \frac{\sqrt{m+n}(mn)^{1/4}}{\sqrt{m}\rho_{\min}}\sqrt{T}\right)$.

The proof of the proposition only differs slightly from the proof of Theorem 8.1. Instead of Theorem G.1, we combine the analysis of SUPLINUCB[19] from (Chu et al., 2011), Lemma 38 in (Kocák et al., 2020), Lemma 3 (Jun et al., 2019), and Lemma 11 in (Abbasi-Yadkori et al., 2011) to show that the second phase of this algorithm has regret as follows:

$$
\mathcal{R}^\pi(T) \lesssim \sqrt{dT}\left(r_{\max} + \sigma\sqrt{\log(eTK/\delta)} + B_2\sqrt{\lambda} + B_\ell\sqrt{\lambda_\perp}\right)\sqrt{\log(1+T/\lambda)\log T}
$$

with probability at least $1 - \delta$. Proposition H.1 follows straightforwardly from this result and our bounds B_2, B_ℓ derived above. Note that in the homogeneous setting (see Definition 3.1), the regret above scales as $\widetilde{O}((m+n)\sqrt{T})$.

---

[19]Here we used SUPLINUCB as given in (Chu et al., 2011) for the analysis.

The reduction presented above has the following limitations:

(i) First, it does not take advantage of the delocalization of the subspace recovery error, i.e., of the fact that this error spreads out along $m + n$ dimensions. This delocalization property is quantified through our subspace recovery error guarantees in the two-to-infinity norm, and it has subtle but important implications. Indeed, it can be leveraged to perform a reduction to an almost low-dimensional linear bandit that is also *almost sparse*, in the sense that $\|\theta_\ell\|_\infty \lesssim \|\theta_\ell\|_2/\sqrt{mn}$. A reduction to a *misspecified* linear bandit leverages this structure and allows us to devise algorithms with tighter regret guarantees.

(ii) Then, since the problem can be reduced to an almost-sparse linear bandit, using algorithms such as LOWDIM-OFUL (Jun et al., 2019; Kang et al., 2022) does not give tight regret guarantees in our setting. Indeed, LOWDIM-OFUL relies on least squares estimators obtained with a weighted Euclidean norm regularization. This type of regularization accounts for the low-dimensionality of the problem, but fails at exploiting the sparsity of the problem. To this aim, one should use an ℓ_1-norm regularization instead. In fact, even adapting algorithms for linear bandits such as SUPLINUCB (Chu et al., 2011) that have typically tighter dimension dependencies than OFUL does not yield tighter bounds. As we state in Proposition H.1, the corresponding regret upper bound scales as $\widetilde{O}((m + n)\sqrt{T})$ in the homogeneous case, compared to $\widetilde{O}((m + n)^{3/4}\sqrt{T})$ achieved using Algorithm 2 in Section 8.

(iii) Finally, a reduction to misspecified linear bandit is more appealing computationally because the resulting feature vectors are of dimension $r(m + n) - r^2$. In the case of a reduction to an almost low-dimensional linear bandit, these vectors are of dimension mn. This means that instead of having to invert[20] $mn \times mn$ matrices, we will only have to invert $(r(m + n) - r^2) \times (r(m + n) - r^2)$ matrices.

---

[20]Indeed, almost all existing algorithms for linear and contextual linear bandits require inverting the features matrix during a regression step that involves least squares estimation (Lattimore & Szepesvári, 2020).

I. Additional discussions

I.1. Knowledge of the context distribution

We decided to assume knowledge of ρ to simplify the exposition of our work, as dealing with unknown ρ does not add any insight to our work nor does it harm any of our guarantees, as we show in this section. After T observations, a natural estimator $\hat{\rho}_T$ of ρ is:

$$\forall i \in [m], \qquad \hat{\rho}_{i,T} = \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\{i_t = 1\}.$$

