arXiv:2402.11156v1 [stat.ML] 17 Feb 2024

Efficient Low-Rank Matrix Estimation, Experimental Design, and Arm-Set-Dependent Low-Rank Bandits

Kyoungseok Jang [1] **Chicheng Zhang** [2] **Kwang-Sung Jun** [2]

Abstract

We study low-rank matrix trace regression and the related problem of low-rank matrix bandits. Assuming access to the distribution of the covariates, we propose a novel low-rank matrix estimation method called LowPopArt and provide its recovery guarantee that depends on a novel quantity denoted by $B(Q)$ that characterizes the hardness of the problem, where Q is the covariance matrix of the measurement distribution. We show that our method can provide tighter recovery guarantees than classical nuclear norm penalized least squares (Koltchinskii et al., 2011) in several problems. To perform efficient estimation with a limited number of measurements from an arbitrarily given measurement set $\mathcal{A}$, we also propose a novel experimental design criterion that minimizes $B(Q)$ with computational efficiency. We leverage our novel estimator and design of experiments to derive two low-rank linear bandit algorithms for general arm sets that enjoy improved regret upper bounds. This improves over previous works on low-rank bandits, which make somewhat restrictive assumptions that the arm set is the unit ball or that an efficient exploration distribution is given. To our knowledge, our experimental design criterion is the first one tailored to low-rank matrix estimation beyond the naive reduction to linear regression, which can be of independent interest.

1. Introduction and related work

In many real-world applications, data exhibit low-rank structure. For example, in the Netflix problem (Bennett et al., 2007), the user-movie rating matrix can be well-approximated by a low-rank matrix; in demographic surveys (Udell et al., 2016), the respondents' answers to the survey questions are also oftentimes modeled as a low-rank matrix. Motivated by these applications, estimation with low-rank structure is one of the central themes in high-dimensional statistics (Wainwright, 2019, Chapter 10).

We study the low-rank trace regression problem (Koltchinskii et al., 2011; Rohde & Tsybakov, 2011; Hamidi & Bayati, 2020) and the related problem of low-rank linear bandits (Jun et al., 2019; Lu et al., 2021). In the low-rank linear bandit problem, a learner sequentially learns to choose arms from a given arm set to maximize reward. For each time step $t \in \{1, \cdots, n\}$, the learner chooses an arm A_t from an arm set $\mathcal{A} \subset \mathbb{R}^{d_1 \times d_2}$, and receives a noisy reward $y_t = \langle \Theta^*, A_t \rangle + \eta_t$, where Θ^* is a rank-r matrix and η_t is σ-subgaussian noise. The learner's objective is to maximize its cumulative reward, $\sum_{t=1}^n y_t$. This low-rank bandit model is applicable to various practical scenarios (Natarajan & Dhillon, 2014; Luo et al., 2017; Jun et al., 2019).

To name a few examples, in drug discovery (Luo et al., 2017), each A_t represent the outer product $u_t v_t^\top$ of the feature representations of a pair of (drug u_t, protein v_t), and Θ^* encodes the interaction between them; in online advertising (Jain & Dhillon, 2013), each A_t represent the outer product of the feature representation of a pair of (user u_t, product v_t), and Θ^* models their interactions. The bandit problem setup naturally induces an exploration-exploitation tradeoff: as the learner does not know the reward predictor matrix Θ^*, she may need to choose arms that are informative in learning Θ^*; on the other hand, since the learner's objective is maximizing the expected reward, it may also be a good idea to choose arms that the learner believes to yield high reward, based on the past observations.

Early studies on low-rank bandits (Jun et al., 2019; Lu et al., 2021; Jang et al., 2021) have designed bandit algorithms with lower regret than naive approaches that view this problem as a $d_1 d_2$-dimensional linear bandit problem (Abbasi-Yadkori et al., 2011; Abe & Long, 1999; Auer, 2002; Dani et al., 2008). However, previous studies lack understandings on the relationship between the geometry of the arm set and regret bounds. Usually they assume that a "nice" exploration distribution over the arm set is given (Jun et al., 2019;

[1]Department of Technology, Operations and Statistics, New York University Stern School of Business, New York, NY, United States [2]Department of Computer Science, University of Arizona, Tucson, AZ, United States. Correspondence to: Kwang-Sung Jun <kjun@cs.arizona.edu>.

Lu et al., 2021; Kang et al., 2022; Li et al., 2022), or assume that the arm set has some curvature property (e.g., the unit Frobenius norm ball) (Lattimore & Hao, 2021; Huang et al., 2021). Also, some of them rely on subprocedures that are either computationally intractable (Lu et al., 2021, Algorithm 1), or nonconvex optimization steps without computational efficiency guarantees (Lattimore & Hao, 2021; Jang et al., 2021); see Appendix A for more related works.[1] To bridge this gap, we ask the following first question:

Can we develop computationally efficient low-rank bandit algorithms that allow generic arm sets and provide guarantees that adapts to the geometry of the arm set?

It is natural to apply efficient low-rank trace regression results for answering this question, since smaller estimation error leads to fewer samples for exploration thus smaller cumulative regret in bandit problems. In the low-rank trace regression problem, where a learner is given a set of measurements (X_i, y_i) that satisfy that $y_i = \langle \Theta^*, X_i \rangle + \eta_i$, where Θ^* is an unknown matrix with rank at most $r \ll \min(d_1, d_2)$, and η_i is a zero-mean σ-subgaussian noise. The goal is to recover Θ^* with low error. Throughout, we will use X_i for the supervised learning setting and A_i for the bandit setting.

The low-rank trace regression problem is one of the extensively studied areas within the field of low-rank matrix recovery problems. Keshavan et al. (2010) provides recovery guarantees for projection based rank-r matrix optimization for matrix completion, and Rohde & Tsybakov (2011); Koltchinskii et al. (2011) provide analysis of nuclear norm regularized estimation method for general trace regression, with Rohde & Tsybakov (2011) providing further analysis on the (computationally inefficient) Schatten-p-norm penalized least squares method. Among these approaches, researchers regarded the nuclear norm penalized least square (Rohde & Tsybakov, 2011; Koltchinskii et al., 2011) as the classic approach and applied this method directly (Lu et al., 2021) to achieve state-of-the-art algorithm for the low-rank bandit with a general arm set. Since better estimation leads to better bandit algorithm, we are interested in investigating the following second question:

For low-rank trace regression, can we design estimation algorithms that can outperform the classical nuclear norm penalized least squares?

In this paper, we make meaningful progress in high-dimensional low-rank regression and low-rank bandits, which provides algorithms with arm-set-adaptive exploration and regret analyses for general operator-norm-bounded arm sets.

[1]We exclude (Kang et al., 2022) from our comparison since their regret bounds involve quantities that have hidden dependence on dimensionality; see Appendix H for details.

We assume that all arms are operator norm-bounded, and the unknown parameter Θ^* is nuclear norm bounded as follows:

Assumption A1 (operator norm-bounded arm set)**.** The arm set $\mathcal{A}$ is such that $\mathcal{A} \subseteq \left\{A \in \mathbb{R}^{d_1 \times d_2} : \|A\|_{\text{op}} \le 1\right\}$.

Assumption A2 (Bounded norm on reward predictor)**.** The reward predictor has a bounded nuclear norm: $\|\Theta^*\|_* \le S_*$.

These two assumptions parallels the standard assumption in the sparse linear model where the covariates are ℓ_∞-norm bounded and the unknown parameter is ℓ_1-norm bounded (Hao et al., 2020).

First, under the additional assumption that the measurement distribution π is accessible to the learner, we propose a novel and computationally efficient low-rank estimation method called LowPopArt (Low-rank POPulation covariance regression with hARd Thresholding) and prove its estimation error guarantee (Theorem 3.4) as follows:

$$\|\hat{\Theta} - \Theta^*\|_{\text{op}} \le \tilde{O}\left(\sigma\sqrt{\frac{B(Q(\pi))}{n_0}}\right),$$

where n_0 is the number of samples used, and $B(Q(\pi))$ (see Eq. (9)) is a quantity that depends on the covariance matrix $Q(\pi)$ of the data distribution π over the measurement set $\mathcal{A}$. We show that the recovery guarantee of LowPopArt is not worse and can sometimes be much better than the classical nuclear norm penalized least squares method (Koltchinskii et al., 2011) (see Section 3).

Second, motivated by the operator norm recovery bound of LowPopArt, we propose a design of experiment objective $B(Q(\pi))$ for finding a sampling distribution that minimizes the error bound of LowPopArt. This is useful in settings when we have control on the sampling distribution, such as low-rank linear bandits, the focus of the latter part of this paper. Applying the recovery bound to the optimal design distribution, we obtain a recovery bound of

$$\|\hat{\Theta} - \Theta^*\|_{\text{op}} \le \tilde{O}\left(\sigma\sqrt{\frac{B_{\min}(\mathcal{A})}{n_0}}\right),$$

where $B_{\min}(\mathcal{A}) := \min_{\pi \in \Delta(\mathcal{A})} B(Q(\pi))$ depends on the geometry of the measurement set $\mathcal{A}$. For example, letting $d := \max\{d_1, d_2\}$, we have $B_{\min}(\mathcal{A}) = \Theta(d^2)$ and $\Theta(d^3)$ when $\mathcal{A}$ is the unit operator norm ball and unit Frobenius norm ball, respectively (See Appendix D for the proof). Moreover, optimizing our experimental design criterion is computationally tractable. In contrast, many prior works on low-rank matrix recovery require finding a sampling distribution that satisfies properties such as restricted isometry property and restricted eigenvalue (Hamdi & Bayati, 2022; Koltchinskii et al., 2011; Wainwright, 2019) - all these are computationally intractable to compute or verify and thus hard to optimize (Bandeira et al., 2013; Juditsky & Nemirovski, 2011), which is even harder when the measurements must be limited to an arbitrarily given set $\mathcal{A}$.