Next, an immediate application of Bernstein's inequality gives

$$\mathbb{P}(|\hat{\rho}_{i,T} - \rho_i| > \varepsilon \rho_i) \leq 2 \exp\left(-\frac{T\rho_i^2 \varepsilon^2}{2\sigma_i^2 + 2\varepsilon \rho_i/3}\right) \leq 2 \exp\left(-\frac{T\rho_i^2 \varepsilon^2}{2\rho_i + 2\rho_i \varepsilon/3}\right) \leq 2 \exp\left(-T\rho_i \left(\frac{\varepsilon^2}{2 + 2\varepsilon/3}\right)\right)$$

where we used $\sigma_i^2 := \mathbb{E}[(\mathbb{1}\{i_t = i\} - \rho_i)^2] = \rho_i - \rho_i^2 \leq \rho_i$. Using a union bound and choosing for instance $\varepsilon = 1/2$, we obtain that

$$\mathbb{P}(\exists i \in [m], \ |\hat{\rho}_{i,t} - \rho_i| > \rho_i/2) \leq 2 \sum_{i=1}^{m} \exp\left(-\frac{3T\rho_i}{16}\right) \leq 2m \exp\left(-\frac{3T\rho_{\min}}{16}\right).$$

Thus, for all $\delta \in (0,1)$,

$$\mathbb{P}(\forall i \in [m], \ |\hat{\rho}_{i,t} - \rho_i| \leq \rho_i/2) \geq 1 - \delta$$

provided that

$$T \geq \frac{16}{3\rho_{\min}} \log\left(\frac{2m}{\delta}\right).$$

In view of the above result, estimating ρ always requires fewer samples than required in Theorem 4.1. To see that, observe that by Lemma B.8 and the inequality $\omega_{\min} \leq \rho_{\min}/n$,

$$\frac{L^2(m+n)}{\sigma_r^2 \omega_{\min}} \log^3\left(\frac{(m+n)T}{\delta}\right) \gtrsim \frac{m+n}{\omega_{\min} mn} \log\left(\frac{2m}{\delta}\right) \gtrsim \frac{1}{\rho_{\min}} \log\left(\frac{2m}{\delta}\right).$$

Consequently, ρ can be assumed to be known without loss of generality.

I.2. Low-rank matrix bandits

Our framework can be applied to address a related and simpler problem known as low-rank matrix bandits. In this scenario, the learner must choose two arms, $i_t \in [m]$ and $j_t \in [n]$, at each time step $t \in T$. In contrast, in the contextual low-rank bandit setting, the row i_t is sampled according to a distribution ρ that is not chosen by the learner.

Our algorithms can be straightforwardly adapted to this setting. For instance, for regret minimization, we could utilize a two-phase algorithm. In the first phase, for $t \leq T_1$, the arm pair (i_t, j_t) is chosen uniformly at random from $[m] \times [n]$. In the second phase, for $t = T_1 + 1, \ldots, T$, (i_t, j_t) is selected by SUPLINUCB with a fixed set of feature vectors $\mathcal{X} = \{\phi(i,j), i \in [m], j \in [n]\}$ where $\phi(i,j)$ is constructed from the estimated singular subspaces according to (7). The parameters used in SUPLINUCB are the same as for contextual low-rank bandits, except that we set $\omega_{\min} = 1/(mn)$.

Similarly to Theorem 8.1, it can be proven that such an algorithm would achieve a regret of at most $\widetilde{O}(L\mu^2 \kappa^2 r^{5/4}(m+n)^{3/4}(mn)^{1/4}\sqrt{T/(m \wedge n)})$ under an appropriate condition on T (see Theorem 8.1). When M is homogeneous, the regret scales as $\widetilde{O}(L(m+n)^{3/4}\sqrt{T})$.

Most existing work on low-rank bandits lacks minimax regret guarantees. (Katariya et al., 2017) introduced RANK1ELIM, achieving regret $\widetilde{O}((m+n)\log(T)/\tilde{\Delta}_{\min})$[21] for a rank-1 reward matrix. (Kveton et al., 2017) extended this result to rank r, with additional strong assumptions on the reward matrix in their LOWRANKELIM algorithm. (Trinh et al., 2020) provided the first algorithm with asymptotically optimal instance-dependent regret for rank-1 bandits. (Bayati et al., 2022) proposed an algorithm enjoying a minimax regret guarantee, but their upper bounds scale at least with mn through the constant C_2 in the worst case (see Theorem 2 in (Bayati et al., 2022)), and requires tuning a filtering resolution h, making their guarantees

---

[21] Here $\tilde{\Delta}_{\min}$ corresponds to a notion of minimum gap different from the standard one - see (Katariya et al., 2017).

hard to evaluate. (Stojanovic et al., 2024) presented SME-AE algorithm, leveraging the entry-wise matrix estimation guarantees to obtain a regret upper bound of order $\widetilde{O}((m+n)(\bar{\Delta}/\Delta_{\min}^2)\log^3(T))$ where $\bar{\Delta}$ is the average reward gap, and $\Delta_{\min}$ is the minimum reward gap.

While these gap-dependent bounds require strong assumptions, e.g., (Stojanovic et al., 2024) assume that $\Delta_{\max}/\Delta_{\min} \leq \zeta$ where $\Delta_{\max}$ is the maximum reward gap, we provide the first algorithm with a minimax regret of order $\widetilde{O}((m+n)^{3/4}\sqrt{T})$ without such assumptions.

I.3. Improved max-norm guarantees

Our two-phase algorithm structure also yields a generic method to estimate a low-rank matrix in more general settings. The low-rank matrix estimator $\overline{M}$ defined by $\overline{M}_{i,j} = \phi_{i,j}^T \hat{\theta}$ (see (7) and (10) for a definition of these terms) enjoys a sharper max-norm guarantee than the standard estimator $\widehat{M}$ defined in (2). Indeed, recall that from Proposition D.1,

$$\|M - \widehat{M}\|_{\max} \lesssim L\sqrt{\frac{\mu^6\kappa^4 r^3(m+n)}{T\omega_{\min}(m\wedge n)^2}\log^3\left(\frac{(m+n)T}{\delta}\right)}$$

with probability larger than $1 - \delta$ when $T \gtrsim \frac{L^2\kappa^2(m+n)}{\sigma_r^2\omega_{\min}}\log^3\left(\frac{(m+n)T}{\delta}\right)$.

Additionally, from (42),

$$\|M - \overline{M}\|_{\max} \lesssim \sigma\mu\sqrt{\frac{r(m+n)}{T\omega_{\min}mn}\log\left(\frac{m+n}{\delta}\right)} + \widetilde{O}\left(\frac{K}{T}\right)$$

with probability larger than $1 - \delta$ under the assumptions of Proposition D.2, where K depends polynomially on the model parameters. When T is large enough to ensure that the first term in the error bound is larger than the second, the scaling in μ and r of the max-norm error bound is improved from $\mu^3 r^{3/2}$ to $\mu r^{1/2}$. Furthermore, the dependency in κ and $\|M\|_{\max}$ is removed in the first term (recall that $L = \|M\|_{\max} \vee \sigma$). This may appear suprising, but note that the term $\widetilde{O}\left(\frac{K}{T}\right)$ still depends on κ and $\|M\|_{\max}$ through K: more samples are required to achieve the stated scaling when κ and $\|M\|_{\max}$ are large.

I.4. Removing the dependence in $\|M\|_{\max}$ in the regret upper bound

We assume that $m = \Theta(n)$ for simplicity as in (Chen et al., 2020). Let $\widehat{M} := Z_{\mathrm{cvx},r}$ from Theorem 2 in (Chen et al., 2020) and let $\widehat{M} = \widehat{U}\widehat{\Sigma}\widehat{V}^\top$. Then, we have:

Lemma I.1. *Under the condition $\|M - \widehat{M}\|_{\mathrm{op}} \leq \sigma_r(M)/4$, it holds that:*

$$\max\left(\|U - \widehat{U}(\widehat{U}^\top U)\|_{2\to\infty}, \|V - \widehat{V}(\widehat{V}^\top V)\|_{2\to\infty}\right) \lesssim \frac{1}{\sigma_r(M)}\left(\sqrt{n}\|M - \widehat{M}\|_{\max} + \mu\sqrt{\frac{r}{n}}\|M - \widehat{M}\|_{\mathrm{op}}\right)$$

Proof. Using repeatedly that $\|AB\|_{2\to\infty} \leq \|A\|_{2\to\infty}\|B\|_{\mathrm{op}}$ for appropriate A, B, we obtain:

$$\|U - \widehat{U}\widehat{U}^\top U\|_{2\to\infty} = \|MV\Sigma^{-1} - \widehat{U}\widehat{U}^\top U\|_{2\to\infty} \leq \frac{1}{\sigma_r(M)}\|MV - \widehat{U}\widehat{U}^\top U\Sigma\|_{2\to\infty}$$

$$\leq \frac{1}{\sigma_r(M)}\|MV - \widehat{M}V\|_{2\to\infty} + \frac{1}{\sigma_r(M)}\|\widehat{M}V - \widehat{U}\widehat{U}^\top U\Sigma\|_{2\to\infty}$$