	Regret bound	Regret when $\mathcal{A} = \mathcal{B}_{\text{op}}(1)$	Regret when $\mathcal{A} = \mathcal{A}_{\text{hard}}$	Limitation
OFUL (Abbasi-Yadkori et al., 2011)	$\tilde{O}(d^2\sqrt{T})$	$\tilde{O}(d^2\sqrt{T})$	$\tilde{O}(d^2\sqrt{T})$	
ESTR (Jun et al., 2019)	$\tilde{O}(\sqrt{\frac{rdT}{\lambda_{\min}(Q(\pi))}}\left(\frac{\lambda_1}{\lambda_r}\right)^3)$	-	-	Bilinear
ε-FALB (Jang et al., 2021)	$\tilde{O}(\sqrt{d^3T})$	-	-	Bilinear & Comp. intractable
rO-UCB (Jang et al., 2021)	$\tilde{O}(\sqrt{rd^3T})$	-	-	Bilinear & Requires oracle
LowLOC (Lu et al., 2021)	$\tilde{O}(\sqrt{rd^3T})$	$\tilde{O}(\sqrt{rd^3T})$	$\tilde{O}(\sqrt{rd^3T})$	Comp. intractable
LowESTR[2] (Lu et al., 2021)	$\tilde{O}(d^{1/4}\sqrt{r\frac{1}{\lambda_{\min}(Q(\pi))^2}T}\left(\frac{S_*}{\lambda_r}\right))$	$\tilde{O}(\sqrt{rd^{5/2}T})$	$\tilde{O}(\sqrt{rd^{13/2}T})$	
Lower bound (Lu et al., 2021)	$\Omega(rd\sqrt{T})$			
LPA-ETC (Algorithm 3)	$\tilde{O}((S_*r^2B_{\min}(\mathcal{A})T^2)^{1/3})$	$\tilde{O}(r^{2/3}d^{2/3}T^{2/3})$	$\tilde{O}(r^{2/3}dT^{2/3})$	
LPA-ESTR (Algorithm 4)	$\tilde{O}(d^{1/4}\sqrt{B_{\min}(\mathcal{A})T}\left(\frac{S_*}{\lambda_r}\right)))$	$\tilde{O}(\sqrt{d^{5/2}T})$	$\tilde{O}(\sqrt{d^{7/2}T})$	

Table 1. A comparison with existing results on low-rank bandits with fixed arm sets and 1-subgaussian noise. Here, λ_r is abbreviation of $\lambda_r(\Theta^*)$, $Q(\pi)$ is the covariance matrix defined in Eq. (1), $\mathcal{B}_{\text{op}}(1)$ is the unit operator norm ball, $\mathcal{A}_{\text{hard}}$ is a special arm set (See Lemma 3.6), and $B_{\min}(\mathcal{A})$ is an arm set dependent constant defined in Eq. (4). When $\mathcal{A} \subseteq \mathcal{B}_{\text{op}}(1)$, we have $B_{\min}(\mathcal{A}) = \Omega(d^2)$ and $\lambda_{\min}(Q(\pi)) = O(\frac{1}{d}), \forall\pi \in \mathcal{P}(\mathcal{A})$. For the third and fourth columns, we set π to be the most favorable sampling distribution for prior results as they did not specify the sampling distribution π but assumed favorable conditions to hold. S_* is an upper bound for $\|\Theta^*\|_*$, see Assumption A2.

Finally, using LowPopArt, we propose two computationally efficient and arm set geometry-adaptive algorithms, for low-rank bandits with general arm sets:

- Our first algorithm, LPA-ETC (LowPopArt-Explore-Then-Commit; Algorithm 3), leverages the classic explore-then-commit strategy to achieve a regret bound of $\tilde{O}((S_*r^2B_{\min}(\mathcal{A})T^2)^{1/3})$ (Theorem 4.1). Compared with the state-of-the-art low-rank bandit algorithms that allow generic arm sets (Lu et al., 2021) that guarantees a regret order $\tilde{O}(\sqrt{rd^3T})$, Algorithm 3's guarantee is better when $T \ll O(\frac{d^9}{B_{\min}(\mathcal{A})^2 r})$ (see Remark 3 for a more precise statement).
- Our second algorithm, LPA-ESTR (LowPopArt-Explore-Subspace-Then-Refine; Algorithm 4), works under the extra condition that the nonzero minimum eigenvalue of Θ^*, denoted by $\lambda_{\min}$, is not too small. Algorithm 4 uses the Explore-Subspace-Then-Refine (ESTR) framework (Jun et al., 2019) and achieves a regret bound of $\tilde{O}(\sqrt{d^{1/2}B_{\min}(\mathcal{A})TS_*/\lambda_{\min}})$ (Theorem 4.2). LPA-ESTR gives a strictly better regret bound than previously-known computationally efficient algorithms. For example, compared to LowESTR (Lu et al., 2021), the regret of our LPA-ESTR algorithm makes not only a factor of $\sqrt{r}$ improvement, but also the dependence on the arm set dependent quantity from $\frac{1}{\lambda_{\min}(Q(\pi))^2}$ to $B_{\min}(\mathcal{A})$; we show that the latter is never larger than the former and that the latter can be a factor of d smaller than the former (Lemma 3.6).
- Both of our algorithms work for general arm sets, unlike many other low-rank bandit algorithms tailored for specific arm sets such as unit sphere (Huang et al., 2021), symmetric unit vector pairs $\{uu^\top : u \in \mathbb{S}^{d-1}\}$ (Kotlowski & Neu, 2019; Lattimore & Szepesvári, 2020), or even one-hot matrices $\{e_ie_j^\top : i,j \in [d]\}$(Katariya et al., 2017; Trinh et al., 2020).

We compare our regret bounds with existing results in Table 1, which showcase how our arm set-dependent regret bounds improve upon prior art in specific arm sets. We also make a meticulous examination of arm set-dependent constants on regret analysis from previous results, which we believe will help future studies.

2. Preliminaries

Basic Notations. For a matrix $M \in \mathbb{R}^{d_1\times d_2}$ and a set of matrices $\mathcal{M} \subseteq \mathbb{R}^{d_1\times d_2}$, let $\text{vec}(M) \in \mathbb{R}^{d_1d_2}$ be the vectorization of the matrix M by vertically stacking its columns and $\text{vec}(\mathcal{M}) := \{\text{vec}(M) : M \in \mathcal{M}\}$. Denote by reshape$(\cdot)$ the inverse map of vec$(\cdot)$; i.e., reshape$(v) = M$ if and only if $\text{vec}(M) = v$. We assume that $\mathcal{A}$ spans $\mathbb{R}^{d_1\times d_2}$. Define $d = \max(d_1, d_2)$. We denote by v_i the i-th component of the vector v and by M_{ij} the entry of a matrix M located at the i-th row and j-th column. Let $\lambda_k(M)$ be the k-th largest singular value, and define $\lambda_{\max}(M) = \lambda_1(M)$, which is also known as $\|M\|_{\text{op}}$, the operator norm of M. Let $\lambda_{\min}(M)$ be the smallest nonzero

[2]Our bound here is a $d^{1/4}$ factor larger from the original paper since our setting is operator norm bounded action set, which is different from their Frobenius norm bounded action set. For details, see Appendix H.3.

singular value of M. Let $\|M\|_F = \sqrt{\sum_{i=1}^{d_1}\sum_{j=1}^{d_2} M_{ij}^2}$ and $\|M\|_* = \sum_{i=1}^{\min(d_1,d_2)} \lambda_i(M)$ be the Frobenius norm of M and nuclear norm, respectively. $\tilde{O}$ is the order notation that hides logarithmic factors. For any set S, let $\mathcal{P}(S)$ be the set of probability distributions on S. For any $\pi \in \mathcal{P}(\mathcal{A})$, define the population covariance matrix of the vectorized matrix $Q(\pi) \in \mathbb{R}^{d_1 d_2 \times d_1 d_2}$ as follows:

$$Q(\pi) = \mathbb{E}_{a\sim\pi}\left[\text{vec}(a)\text{vec}(a)^\top\right] \tag{1}$$

We define $\mathcal{B}_{\text{op}}(R) := \{a \in \mathbb{R}^{d_1\times d_2} : \|a\|_{\text{op}} \le R\}$.

Low-rank bandits. Throughout, we assume that the learning agent interacts with the environment in the following manner. At every time step $t \in \{1,\dots,T\}$, the learner chooses an arm A_t from the arm set $\mathcal{A} \subset \mathbb{R}^{d_1\times d_2}$ and receives reward $y_t = \langle \Theta^*, A_t\rangle + \eta_t$, where Θ^* is an unknown matrix with a known upper bound of the rank at most $r \ll \min(d_1,d_2)$. η_t is an independent zero-mean σ-subgaussian noise, and the inner product of two matrices are defined as $\langle A, B\rangle = \langle \text{vec}(A), \text{vec}(B)\rangle = \text{tr}(A^\top B)$. The goal of the learner is to minimize its (pseudo-)regret:

$$\text{Reg}(T) := T\max_{A\in\mathcal{A}}\langle \Theta^*, A\rangle - \sum_{t=1}^{T}\langle\Theta^*, A_t\rangle.$$

The following matrix generalization of Catoni's robust mean estimator proposed by (Minsker, 2018) will be useful for our novel estimator.