Note that the first term is bounded by $\sqrt{n}\|M - \widehat{M}\|_{\max}$ since $\|MV - \widehat{M}V\|_{2\to\infty} \leq \|M - \widehat{M}\|_{2\to\infty}\|V\|_{\mathrm{op}} \leq \sqrt{n}\|M - \widehat{M}\|_{\max}$. Before we bound the second term, note that:

$$\widehat{U}\widehat{U}^\top U\Sigma = \widehat{U}\widehat{U}^\top MV = \widehat{U}\widehat{U}^\top(M - \widehat{M})V + \widehat{U}\widehat{U}^\top\widehat{M}V$$

and $\widehat{U}\widehat{U}^\top\widehat{M}V = \widehat{U}\widehat{\Sigma}\widehat{V}^\top V = \widehat{M}V$. Thus:

$$\|\widehat{M}V - \widehat{U}\widehat{U}^\top U\Sigma\|_{2\to\infty} = \|\widehat{U}\widehat{U}^\top(M - \widehat{M})V\|_{2\to\infty} \leq \|\widehat{U}\|_{2\to\infty}\|M - \widehat{M}\|_{\mathrm{op}}.$$

Lastly, we bound $\|\widehat{U}\|_{2\to\infty}$ as follows:

$$\|\widehat{U}\|_{2\to\infty} \leq \|\widehat{U}\widehat{U}^\top U\|_{2\to\infty}\|(\widehat{U}^\top U)^{-1}\|_{\mathrm{op}} \leq (\|U\|_{2\to\infty} + \|U - \widehat{U}\widehat{U}^\top U\|_{2\to\infty})\|(\widehat{U}^\top U)^{-1}\|_{\mathrm{op}}$$

and similarly to the equation between (49) and (50) in (Stojanovic et al., 2024) we have:

$$\|(\widehat{U}^\top U)^{-1}\|_{\mathrm{op}} = \frac{1}{\sigma_r(\widehat{U}^\top U)} \leq \frac{1}{1 - \|\widehat{U}^\top U - \mathrm{sgn}(\widehat{U}^\top U)\|_{\mathrm{op}}} \leq \frac{1}{1 - \frac{2\|M-\widehat{M}\|_{\mathrm{op}}^2}{\sigma_r^2(M)}}$$

Under the condition $\|M - \widehat{M}\|_{\mathrm{op}} \leq \sigma_r(M)/4$ we consequently have:

$$\|U - \widehat{U}\widehat{U}^\top U\|_{2\to\infty} \lesssim \frac{1}{\sigma_r(M)}\left(\|M - \widehat{M}\|_{2\to\infty} + \|U\|_{2\to\infty}\|M - \widehat{M}\|_{\mathrm{op}}\right)$$

$\square$

Then, as a consequence of Lemma I.1 and Theorem 2 from (Chen et al., 2020), we obtain analogously to Theorem 4.1:

Corollary I.2. *Let us denote* $\epsilon_{\text{Sub-Rec}} := \max(d_{2\to\infty}(U, \widehat{U}), d_{2\to\infty}(V, \widehat{V}))$. *For any* $\delta \in (0, 1)$, *the following event:*

$$\epsilon_{\text{Sub-Rec}} = \widetilde{O}\left(\frac{n\sigma}{\sigma_r}\frac{1}{\sqrt{T}}\mu^3\kappa^{5/2}r^{3/2}\right)$$

holds with probability $\geq 1 - O(n^{-3})$, *provided that* $T \gtrsim \kappa^4\mu^2 rn\log n\left(\mu^2 r\log^2 n + \frac{\sigma^2 n^2}{\sigma_r^2}\right)$.

Considering the setting of Corollary I.2, the bound from Theorem 4.1 scales with $\widetilde{O}\left(\frac{nL}{\sigma_r}\frac{1}{\sqrt{T}}\mu\kappa r^{1/2}\right)$. Note that even though the dependence on $L = \|M\|_{\max} \vee \sigma$ is reduced to σ in Corollary I.2, the scaling in μ, κ and r is worse than in Theorem 4.1.

To derive the corresponding regret bound, we can rewrite the requirement on the number of samples from Corollary I.2 as $T_1 \gtrsim n\left(1 + \frac{\sigma^2}{\|M\|_{\max}^2}\right)\text{poly}(\kappa, \mu, r, \log n)$. Next, similarly to (51) in the proof of Theorem 8.1 we obtain that the regret is minimized for:

$$T_1^\star \gtrsim \frac{\sigma}{\|M\|_{\max}}\sqrt{T}n^{3/4}\text{poly}(\kappa, \mu, r, \log n)$$

Combining this with the requirement on T_1, we set:

$$T_1 := \max\left\{\frac{\sigma}{\|M\|_{\max}}\sqrt{T}n^{3/4}, n\left(1 + \frac{\sigma^2}{\|M\|_{\max}^2}\right)\right\}\text{poly}(\kappa, \mu, r, \log n)$$

Next, following the arguments from the proof of Theorem 8.1 in Appendix G.2 we obtain that combining our analysis with the guarantee from Corollary I.2 achieves:

$$R^\pi(T) \lesssim \max\left\{\sigma\sqrt{T}n^{3/4}, n\left(\|M\|_{\max} + \sigma\frac{\sigma}{\|M\|_{\max}}\right)\right\}\text{poly}(\kappa, \mu, r, \log n),$$

Comparing this result to the regret upper bound from Theorem 8.1, namely

$$R^\pi(T) = \widetilde{O}\left((\|M\|_{\max} \vee \sigma)\mu^2\kappa^2 r^{5/4}n^{3/4}\sqrt{T}\right),$$

we see that in the regime $\sigma \lesssim \|M\|_{\max}$, the regret upper bound of Theorem 8.1 scales with $\|M\|_{\max}n^{3/4}\sqrt{T}$, whereas the regret upper bound obtained by using Corollary I.2 scales with $\sigma n^{3/4}\sqrt{T} + \|M\|_{\max}n$ up to $\text{poly}(\kappa, \mu, r, \log n)$ terms.

Finally, we note that we could also use the improved max-norm bound recalled in Section I.3 to estimate the subspace recovery error related to $\overline{M}$ and that, for T large enough, it would have a milder dependence in μ, κ and r compared to Corollary I.2.

J. Numerical experiments

Notation. In this appendix, we denote by $\overline{M}$ the reward matrix estimator that `RS-PE` and `RS-BPI` leverage. It is defined by

$$\overline{M}_{i,j} = \phi_{i,j}^T \hat{\theta}$$

for $(i,j) \in [m] \times [n]$. We also denote by `SBPI` the algorithm that outputs the policy $\hat{\pi}_{\text{SBPI}}$ defined by $\hat{\pi}_{\text{SBPI}}(i) = \arg\max_{1 \le j \le n} \widehat{M}_{i,j}$. We finally note that `SIPS` (resp. `SBPI`) is denoted by DSM-PE (resp. DSM-BPI) on the graphs.

We perform numerical experiments on synthetic data with a uniform context distribution. Unless specified otherwise, the behavior policy is uniform, the target policy is chosen as the best policy: $\pi(i) = \arg\max_j M_{i,j}$ (ties are broken arbitrarily), and we generate noisy entries $M_{i_t,j_t} + \xi_t$ where $\xi_t \sim \mathcal{N}(0,1)$ is standard Gaussian, and where $M = PDQ$ for two invertible matrices $P \in \mathbb{R}^{m \times m}$, $Q \in \mathbb{R}^{n \times n}$, and $D \in \mathbb{R}^{m \times n}$ defined by $D_{i,j} = \mathbb{1}_{i=j} \mathbb{1}_{i \le r}$ (note that M is consequently of rank r). P and Q are initially generated at random with uniform entries in $[0,1]$ and their diagonal elements are replaced by the sum of the corresponding row to ensure invertibility. All experiments are performed 50 times and the shaded regions are the corresponding $[5\%, 95\%]$ confidence intervals[22].