Definition 2.1. Given a symmetric matrix M with its eigenvalue decomposition $M = U\Lambda U^\top$ where $\Lambda = \text{diag}(\lambda_1,\cdots,\lambda_d)$, we first define $\phi_0 : \mathbb{R}\to\mathbb{R}$ as

$$\phi_0(x) = \begin{cases} \log(1+x+\frac{x^2}{2}) & \text{if } x>0 \\ -\log(1-x+\frac{x^2}{2}) & \text{otherwise}\end{cases}$$

and $\phi : \mathbb{R}^{d\times d}\to\mathbb{R}^{d\times d}$ as

$$\phi(M) = U\left[\text{diag}(\phi_0(\lambda_1),\phi_0(\lambda_2),\cdots,\phi_0(\lambda_d))\right]U^\top$$

Finally, for any matrix $A\in\mathbb{R}^{d_1\times d_2}$, define the dilation operator $\mathcal{H} : \mathbb{R}^{d_1\times d_2}\to\mathbb{R}^{(d_1+d_2)\times(d_1+d_2)}$ as

$$\mathcal{H}(A) = \begin{bmatrix} 0_{d_1\times d_1} & A \\ A^\top & 0_{d_2\times d_2}\end{bmatrix}.$$

Dilation is a common trick to allow existing estimation tools built for real symmetric matrices to work on rectangular matrices, as in (Huang et al., 2021; Minsker, 2018). For a dilated matrix $M\in\mathbb{R}^{(d_1+d_2)\times(d_1+d_2)}$, $(M)_{\text{ht}}$ refers to the shorthand of $M_{1:d_1, d_1+1:d_1+d_2}$.

3. LowPopArt: A novel low-rank matrix estimator

In this section, we will present our novel low-rank matrix estimation algorithm, LOW-rank Population Covariance regression with hARd Thresholding (LowPopArt; Algorithm 1), which is inspired by a recent sparse linear estimation algorithm called PopArt (Jang et al., 2022). We discuss the differences between LowPopArt and PopArt in detail at the end of this section.

LowPopArt takes samples $\{X_i, Y_i\}_{i=1}^{n_0}$, sample size n_0, the population covariance matrix of the vectorized matrix $Q(\pi)$, pilot estimator Θ_0 and pilot estimation error bound R_0 s.t. $\max_{A\in\mathcal{A}}\left|\langle\Theta_0-\Theta, A\rangle\right| \le R_0$ as its input. It consists of three stages. In the first stage, PopArt creates a collection of one-sample estimator $\{\tilde{\Theta}_i\}_{i=1}^{n_0}$ from the input data $\{(X_i,Y_i)\}_{i=1}^{n_0}$ as follows:

$$\tilde{\Theta}_i := Q(\pi)^{-1}(Y_i - \langle\Theta_0, X_i\rangle)\text{vec}(X_i) \tag{2}$$

Note that each $\tilde{\Theta}_i$ is an unbiased estimator of $\text{vec}(\Theta^*-\Theta_0)$.

Naively, one could use the average $\bar{\Theta} := \frac{1}{n_0}\sum_{i=1}^{n_0}\tilde{\Theta}_i$ as an estimator for $\Theta^*-\Theta_0$. When the number of samples is large enough, the empirical covariance matrix $\tilde{Q} = \frac{1}{n_0}\sum_{i=1}^{n_0}\text{vec}(X_t)\text{vec}(X_t)^\top$ is close to $Q(\pi)$, which makes $\bar{\Theta}$ close to the d_1d_2-dimensional ordinary least squares (OLS) estimator. However, it is not easy to control the tail behavior of $\bar{\Theta}$, and consequently it is hard to exploit the low-rank property when one naively uses $\bar{\Theta}$. Instead, we use the estimator of Minsker (2018, Corollary 3.1) which symmetrizes the original matrix and computes the Catoni function for each eigenvalue (Definition 2.1), which has the effect of lightening the tail distribution of singular values. We call the resulting matrix Θ_1. Finally, we run SVD on Θ_1 and zero out all the singular values smaller than the threshold, to exploit the knowledge that Θ^* is low-rank.

Remark 1. In the general estimation problem, we do not have prior knowledge of the inverse covariance matrix of the data, but one may attempt to estimate it if having sample access to the covariate distribution; e.g., matrix geometric sampling (Neu & Olkhovskaya, 2020). On the other hand, there are some problems (such as bandits or compressed sensing) where the agent has full control over the distribution of the dataset. In these cases, LowPopArt can be directly applied. Obtaining a precise performance guarantee when the covariance matrix is estimated from the observed samples is left as future work.

Analysis of Algorithm 1 We start by stating the following recovery guarantee of the estimator Θ_1. Detailed proofs of this part are mainly in Appendix B.

Theorem 3.1. *Suppose we run Algorithm 1 with the arm set $\mathcal{A}$, sample size n_0, population covariance matrix of vectorized matrices Q, pilot estimator Θ_0 and pilot estimation error bound R_0, such that $\max_{A\in\mathcal{A}}\left|\langle\Theta_0-\Theta^*, A\rangle\right| \le R_0$, then Θ_1 satisfies the following error bound with probability at least $1-\delta$:*

$$\|\Theta_1-\Theta^*\|_{\text{op}} \le O\left((\sigma+R_0)\sqrt{\frac{B(Q)}{n_0}\ln\frac{2d}{\delta}}\right). \tag{3}$$

Algorithm 1 LowPopArt

Input: Samples $\{X_i, Y_i\}_{i=1}^{n_0}$, sample size n_0, the population covariance matrix of the vectorized matrix $Q(\pi)$, pilot estimator Θ_0 and pilot estimation error bound R_0.

`Step 1: Compute one-sample estimators.`

for $t = 1, \dots, n_0$ **do**

  Compute $\tilde{\Theta}_i$ as in Eq. (2).

end for

`Step 2: Compute the matrix Catoni estimator (Minsker, 2018) using` $\{\tilde{\Theta}_i\}_{i=1}^{n_0}$

Compute:

$$\Theta_1 = \Theta_0 + \left(\frac{1}{n_0 \nu} \sum_{i=1}^{n_0} \psi\left(\nu \mathcal{H}\left(\text{reshape}\left(\tilde{\Theta}_i\right)\right)\right)\right)_{\text{ht}}$$

where $\nu = \frac{1}{\sigma + R_0} \sqrt{\frac{2}{B(Q) n_0} \ln \frac{2d}{\delta}}$.

`Step 3: Hard-thresholding eigenvalues.`

Let $U_1 \Sigma_1 V_1^\top$ be Θ_1's SVD. Let $\tilde{\Sigma}_1$ be a modification of Σ that zeros out its diagonal entries that are at most $\lambda_{\text{th}} := 2(R_0 + \sigma)\sqrt{\frac{\left(B(Q) \ln \frac{2d}{\delta}\right)}{n_0}}$ where $B(Q)$ is in Eq. (4).

Return: Estimator $\hat{\Theta} = U_1 \tilde{\Sigma}_1 V_1^\top$.

where

$$B(Q) := \max\left(\lambda_{\max}\left(\sum_{i=1}^{d_2} D_i^{(\text{col})}\right), \lambda_{\max}\left(\sum_{i=1}^{d_1} D_i^{(\text{row})}\right)\right) \quad (4)$$

where $D_i^{(\text{col})} = (Q^{-1})_{[i \cdot d_s + 1 : (i+1) \cdot d_s], [i \cdot d_s + 1 : (i+1) \cdot d_s]}$ *and* $D_i^{(\text{row})} := [(Q^{-1})_{jk}]_{j,k \in \{i + d_1(\ell - 1) : \ell \in [d_2]\}}$; *see Figure 1 for illustrations.*

Remark 2. The intuition underlying $B(Q)$ is as follows. When $d = 1$, $B(Q)$ is proportional to the variance of $\tilde{\Theta}_1$; for $d \geq 1$, $B(Q)$ is, informally, proportional to the largest variance of $\tilde{\Theta}_1$ projected onto rank-1 dyads $\{uv^\top : u, v \in \mathbb{S}^{d-1}\}$.

From the above Theorem 3.1, one could deduce the final operator norm bound of the output $\hat{\Theta}$.

Theorem 3.2. *Under the same assumption in Theorem 3.1,* $\text{rank}(\hat{\Theta}) \leq r$*, and the following operator norm bounds hold with probability at least* $1 - \delta$*:*

$$\|\hat{\Theta} - \Theta^*\|_{\text{op}} \leq O\left((\sigma + R_0)\sqrt{\frac{B(Q)}{n_0} \ln \frac{2d}{\delta}}\right) \quad (5)$$

Theorem 3.2 implies the following error bounds in nuclear norm and Frobenius norm recovery errors:

Corollary 3.3. *Under the same assumption as in Theorem 3.1, the following nuclear norm and Frobenius norm bounds hold with probabiliy at least* $1 - \delta$*:*

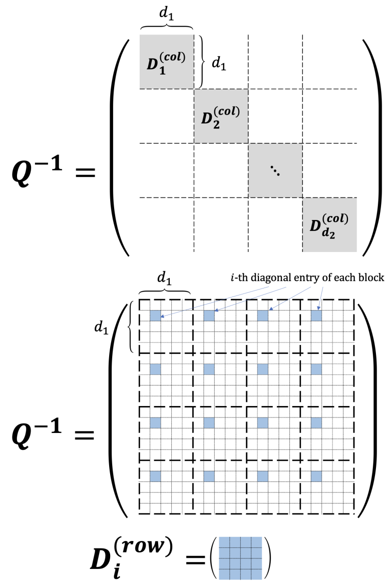

Figure 1. Illustration of $D_i^{(\text{col})}$ and $D_i^{(\text{row})}$

$$\|\hat{\Theta} - \Theta^*\|_* \leq O\left((\sigma + R_0)\sqrt{\frac{r^2 B(Q)}{n_0} \ln \frac{2d}{\delta}}\right) \quad (6)$$

$$\|\hat{\Theta} - \Theta^*\|_F \leq O\left((\sigma + R_0)\sqrt{\frac{r B(Q)}{n_0} \ln \frac{2d}{\delta}}\right) \quad (7)$$

A naive application of LowPopArt with pilot estimator $0_{d_1 \times d_2}$ gives an estimator $\hat{\Theta}$ such that $\|\hat{\Theta} - \Theta^*\|_{\text{op}} \leq \tilde{O}\left((\sigma + S_*)\sqrt{\frac{B(Q)}{n_0}}\right)$; the dependence on S_* is somewhat undesirable when $S_* \gg \sigma$. Motivated by this, we propose an improved version of LowPopArt whose estimation error guarantee is $\tilde{O}\left(\sigma\sqrt{\frac{B(Q)}{n_0}}\right)$ under mild assumptions, i.e. Warm-LowPopArt (Algorithm 2). Its key idea is to first use LowPopArt to construct a coarse estimator Θ_0 such that $\|\Theta_0 - \Theta^*\|_* \leq \sigma$, which ensures that $\max_{A \in \mathcal{A}} |\langle \Theta_0 - \Theta^*, A \rangle| \leq \sigma$; it subsequently calls LowPopArt again with Θ_0 as a pilot estimator, to obtain the final estimate $\hat{\Theta}$. Formally, we have the following theorem:

Theorem 3.4. *Suppose that Algorithm 2 is run with arm set* $\mathcal{A}$*, sample size* n_0*, failure rate* δ*, such that* $\max_{A \in \mathcal{A}} |\langle \Theta_0 - \Theta, A \rangle| \leq R_0$*, and* $n_0 \geq \tilde{O}\left(r^2 B(Q) \cdot (\frac{\sigma + S_*}{\sigma})^2\right)$*, then its output* $\hat{\Theta}$ *is such that rank*$(\hat{\Theta}) \leq r$*, and:*

$$\|\hat{\Theta} - \Theta^*\|_{\text{op}} \leq O\left(\sigma\sqrt{\frac{2B(Q)}{n_0} \ln \frac{2d}{\delta}}\right). \quad (8)$$

Comparison with nuclear norm penalty methods An alternative and popular approach for matrix estimation is nuclear norm penalized least squares (Koltchinskii et al., 2011), which yields a recovery guarantee of $\|\hat{\Theta} - \Theta^*\|_F \leq$

Algorithm 2 Warm-LowPopArt: a bootstrapped version of LowPopArt

1: **Input:** Samples $\{X_i, Y_i\}_{i=1}^{n_0}$, sample size n_0, population covariance matrix of the vectorized matrix Q, failure rate δ.
2: $\Theta_0 \leftarrow \text{LowPopArt}(\{X_i, Y_i\}_{i=1}^{\frac{n_0}{2}}, n_0/2, Q, 0_{d_1\times d_2}, S_*, \delta/2)$
3: $\hat{\Theta} \leftarrow \text{LowPopArt}(\{X_i, Y_i\}_{i=\frac{n_0}{2}+1}^{n_0}, n_0/2, Q, \Theta_0, \sigma, \delta/2)$
4: **Return:** $\hat{\Theta}$

$\tilde{O}(\sqrt{\frac{r}{n\lambda_{\min}(Q)^2}})$ and $\|\hat{\Theta}-\Theta^*\|_* \le \tilde{O}(\sqrt{\frac{r^2}{n\lambda_{\min}(Q)^2}})$. We show in Appendix I that under Assumption A1, $\lambda_{\min}(Q) \le \frac{1}{d}$, and by Lemma 3.5 below, our error bound of LowPopArt is always tighter than that of (Koltchinskii et al., 2011).

Lemma 3.5. $B(Q) \le \frac{d}{\lambda_{\min}(Q)}$

Thus, $B(Q)$ can be viewed as a tighter measurement-distribution-dependent quantity that characterizes the hardness of the low-rank matrix recovery,

However, we can go even further – it is a natural question to consider how much the recovery error can be reduced when applying the best experimental design tailored to each estimation method.

Experimental design As can be seen from Theorem 3.2, the recovery guarantee of the LowPopArt algorithm depends on the hardness $B(Q)$. Therefore, if the agent can design the sampling distribution over the given measurement set $\mathcal{A}$, a natural choice would be one that minimizes the $B(Q)$ value. Formally, we define the optimal $B(Q)$ as:

$$B_{\min}(\mathcal{A}) := \min_{\pi\in\mathcal{P}(\mathcal{A})} B(Q(\pi)) \quad (9)$$

where $Q(\pi)$ is defined in Eq. (1).

Intuitively, this quantity can be understood as a single metric capturing the geometry of the measurement set. This optimization problem is convex and can be efficiently computed using common convex optimization tools such as cvxpy (Diamond & Boyd, 2016).

Research on the experimental design for low-rank matrix estimation is surprisingly scarce. One reasonable comparison point for our experimental design is the classical E-optimal design (Lattimore & Hao, 2021; Hao et al., 2020; Soare et al., 2014), well-known in experimental design for linear regression. E-optimality aims to maximize the minimum eigenvalue of the sampling distribution's covariance matrix, with optimal objective value formally defined as follows:

$$C_{\min}(\mathcal{A}) = \max_{\pi\in\mathcal{P}(\mathcal{A})} \lambda_{\min}(Q(\pi)) \quad (10)$$

Now, the important question is how the recovery bounds of LowPopArt and nuclear norm penalized least squares differ when written in terms of $C_{\min}(\mathcal{A})$ and $B_{\min}(\mathcal{A})$, respectively. We have established the following results between $C_{\min}(\mathcal{A})$ and $B_{\min}(\mathcal{A})$:

Lemma 3.6. *Suppose Assumption A1 holds. Then $d^2 \le B_{\min}(\mathcal{A}) \le \frac{d}{C_{\min}}$, and there exists an arm set $\mathcal{A}_{\textsf{hard}}$ for which $B_{\min}(\mathcal{A}_{\textsf{hard}}) \approx \frac{1}{C_{\min}}$.*

See Appendix C for the proof of Lemma 3.5, 3.6 and the construction of $\mathcal{A}_{\textsf{hard}}$. For the arm set $\mathcal{A}_{\textsf{hard}}$ our guarantee is $\frac{1}{d^{3/2}}$ times tighter than the guarantee of (Koltchinskii et al., 2011), which shows the importance of using the right arm set geometry quantity.

Main novelty of LowPopArt compared to PopArt (Jang et al., 2022). The major challenge is the absence of the knowledge of a well-structured basis that the agent could exploit a low-rank property of Θ^* to do better estimation. In sparse linear bandits, the basis for testing the zeroness is known to the agent (i.e. the canonical basis), so the estimation procedure can simply focus on controlling the estimation error over the d coordinates. On the other hand, in low-rank bandits, we need to control the subspace estimation error, but the potential number of subspace directions (i.e., $\mathcal{F} = \{uv^\top : u \in \mathbb{S}^{d_1-1}, v \in \mathbb{S}^{d_2-1}\}$ or its ε-net) is infinite or exponentially large ($\sim \exp(d_1+d_2)$). Indeed, one of the naive extension of (Jang et al., 2022) for estimation, which considers all possible directions in an ε-net of $\mathcal{F}$. However, this causes computational intractability. To get around this issue, we propose to directly upper bound $\|\hat{\Theta}-\Theta^*\|_{\text{op}}$ for establishing Frobenius and nuclear norm recovery error guarantees, which can be performed via the method of (Minsker, 2018) in a computationally efficient manner. This was the key observation that led to our main result. We reiterate that our key contribution is to identify a novel subspace exploration hardness measure of a *sampling distribution*. If one has control of the sampling distribution (e.g., in bandits), then this measure (i.e., our experimental design criterion) can be efficiently minimized to represent the subspace exploration hardness of the *measurement set*, which, as we proved, is strictly better than a naive attempt like E-optimality (Eq. (10)). *We believe this is a significant result given that there were no prior studies on experimental design in the low-rank matrix estimation setting.*

4. Low rank bandit algorithms

We now leverage LowPopArt to design two computationally efficient algorithms for low-rank bandits.

Explore-then-commit based algorithm. Algorithm 3 is based on the well-known Explore-then-Commit framework. It uses Warm-LowPopArt as its exploration method to obtain $\hat{\Theta}$, an estimate of Θ^*, and subsequently takes the greedy arm with respect to $\hat{\Theta}$.