J.1. Policy Evaluation

J.1.1. CHOICE OF HYPERPARAMETERS

Experiment 1: Impact of data splitting. For a regularization parameter of $\tau = 10^{-4}$, we compare the performance of `RS-PE` for $\alpha \in \{1/5, 1/2, 4/5\}$, where α is the proportion of samples used in the first phase of the algorithm. We also plot the error of `RS-PE` when no data splitting is performed, that is to say when all of the samples are used in both phases of the algorithm. Finally, we plot the dominant term $\frac{\|\psi_\pi\|_2}{\sqrt{\omega_{\min}}} \sqrt{\frac{2\log(16/\delta)}{(1-\alpha)T}}$ in the instance-dependent upper bound on the PE error of `RS-PE` for $\alpha = 4/5$ and $\delta = 10^{-2}$ (see Theorem D.7), which approaches its true error bound when T grows large. Note that consequently, the error of `RS-PE` can be below this dominant term. Additionally, the scaling in $1/\sqrt{1-\alpha}$ of this term suggests that using more samples in the second phase should yield better asymptotic performance. The results are presented in Figure 1.
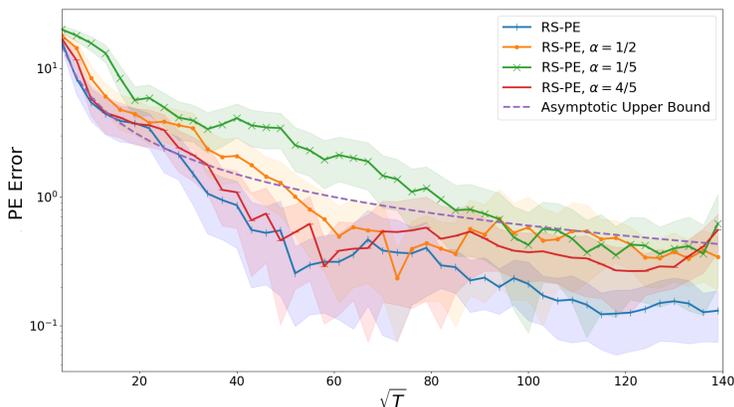

Figure 1. Impact of data splitting ($m = n = 50$, $r = 2$).

Empirically, using more samples in the first phase appears to yield faster initial performance, but this advantage is negated when the number of samples grows large. Finally, although this is not justified theoretically, not splitting the data (i.e., using the entire dataset for both phases) appears empirically more efficient.

Experiment 2: Impact of the regularization parameter τ. We compare the performance of `RS-PE` without data splitting for different values of the regularization parameter of $\hat{\Lambda}_\tau$, namely $\tau \in \{10^{-4}, 10^{-2}, 10^{-1}\}$. The results are presented in Figure 2.

---

[22]The code used in the experiments can be accessed at https://github.com/wilrev/LowRankBanditsTwoToInfinity.
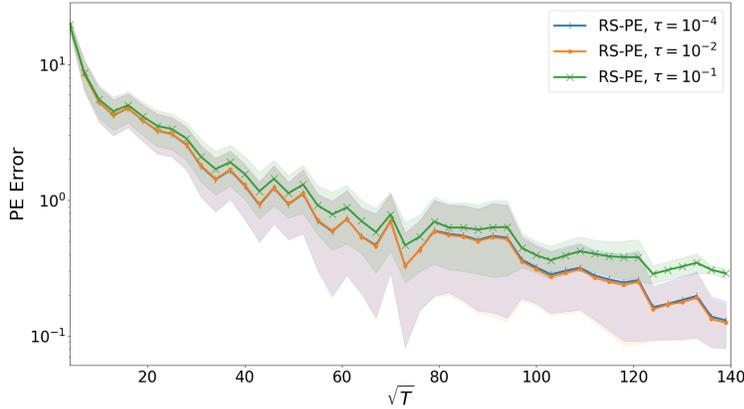
Figure 2. Impact of regularization ($m = n = 50, r = 2$)

We note that the regularization parameter does not appear to impact the error significantly as long as it is chosen small enough.

Given the previous results, in all future experiments, we do not perform data splitting: all samples are used in both phases of the algorithms. Furthermore the regularization parameter of $\hat{\Lambda}_\tau$ will be chosen small: $\tau = 10^{-4}$.

J.1.2. COMPARISON OF RS-PE WITH BENCHMARK ESTIMATORS.