We prove the following regret guarantee:

Algorithm 3 LPA-ETC (LowPopArt based Explore then commit)

Input: time horizon T, arm set $\mathcal{A}$, exploration lengths n_0, regularization parameter ν, pilot estimator Θ_0
Solve the optimization problem in Eq. (9) and denote the solution as π^*
for $t = 1, \dots, n_0$ **do**
  Independently pull the arm A_t according to π^* and receives the reward Y_t
end for
Run Warm-LowPopArt$(\{A_i, Y_i\}_{i=1}^{n_0}, n_0, Q(\pi^*), \delta)$ and get $\hat{\Theta}$
for $t = n_0 + 1, \dots, T$ **do**
  Pull the arm $A_t = \arg\max_{A \in \mathcal{A}} \langle \hat{\Theta}, A \rangle$
end for

Theorem 4.1 (Regret upper bound)**.** *Suppose $T \geq rB_{\min}(\mathcal{A})(\frac{\sigma+S_*}{\sigma})^4$. The regret upper bound of Alg. 3 with $n_0 = \min(T, \left(\sigma^2 r^2 B_{\min}(\mathcal{A}) T^2 / S_*^2\right)^{1/3})$ is as follows:*

$$\mathsf{Reg}(T) \leq \tilde{O}((\sigma^2 S_* r^2 T^2 B_{\min}(\mathcal{A}))^{1/3}) \qquad (11)$$

Remark 3. To the best of our knowledge, the only algorithms that can handle general arm sets with $\lambda_{\min}(\Theta^*)$-free regret bounds are LowLOC (Lu et al., 2021) and rO-UCB (Jang et al., 2021). Both algorithms have regret bounds of $O(\sigma r^{1/2} d^{3/2} \sqrt{T})$ but are not computationally tractable. On the other hand, our ETC-based algorithm is computationally efficient and achieves a better regret bound when $T \leq O\left(\sigma^2 d^9 S_*^{-2} B_{\min}(\mathcal{A})^{-2} r^{-1}\right)$.

Explore-Subspace-Then-Refine (ESTR) based algorithm. Although general, Algorithm 3 overlooks a favorable structure underlying many low-rank bandit problems: Θ^* is well-conditioned in many settings, e.g. $\lambda_{\min} \geq \Omega(S_*/r)$. Such structure has been exploited by many prior works (Jun et al., 2019; Lu et al., 2021; Kang et al., 2022) to design $\sqrt{T}$-regret algorithms. In this part, we assume that $\lambda_{\min}(\Theta^*) \geq S_r$ for some known $S_r > 0$.

In this section, we use Warm-LowPopArt to design an efficient algorithm with $O(\sqrt{T})$ regret (Algorithm 4). Algorithm 4 is based on the Explore-Subspace-Then-Refine (ESTR) framework (Jun et al., 2019). In ESTR, we use Warm-LowPopArt to find an estimate $\hat{\Theta}$ such that it closely approximates Θ in operator norm. We then estimate the row and column spaces of Θ using an SVD over $\hat{\Theta}$, represented by their orthonormal bases $\hat{U}$ and $\hat{V}$. Then, we rotate the arm set using $\hat{U}$ and $\hat{V}$. After this transformation, the original linear bandit problem becomes a $d_1 d_2$-dimensional linear bandit problem with arm set $\mathcal{A}'$ and reward predictor

$$\theta^* = (\operatorname{vec}(\hat{U}^\top \Theta^* \hat{V}); \operatorname{vec}(\hat{U}_\perp^\top \Theta^* \hat{V}); \operatorname{vec}(\hat{U}^\top \Theta^* \hat{V}_\perp); \operatorname{vec}(\hat{U}_\perp^\top \Theta^* \hat{V}_\perp))$$

Algorithm 4 LPA-ESTR (LowPopArt based Explore Subspace Then Refine)

Input: time horizon T, arm set $\mathcal{A}$, exploration lengths n_0, singular value lower bound S_r
Solve the optimization problem in Eq. (9) and denote the solution as π
for $t = 1, \dots, n_0$ **do**
  Independently pull the arm A_t according to π and receives the reward Y_t
end for
Run Warm-LowPopArt$(\{A_i, Y_i\}_{i=1}^{n_0}, n_0, Q(\pi), \delta)$ and get $\hat{\Theta}$ with SVD result $\hat{\Theta} = \hat{U}\hat{\Sigma}\hat{V}^\top$.
Let $\hat{U}_\perp$ and $\hat{V}_\perp$ be the orthonormal bases of the orthogonal complement subspaces of $\hat{U}$ and $\hat{V}$, respectively.
Rotate whole arm feature set $\mathcal{A}' := \{[\hat{U}\ \hat{U}_\perp] A [\hat{V}\ \hat{V}_\perp]^\top : A \in \mathcal{A}\}$
Define a vectorized arm feature set so that the last $(d_1 - r)(d_2 - r)$components are from the complementary subspaces:

$$\mathcal{A}'_{vec} := \{(\operatorname{vec}(A'_{1:r,1:r}); \operatorname{vec}(A'_{r+1:d_1,1:r}); \operatorname{vec}(A'_{1:r,r+1:d_2}); \operatorname{vec}(A'_{r+1:d_1,r+1:d_2})) : A' \in \mathcal{A}'\}$$

Invoke LowOFUL with time horizon $T - n_0$, arm set $\mathcal{A}'_{\text{vec}}$, the low dimension $k = r(d_1 + d_2 - r)$, $\lambda = \frac{\sigma^2}{S_*^2} dr$, $\lambda_\perp = \frac{T}{r}$, $B = S_*$, and $B_\perp = \frac{B_{\min}(\mathcal{A})\sigma^2 S_*}{n_0 S_r^2}$.

Crucially, by the recovery guarantee of Warm-LowPopArt and Wedin's Theorem (Stewart & Sun, 1990), $\|\hat{U}_\perp^\top U\|_{\text{op}}$ and $\|\hat{V}_\perp^\top V\|_{\text{op}}$ are both small; as a consequence, $\|\theta^*_{r(d_1+d_2-r)+1:d_1 d_2}\|_2 = \|\operatorname{vec}(\hat{U}_\perp^\top \Theta^* \hat{V}_\perp)\|_F \leq \|\hat{U}_\perp^\top U\|_{\text{op}} \|\Theta^*\|_F \|\hat{V}_\perp^\top V\|_{\text{op}}$, which is also small. In other words, we are now faced with a linear bandit problem with the prior knowledge that a large subset of the coordinates of the reward predictor is small.

This motivates the usage of the LowOFUL algorithm (Jun et al., 2019)[3] in the second stage, which is a modification of OFUL (Abbasi-Yadkori et al., 2011) with heavy penalizations on the reward predictor on insignificant coordinates. Theorem 4.2 states the overall regret upper bound of Algorithm 4.

Theorem 4.2. *Suppose $T \geq \frac{16 B_{\min}(\mathcal{A})\sigma^4}{d^{0.5} S_r (\Theta^*)^2}$. The regret upper bound of Algorithm 4 with $n_0 = \sqrt{\frac{d^{0.5} B_{\min}(\mathcal{A})}{S_r^2}} T$ is*

$$\mathsf{Reg}(T) \leq O\left(\sigma \sqrt{\frac{S_*^2}{S_r^2} B_{\min}(\mathcal{A}) d^{0.5} T}\right)$$

with probability at least $1 - 2\delta$.

Algorithm 4 attains a $\sqrt{T}$-order regret bound, at the cost of introducing a dependence of S_r factor in the regret bound.

[3] Pseudocode of LowOFUL is in Appendix F.1, Algorithm 5.

Remark 4. When Θ^* is well conditioned, i.e. $S_r \geq \Omega(S_*/r)$, the above regret bound can be simplified to $O(\sigma\sqrt{r^2 d^{0.5} B_{\min}(\mathcal{A}) T})$. For the case where $\mathcal{A} = \mathcal{B}_{\text{op}}(1)$, we can prove $B_{\min}(\mathcal{A}) \leq d^2$, and we have the upper bound of order $\tilde{O}(\sqrt{r^2 d^{2.5} T})$ when Θ^* is well-conditioned, which is an improved result compared to $\sqrt{r^3 d^{2.5} T}$ of Lu et al. (2021) and even to the computationally inefficient result $\sqrt{rd^3 T}$ of Lu et al. (2021). Plus, our algorithm is strictly better than LowESTR (Lu et al., 2021) in any cases because $B_{\min}(\mathcal{A}) \leq \frac{1}{\lambda_{\min}(Q(\pi))^2}, \forall \pi \in \mathcal{P}(\mathcal{A})$ by Lemma 3.6.

Remark 5. In addition to arm set dependent constant, LPA-ESTR also achieves an improved regret guarantee over LowESTR (Lu et al., 2021) w.r.t. r. This is because our LowPopArt estimator provides improved bounds on $\|\hat{U}_\perp^\top U\|_{\text{op}}$ and $\|\hat{V}_\perp^\top V\|_{\text{op}}$, which are a factor of $\sqrt{r}$ lower than their respective bounds in (Lu et al., 2021). This is enabled by the unique operator-norm based recovery guarantee of LowPopArt and the operator norm-version of Wedin's Theorem; to the best of our knowledge, we are not aware of an operator-norm-based recovery guarantee for nuclear norm penalized least squares regression.

5. Experiments

We now evaluate the empirical performance of LowPopArt and our proposed experimental design to validate our improvement. We defer unimportant details of the experimental setup in Appendix J.

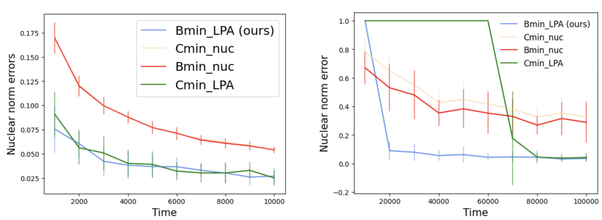

Figure 2. Experiment results on nuclear norm error

Low-rank matrix recovery. Figure 2 presents the results on the nuclear norm recovery error (y-axis) as a function of the sample size (x-axis). The prefix of each line (Cmin, Bmin) represents the experimental design for the sampling distribution (optimal solutions of Eq. (10) and Eq. (9), respectively). The suffix (LPA, nuc) indicates the estimation method employed (LowPopArt and nuclear norm regularized least squares, respectively.) In the left plot, we draw arm matrices uniformly at random from $\mathcal{B}_{\text{Frob}}(1)$. In the right figure, we consider the arm set $\mathcal{A}_{\text{hard}}$ from Lemma 3.6 that has a significant disparity between $B_{\min}(\mathcal{A})$ and $C_{\min}(\mathcal{A})$ values (see Appendix C for the definition).