Experiment 3: Scaling of the PE error with the sample size. We compare the RS-PE estimator $\hat{v}_{\text{RS-PE}} = \sum_{i,j} w^\pi_{i,j} \overline{M}_{i,j}$ with the SIPS estimator $\hat{v}_{\text{SIPS}} = \sum_{i,j} w^\pi_{i,j} \widehat{M}_{i,j}$ and the IPS estimator (Wang et al., 2017) which can also be defined as $\hat{v}_{\text{IPS}} = \sum_{i,j} w^\pi_{i,j} \widetilde{M}_{i,j}$. Similarly to RS-PE and SIPS, IPS can be interpreted as a plug-in estimator, i.e., it only relies on an estimator of the reward matrix. It is thus a suitable benchmark to demonstrate the efficiency of the reward matrix estimator $\overline{M}$ for the PE task. We also plot the asymptotic upper bound $\frac{\|\psi_\pi\|_2}{\sqrt{\omega_{\min}}} \sqrt{\frac{2\log(16/\delta)}{T}}$ of RS-PE suggested by Theorem D.7 for $\delta = 10^{-2}$. More specifically, this is the value of the dominant term in the upper bound for $\alpha = 0$, which one would expect to match the asymptotic behavior of the error of RS-PE when no data splitting is performed. The results are presented in Figure 3.
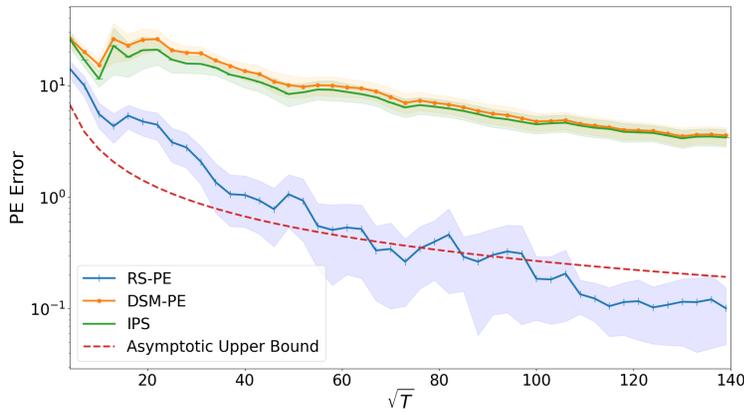

Figure 3. Sample size scaling of the PE error ($m = n = 50, r = 2$)

We note that the `RS-PE` estimator outperforms the `SIPS` and `IPS` estimators by a significant margin. The `SIPS` estimator appears to perform comparably to the `IPS` estimator, which is surprising since the latter does not leverage the low-rank structure. Furthermore, despite the restrictive theoretical condition on the number of samples to ensure that the higher-order term in the bound of Theorem D.7 is negligible, the asymptotic upper bound appears to closely match the behavior of the `RS-PE` error when the number of samples is reasonably large. We finally note that the confidence intervals for `RS-PE` only appear larger because of the logarithmic scale on the y-axis.

Experiment 4: Scaling of the PE error with the matrix size. Note that when $m = n$, Theorems 6.1 and 6.2 suggest that the PE error of `RS-PE` and `SIPS` scale with $\sqrt{m}$, while the error bounds of estimators that do not leverage the low-rank structure would typically scale with m (Yin & Wang, 2020). We perform an experiment to determine if the matrix size scaling of the error is also improved experimentally. To isolate the dependency in the size of the matrix, we do not generate a random reward matrix for each m. Rather, we ensure that the reward matrix retains the same incoherence parameter, condition number and max-norm for each m by choosing $M \in \mathbb{R}^{m \times m}$ defined by $M_{i,j} = 1$ for all $i, j \in [m]$. Furthermore, for this choice of M and our policy choices, it can be checked that the instance-dependent term in the `RS-PE` guarantee of Theorem 6.2 is simply $\frac{\|\psi_\pi\|}{\sqrt{\omega_{\min}}} = \sqrt{m}$, which further suggests that we should expect the `RS-PE` error to scale with $\sqrt{m}$ for this particular instance. The PE error of `RS-PE`, `SIPS` and `IPS` are plotted as a function of $m \in \{1, \ldots, 300\}$ on a log-log scale for $T = 10\,000$. The results are presented in Figure 4.
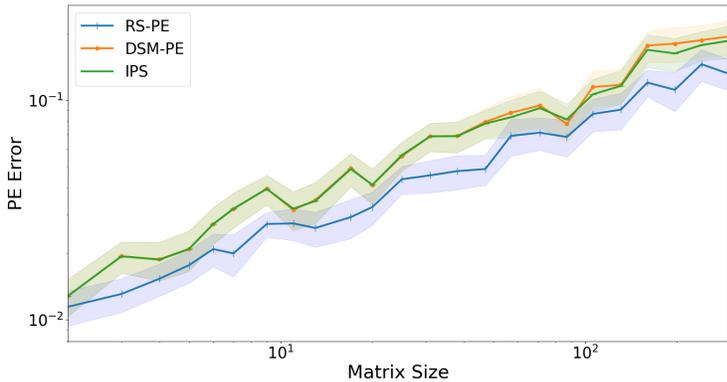

Figure 4. Matrix size scaling of the PE error ($T = 10\,000$, $r = 1$)

Surprisingly, all PE estimators appear to have a comparable scaling in m. Nonetheless, `RS-PE` outperforms the two other estimators for every m, even though we have chosen a reward matrix for which $\|M\|_{\max} = \mu = \kappa = r = 1$, so that the error guarantee of `SIPS` matches the `RS-PE` one up to logarithmic factors.