As one can see in the above figures, in all cases, $B_{\min}(\mathcal{A})$ based exploration generally outperforms naive E-optimal design, and LowPopArt tends to show a better nuclear norm recovery error than nuc.

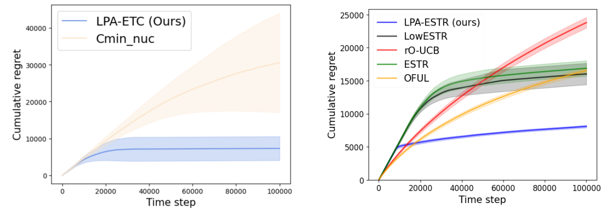

Figure 3. Experiment results on bandits with ETC-based (left) and ESTR-based algorithms (right)

Low-rank matrix bandits. Figure 3 presents the results of applying LowPopArt-based algorithms (Algorithm 3 and 4) to the low-rank bandit problem. The first graph (left) compares Algorithm 3 with another ETC-based algorithm, which is based on nuclear norm regularized least squares. Please check Appendix J for the pseudocode of this algorithm. Algorithm 3 achieves a significantly lower regret with a much shorter exploration length, demonstrating more stable results than nuclear norm regularization.

The second graph (right) compares our Algorithm 4 with state-of-the-art algorithms based on OFUL, such as ESTR (Jun et al., 2019), rO-UCB (Jang et al., 2021), LowESTR(Lu et al., 2021), and OFUL on the flattened $d_1 d_2$-dimensional linear bandit problem itself (Abbasi-Yadkori et al., 2011). Once again, it is apparent that our LPA-ESTR (Algorithm 4) outperforms other OFUL based algorithms, showing lower and more stable cumulative regret.

6. Conclusion

We have proposed a novel low-rank estimation algorithm called LowPopArt, along with a novel experimental design that aims at minimizing LowPopArt's recovery guarantees. This new algorithm utilizes the geometry of the arm set to conduct estimation in a different manner than conventional approaches. Based on LowPopArt, we have designed two low-rank bandit algorithms with general arm sets, improving the dimensionality dependence in regret bounds.

Although general, one drawback of our algorithms is that, when applied to special arm sets (e.g. the unit Frobenius norm ball), its guarantees are inferior than algorithms designed specifically for these settings (Lattimore & Hao, 2021; Huang et al., 2021). Designing algorithms that can match these guarantees in these specialized settings while maintaining generality is an interesting future direction. Another interesting open question is establishing regret lower bound that depends on the geometry of the arm set in the low-rank bandit problem.

7. 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., Pal, D., and Szepesvari, C. Improved Algorithms for Linear Stochastic Bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1–19, 2011.

Abe, N. and Long, P. M. Associative reinforcement learning using linear probabilistic concepts. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 3–11, 1999.

Auer, P. Using Confidence Bounds for Exploitation-Exploration Trade-offs. *Journal of Machine Learning Research*, 3:397–422, 2002.

Bandeira, A. S., Dobriban, E., Mixon, D. G., and Sawin, W. F. Certifying the restricted isometry property is hard. *IEEE Transactions on Information Theory*, 59(6):3448–3450, 2013.

Bennett, J., Lanning, S., et al. The netflix prize. In *Proceedings of KDD cup and workshop*, volume 2007, pp. 35. New York, 2007.

Camilleri, R., Jamieson, K., and Katz-Samuels, J. High-dimensional experimental design and kernel bandits. In *International Conference on Machine Learning*, pp. 1227–1237. PMLR, 2021.

Dani, V., Hayes, T. P., and Kakade, S. M. Stochastic Linear Optimization under Bandit Feedback. In *Proceedings of the Conference on Learning Theory (COLT)*, pp. 355–366, 2008.

Diamond, S. and Boyd, S. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

Eftekhari, H., Banerjee, M., and Ritov, Y. Design of c-optimal experiments for high-dimensional linear models. *Bernoulli*, 29(1):652–668, 2023.

Gales, S. B., Sethuraman, S., and Jun, K.-S. Norm-agnostic linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 73–91. PMLR, 2022.

Hamdi, N. and Bayati, M. On low-rank trace regression under general sampling distribution. *The Journal of Machine Learning Research*, 23(1):14424–14472, 2022.

Hamidi, N. and Bayati, M. On worst-case regret of linear thompson sampling. *arXiv preprint arXiv:2006.06790*, 2020.

Hao, B., Lattimore, T., and Wang, M. High-dimensional sparse linear bandits. *Advances in Neural Information Processing Systems*, 33:10753–10763, 2020.

Hardt, M. and Price, E. The noisy power method: A meta algorithm with applications. *Advances in neural information processing systems*, 27, 2014.

Horn, R. A. and Johnson, C. R. *Matrix analysis*. Cambridge university press, 2012.

Huang, B., Huang, K., Kakade, S., Lee, J. D., Lei, Q., Wang, R., and Yang, J. Optimal gradient-based algorithms for non-concave bandit optimization. *Advances in Neural Information Processing Systems*, 34:29101–29115, 2021.

Huang, R., Lattimore, T., György, A., and Szepesvári, C. Following the leader and fast rates in linear prediction: Curved constraint sets and other regularities. *Advances in Neural Information Processing Systems*, 29, 2016.

Jain, P. and Dhillon, I. S. Provable inductive matrix completion. *arXiv preprint arXiv:1306.0626*, 2013.

Jang, K., Jun Kwang-Sung, Y. S. Y., and Kang, W. Improved Regret Bounds of Bilinear Bandits using Action Space Dimension Analysis. In *Proceedings of the International Conference on Machine Learning (ICML), accepted*, pp. 3163–3172, 2021.

Jang, K., Zhang, C., and Jun, K.-S. Popart: Efficient sparse regression and experimental design for optimal sparse linear bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2102–2114. Curran Associates, Inc., 2022.

Juditsky, A. and Nemirovski, A. On verifiable sufficient conditions for sparse signal recovery via l_1 minimization. *Mathematical programming*, 127:57–88, 2011.

Jun, K.-S., Willett, R., Wright, S., and Nowak, R. Bilinear Bandits with Low-rank Structure. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 97, pp. 3163–3172, 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.

Katariya, S., Kveton, B., Szepesvári, C., Vernade, C., and Wen, Z. Bernoulli Rank-1 Bandits for Click Feedback. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 2001–2007, 2017.

Keshavan, R. H., Montanari, A., and Oh, S. Matrix Completion from Noisy Entries. *J. Mach. Learn. Res.*, 11: 2057–2078, 2010. ISSN 1532-4435.

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

Kotlowski, W. and Neu, G. Bandit Principal Component Analysis. In Beygelzimer, A. and Hsu, D. (eds.), *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 1994–2024, Phoenix, USA, 2019. PMLR.

Kveton, B., Szepesvári, C., Rao, A., Wen, Z., Abbasi-Yadkori, Y., and Muthukrishnan, S. Stochastic Low-Rank Bandits. *arXiv:1712.04644*, 2017.

Lattimore, T. and Hao, B. Bandit Phase Retrieval, 2021.

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

Li, W., Barik, A., and Honorio, J. A simple unified framework for high dimensional bandit problems. In *International Conference on Machine Learning*, pp. 12619–12655. PMLR, 2022.

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.

Luo, Y., Zhao, X., Zhou, J., Yang, J., Zhang, Y., Kuang, W., Peng, J., Chen, L., and Zeng, J. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nature communications*, 8(1):573, 2017.

Mason, B., Camilleri, R., Mukherjee, S., Jamieson, K., Nowak, R., and Jain, L. Nearly optimal algorithms for level set estimation. *arXiv preprint arXiv:2111.01768*, 2021.

Minsker, S. Sub-gaussian estimators of the mean of a random matrix with heavy-tailed entries. *The Annals of Statistics*, 46(6A):2871–2903, 2018.

Natarajan, N. and Dhillon, I. S. Inductive matrix completion for predicting gene–disease associations. *Bioinformatics*, 30(12):i60–i68, 2014.

Neu, G. and Olkhovskaya, J. Efficient and robust algorithms for adversarial linear contextual bandits. In *Conference on Learning Theory*, pp. 3049–3068. PMLR, 2020.

Rohde, A. and Tsybakov, A. B. Estimation of high-dimensional low-rank matrices. 2011.

Rusmevichientong, P. and Tsitsiklis, J. N. Linearly Parameterized Bandits. *Math. Oper. Res.*, 35(2):395–411, 2010.

Soare, M., Lazaric, A., and Munos, R. Best-arm identification in linear bandits. *Advances in Neural Information Processing Systems (NeurIPS)*, 27:828–836, 2014.

Stewart, G. W. and Sun, J.-g. *Matrix perturbation theory*. Academic press, 1990.

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.

Udell, M., Horn, C., Zadeh, R., Boyd, S., et al. Generalized low rank models. *Foundations and Trends® in Machine Learning*, 9(1):1–118, 2016.

Valko, M., Munos, R., Kveton, B., and Kocak, T. Spectral Bandits for Smooth Graph Functions. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 46—-54, 2014.

Wainwright, M. J. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. doi: 10.1017/9781108627771.