Experiment 5: Scaling of the max-norm error with the matrix size. Similarly to the PE error, when $m = n$, the max-norm error bounds of $\widehat{M}$ and $\overline{M}$ summarized in Section I.3 scale with $\sqrt{m}$. In contrast, the error bounds of matrix estimators that do not leverage the low-rank structure would typically scale with m. To determine if the scaling in the size of the matrix is also improved experimentally, we retain the same setting as Experiment 4, but we instead plot $\|M - \overline{M}\|_{\max}$, $\|M - \widehat{M}\|_{\max}$ and $\|M - \widetilde{M}\|_{\max}$ as a function of m. The results are presented in Figure 5.
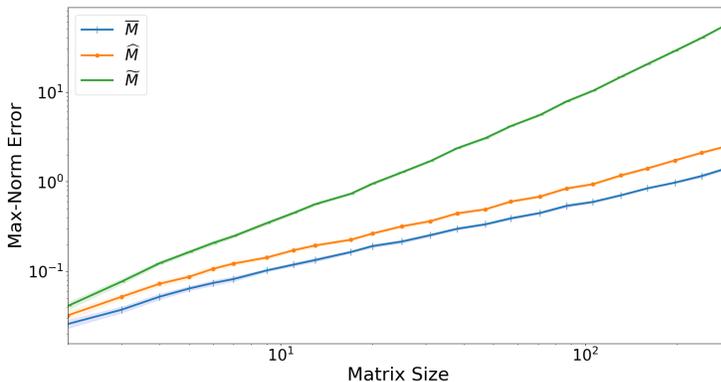
Figure 5. Matrix size scaling of the max-norm error ($T = 10\,000$, $r = 1$)

We note that $\widehat{M}$ and $\overline{M}$, the reward estimators that leverage the low-rank structure, have a max-norm error that scales noticeably better than $\widetilde{M}$. Furthermore, $\overline{M}$ consistently outperforms $\widehat{M}$, even though their max-norm error bounds match up to logarithmic factors for our reward matrix choice (see Appendix I.3 for a discussion on this matter).

J.2. Best Policy Identification

Experiment 6 : Comparison of `RS-BPI` with benchmark algorithms. We compare the value of the policy learned by `RS-BPI`, `SBPI`, and a benchmark algorithm that corresponds to `SBPI` without taking the rank-r approximation of the estimated matrix. Specifically, this benchmark outputs the policy $\hat{\pi}$ defined by $\hat{\pi}(i) = \arg\max_{1 \leq j \leq m} \widetilde{M}_{i,j}$ for all $i \in [n]$. The results are presented in Figure 6.
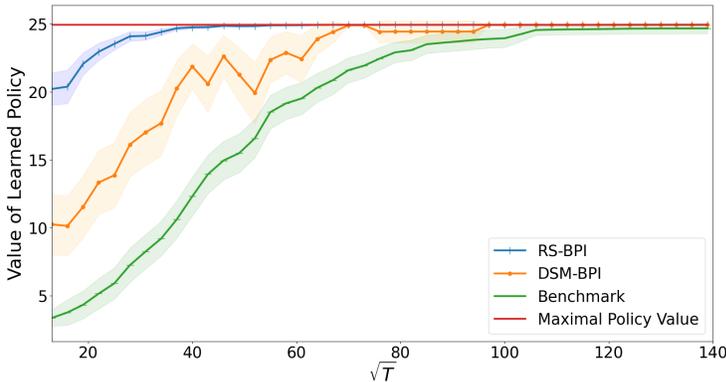

Figure 6. Value of the learned policy against the maximal policy value ($m = n = 50, r = 2$)

Both low-rank algorithms display improved performance compared to the benchmark, yet the value of the policy learned by `RS-BPI` appears to converge much quicker towards the value of the best policy than the ones of `SBPI` and the benchmark algorithm.