Appendix

Table of Contents

A. Additional Related Work

Low-rank bandits with general arm sets The first low-rank bandit algorithm that can work with a broad range of arm sets is proposed by Jun et al. (2019). They studied the bilinear bandit model, where the arm set $\mathcal{A}$ is of the form $\left\{xz^\top, x \in \mathcal{X}, z \in \mathcal{Z}\right\}$, and $\mathcal{X}, \mathcal{Z}$ are subsets of $\left\{x \in \mathbb{R}^{d_1} : \|x\|_2 \le 1\right\}, \left\{z \in \mathbb{R}^{d_2} : \|z\|_2 \le 1\right\}$, respectively. They proposed the Explore-Subspace-Then-Refine algorithm that has a regret of $\tilde{O}(\sqrt{\frac{rdT}{\lambda_{\min}(Q(\pi))}\frac{\lambda_{\max}(\Theta^*)}{\lambda_{\min}(\Theta^*)}})$; this is the first algorithm that enjoys regret rate improvements over the naive rate of $\tilde{O}(d^2\sqrt{T})$ obtained by a direct reduction to $d_1 d_2$-dimensional linear bandits, which ignores the low-rank structure. Lu et al. (2021) extended the bilinear arm set to generic matrix arm sets and proposed LowLOC, a computationally inefficient algorithm with $\tilde{O}(\sqrt{rd^3T})$ regret and a computationally efficient algorithm LowESTR with $\tilde{O}(\sqrt{rd^3T}/\lambda_{\min})$ regret. They also proved a $\Omega(rd\sqrt{T})$ regret lower bound for this setting. Kang et al. (2022) designed low-rank bandit algorithms by combining Stein's method for matrix estimation and the Explore-Subspace-Then-Refine framework of (Jun et al., 2019), assuming the existence of a nice exploration distribution over the arm set; their regret bound is $\tilde{O}(\sqrt{rd^2MT}/\lambda_{\min})$, where M is an arm set-dependent constant. However, the M from their given example can have hidden dimensionality dependence – when specialized to the setting of $\mathcal{A}$ being the unit Frobenius norm ball, it is of order $d_1 d_2$, which induces higher regret compared to the previous works with general arm sets (Jun et al., 2019; Lu et al., 2021). See Appendix H for a detailed derivation. In addition, there is no known method to optimize M. As far as we know, (Kang et al., 2022) is the first low-rank bandit paper that applies the techniques of (Minsker, 2018). For the Catoni's estimator, several studies use Catoni's estimator to get a variance-dependent bound on regret bound, such as (Camilleri et al., 2021; Mason et al., 2021; Eftekhari et al., 2023).

Low-rank bandits with specific arm sets There have been lots of other variants of the low-rank bandit, exploiting more specific structures. Katariya et al. (2017) and Trinh et al. (2020) studied rank-1 bilinear bandit problem with canonical arms, which means $\mathcal{A} = \left\{e_i e_j^\top : i \in [d_1], j \in [d_2]\right\}$. Kveton et al. (2017) studied about low-rank bandit where the hidden matrix is a hott topic matrix and arm set is $\{UV^\top : U^\top = [u_1; u_2; \cdots ; u_r], u_i \in \Delta([d_1]), V^\top = [v_1; v_2; \cdots ; v_r], v_i \in \Delta([d_2])\}$. where $[u_1; u_2; \cdots ; u_r]$ refers to concatenation of r vectors to create a matrix. Kotlowski & Neu (2019); Lattimore & Hao (2021); Huang et al. (2021) studied the low-rank bandit with a sphere or unit ball arm set. Though Lattimore & Hao (2021) and Huang et al. (2021) dramatically improved the regret bounds (see Table 1), as Rusmevichientong & Tsitsiklis (2010) have pointed out, the curvature property of the arm set (Huang et al., 2016) can help the agent to improve the regret bound - the regret bound of ETC can be $\sqrt{T}$ when the arm set satisfies certain curvature property. We show in Appendix H that even when the arm set is modified slightly, the regret analysis in these works may no longer go through. In contrast, our algorithm is applicable to general arm sets.

Low-rank contextual bandits with time-varying arm sets Li et al. (2022) studied high-dimensional contextual bandits where at each time step, the set of available arms are drawn iid from some fixed distribution; when specialized to the low-rank linear bandit setting, their setup is different ours due to the nature of time-varying arm sets in their work.

Sparse linear bandits As previously discussed in Section 1, the algorithm presented in this paper draws inspiration from sparse linear bandit algorithms. Reserachers have made significant development on the field of sparse linear bandit algorithms, e.g. (Hao et al., 2020; Jang et al., 2022). These papers extensively utilize the geometry of the arm set and effectively mitigate the dependence on dimensionality in the regret bound.

Low-rank matrix estimation It is natural to apply efficient low-rank matrix recovery results for solving low-rank bandit, since smaller estimation error leads fewer samples for exploration which leads smaller cumulative regret in bandit problems. Keshavan et al. (2010) provides recovery guarantees for projection based rank-r matrix optimization for matrix completion, and Rohde & Tsybakov (2011); Koltchinskii et al. (2011) provide analysis of nuclear norm regularized estimation method for general trace regression, with Rohde & Tsybakov (2011) providing further analysis on the (computationally inefficient) Schatten-p-norm penalized least squares method. In this paper, we mainly use the robust matrix mean estimator of (Minsker, 2018) us it to provide efficient matrix recovery.

B. Proof of Section 3

B.1. Proof of Theorem 3.1

Proof. First, we recall the following lemma of Minsker (2018) on robust matrix mean estimation:

Lemma B.1 (Modification of Corollary 3.1, Minsker (2018))**.** *For a sequence independent, identically distributed random matrices* $(M_i)_{i=1}^n$*, let*

$$\sigma_n^2 = \max\left(\left\|\sum_{i=1}^n \mathbb{E}[M_i M_i^\top]\right\|_{\text{op}}, \left\|\sum_{i=1}^n \mathbb{E}[M_i^\top M_i]\right\|_{\text{op}}\right)$$

Given $\nu = \frac{t\sqrt{n}}{\sigma_n^2}$*, let* $X_i = \phi(\nu\mathcal{H}(M_i))$ *and let* $\hat{T} = \frac{1}{n\nu}(\sum_{i=1}^n X_i)_{\mathsf{ht}}$*. Then, with probability at least* $1-\delta$*,*

$$\|\hat{T} - \mathbb{E}[M_i]\|_{\text{op}} \le \frac{t}{\sqrt{n}}$$

To utilize this Lemma B.1, we choose M_i's so that

- $\mathbb{E}[M_i] = \Theta^* - \Theta_0$ so that $\hat{T}$ estimates the hidden parameter $\Theta^* - \Theta_0$
- σ_m^2 is well-controlled.

It can be checked that $M_i = \text{reshape}\left(\tilde{\Theta}_i\right)$ satisfies the condition with $\sigma_n^2 \le 2(\sigma^2 + R_0^2)B(Q)n_0$ (See Appendix B.2 for the proof.) Substituting σ_n^2 by $2(\sigma^2 + R_0^2)B(Q)n_0$, and setting $t = \sqrt{\frac{2\sigma_n^2}{n_0}\ln\frac{2d}{\delta}}$ leads the desired result. $\square$

B.2. Proof of $\sigma_n^2 \le 2(\sigma^2 + R_0^2)B(Q)$ in Theorem 3.1

Lemma B.2.

$$\sigma_n^2 = \max(\sum_{i=1}^n \|\mathbb{E}[M_i M_i^\top]\|_{\text{op}}, \sum_{i=1}^n \|\mathbb{E}[M_i^\top M_i]\|_{\text{op}}) \le 2nB(Q)(\sigma^2 + R_0^2)$$

Proof. Note that $M_i = \text{reshape}\left(Q(\pi^*)^{-1}(Y_i - \langle\Theta_0, X_i\rangle)\text{vec}(X_i)\right)$, and all M_i are i.i.d. Therefore, $\sigma_n^2 = n \cdot \max(\|\mathbb{E}[M_1 M_1^\top]\|_{\text{op}}, \|\mathbb{E}[M_1^\top M_1]\|_{\text{op}})$, and to compute the first term in the max,

$$\begin{aligned}
\mathbb{E}[M_i M_i^\top] &= \mathbb{E}\left[(Y_i - \langle\Theta_0, X_i\rangle)^2 \text{reshape}\left(Q(\pi)^{-1}\text{vec}(X_i)\right)\text{reshape}\left(Q(\pi)^{-1}\text{vec}(X_i)\right)^\top\right] \\
&\preceq 2\,\mathbb{E}\left[(\eta_i^2 + \langle\Theta_0 - \Theta^*, X_i\rangle^2)\text{reshape}\left(Q(\pi)^{-1}\text{vec}(X_i)\right)\text{reshape}\left(Q(\pi)^{-1}\text{vec}(X_i)\right)^\top\right] \\
&\preceq 2(\sigma^2 + R_0^2)\cdot\mathbb{E}\left[\text{reshape}\left(Q(\pi)^{-1}\text{vec}(X_i)\right)\text{reshape}\left(Q(\pi)^{-1}\text{vec}(X_i)\right)^\top\right]
\end{aligned}$$

where the first inequality holds since $(Y_i - \langle\Theta_0, X_i\rangle)^2 = (\eta_i + \langle\Theta^* - \Theta_0, X_i\rangle)^2 \le 2\eta_i^2 + 2\langle\Theta^* - \Theta_0, X_i\rangle^2$. Now the main task is how to compute $\|\mathbb{E}\left[\text{reshape}\left(Q(\pi^*)^{-1}\text{vec}(X_i)\right)\text{reshape}\left(Q(\pi^*)^{-1}\text{vec}(X_i)\right)^\top\right]\|_{\text{op}}$. Here, we will simply use the definition of the operator norm.

$$\begin{aligned}
&\left\|\mathbb{E}\left[\text{reshape}\left(Q(\pi^*)^{-1}\text{vec}(X_i)\right)\text{reshape}\left(Q(\pi^*)^{-1}\text{vec}(X_i)\right)^\top\right]\right\|_{\text{op}} \\
&= \max_{u\in\mathbb{S}^{d_1-1}} u^\top\,\mathbb{E}\left[\text{reshape}\left(Q(\pi^*)^{-1}\text{vec}(X_i)\right)\text{reshape}\left(Q(\pi^*)^{-1}\text{vec}(X_i)\right)^\top\right]u \\
&= \max_{u\in\mathbb{S}^{d_1-1}} u^\top\,\mathbb{E}\left[\text{reshape}\left(Q(\pi^*)^{-1}\text{vec}(X_i)\right)\cdot\left(\sum_{i=1}^{d_2} e_i^{d_2}(e_i^{d_2})^\top\right)\cdot\text{reshape}\left(Q(\pi^*)^{-1}\text{vec}(X_i)\right)^\top\right]u \\
&= \max_{u\in\mathbb{S}^{d_1-1}}\mathbb{E}\left[\sum_{i=1}^{d_2}\langle(e_i^{d_2}\otimes u), \left(Q(\pi^*)^{-1}\text{vec}(X_i)\right)\rangle^2\right]
\end{aligned}$$

$$= \max_{u\in\mathbb{S}^{d_1-1}} \left[\sum_{i=1}^{d_2} (e_i^{d_2}\otimes u)^\top Q^{-1}(\pi)(e_i^{d_2}\otimes u)\right]$$
$$= \max_{u\in\mathbb{S}^{d_1-1}} \left[u^\top \left(\sum_{i=1}^{d_2} D_i^{(\mathrm{col})}\right) u\right] = \lambda_{\max}(\sum_{i=1}^{d_2} D_i^{(\mathrm{col})})$$

Therefore, we can conclude $\|\mathbb{E}[M_i M_i^\top]\|_{\mathrm{op}} \le (\sigma^2+R_0^2)\lambda_{\max}(\sum_{i=1}^{d_2} D_i^{(\mathrm{col})})$, and similarly $\|\mathbb{E}[M_i^\top M_i]\|_{\mathrm{op}} \le d_1(\sigma^2+R_0^2)\lambda_{\max}(D_i^{(\mathrm{row})})$. Thus,

$$\sigma_n^2 \le 2\max\left(\lambda_{\max}(\sum_{i=1}^{d_2} D_i^{(\mathrm{row})}), \lambda_{\max}(\sum_{i=1}^{d_1} D_i^{(\mathrm{col})})\right)(\sigma^2+R_0^2)n = 2B(Q)(\sigma^2+R_0^2)n.$$

This concludes the proof. □

B.3. Proof of Theorem 3.2

Proof. Note that for all $j \ge r+1$, $\sigma_j(\Theta^*) = 0$. By Weyl's Theorem (Horn & Johnson, 2012), for all $j \ge r+1$, we have that $\sigma_j(\Theta_1) \le 2\sqrt{\frac{((\sigma^2+R_0^2))B(Q)\left(\ln\frac{2d}{\delta}\right)}{n_0}} = \lambda_{\mathrm{th}}$. As a consequence, $\hat{\Theta}$ has rank at most r.

Moreover, by construction, $\|\hat{\Theta}-\Theta_1\|_{\mathrm{op}} \le \lambda_{\mathrm{th}}$. By triangle inequality, we have $\|\hat{\Theta}-\Theta^*\|_{\mathrm{op}} \le 2\lambda_{\mathrm{th}}$. □

B.4. Proof of Corollary 3.3

Proof. For any matrix M, $\|M\|_* \le r\|M\|_{\mathrm{op}}$ and $\|M\|_* \le \sqrt{r}\|M\|_F$. Substitute M to $\hat{\Theta}-\Theta^*$ leads the desired property. □

B.5. Proof of Theorem 3.4

Proof. By Corollary 3.3, the assumption $n_0 \ge \tilde{O}\left(r^2 B(Q)\cdot(\frac{\sigma+S_*}{\sigma})^2\right)$ guarantees that $\|\Theta_0-\Theta^*\|_* \le O(\sigma)$ where Θ_0 is the pilot estimator in Line 2 of Algorithm 2. Therefore, $\max_{A\in\mathcal{A}}|\langle\Theta_0-\Theta, A\rangle| \le \max_{A\in\mathcal{A}}|\Theta_0-\Theta\|_*\|A\|_{\mathrm{op}}| \le O(\sigma)$. We can get our final result by substituting R_0 to $O(\sigma)$ in Theorem 3.2. □

C. Proofs of Lemma 3.5 and 3.6

C.1. Preliminaries - Relationship between $D_i^{(\mathrm{col})}$ and $D_i^{(\mathrm{row})}$

In Figure 1, $D_i^{(\mathrm{col})}$ and $D_i^{(\mathrm{row})}$ looks quite different. However, it turns out that they are coming from the similar logic, due to the nature of the low-rank bandit problem.

Recall the definition of the low-rank bandit problem. For each time, the agent pulls action $A_t \in \mathbb{R}^{d_1\times d_2}$ and receives reward $\langle\Theta^*, A_t\rangle + \eta_t$. However, one could simply transpose all the actions and define $\mathcal{A}^\top := \{a^\top : a\in\mathcal{A}\}$, and think of the reward as $\langle(\Theta^*)^\top, A_t^\top\rangle + \eta_t$. This does not change the nature of the problem. The definition of $D_i^{(\mathrm{col})}$ and $D_i^{(\mathrm{row})}$ comes from this fact.

Since we need to compare the original low-rank bandit problem with 'transposed version' of the low-rank bandit problem, we will redefine our $D_i^{(\mathrm{col})}$ and $D_i^{(\mathrm{row})}$ as follows:

Definition C.1 (Redefine $D_i^{(row)}(Q)$ and $D_i^{(col)}(Q)$). For a covariance matrix Q of the action set $\mathcal{S}\subset\mathbb{R}^{s_1\times s_2}$, define

$$D_i^{(\mathcal{S},col)}(Q) := (Q^{-1})_{[i\cdot s_1+1:(i+1)\cdot s_1],[i\cdot s_1+1:(i+1)\cdot s_1]}$$
$$D_i^{(\mathcal{S},row)}(Q) := [(Q^{-1})_{jk}]_{j,k\in\{i+s_1(\ell-1):\ell\in[s_2]\}}$$

One can check in the case of $\mathcal{A}$, $D_i^{(\mathcal{A},col)}(Q)$ and $D_i^{(\mathcal{A},row)}(Q)$ are exactly same as $D_i^{(\mathrm{col})}(Q)$ and $D_i^{(\mathrm{row})}$, respectively. Moreover, the following properties also hold:

- $\text{vec}(a) = P\text{vec}(a^\top)$ for a fixed permutation matrix $P \in \mathbb{R}^{d_1 d_2 \times d_1 d_2}$.
- Let $Q_{trans}(\pi) := \mathbb{E}_{a\sim\pi}[\text{vec}(a^\top)\text{vec}(a^\top)^\top]$. Then, one could see that $\lambda_{\min}(Q) = \lambda_{\min}(Q_{trans})$ since $Q_{trans} = P^\top Q P$.
- One could check $D_i^{(\mathcal{A},row)}(Q) = D_i^{(\mathcal{A}^\top,col)}(Q_{trans})$ and $D_i^{(\mathcal{A},col)}(Q) = D_i^{(\mathcal{A}^\top,row)}(Q_{trans})$.

Which means, though $D_i^{(\text{col})}$ and $D_i^{(\text{row})}$ looks quite different, $D_i^{(\text{row})}$ is the matrix that come from the same logic as $D_i^{(\text{col})}$, but from the transposed problem.

Therefore, from now on, we will only compute $D_i^{(\text{col})}$ related quantity for the scale comparison in this Section C.

C.2. Proof of Lemma 3.5

Proof. For any vector $v \in \mathbb{R}^{d_1}$, define $\mathsf{Ext}(v,i) \in \mathbb{R}^{d_1 d_2}$ as follows:

$$\mathsf{Ext}(v,i) := e_i^{d_2} \otimes v$$

Then,

$$\begin{aligned}
\lambda_{\max}(\sum_{i=1}^{d_2} D_i^{(\text{col})}) &\le \sum_{i=1}^{d_2} \lambda_{\max}(D_i^{(\text{col})}) && \text{(Homogeneity of degree 1 and convexity of maximum eigenvalue.)}\\
&= \sum_{i=1}^{d-2} \max_{v\in\mathbb{S}^{d_1-1}} v^\top (D_i^{(\text{col})}) v\\
&= \sum_{i=1}^{d-2} \max_{v\in\mathbb{S}^{d_1-1}} \mathsf{Ext}(v,i)^\top Q^{-1} \mathsf{Ext}(v,i)\\
&\le \sum_{i=1}^{d-2} \max_{u\in\mathbb{S}^{d_1 d_2-1}} u^\top Q^{-1} u\\
&= d_2 \lambda_{\max}(Q^{-1}) = \frac{d_2}{\lambda_{\min}(Q)}
\end{aligned}$$

and the proof follows. ☐

C.3. Proof of Lemma 3.6

In this section, we will consider a setting where $d_1 = d_2 = d$, and the following action set, $\mathcal{A}_{\mathsf{hard}} = \{\text{reshape}(a_1), \cdots, \text{reshape}(a_{d^2})\} \subset \mathbb{R}^{d\times d}$ where

$$a_i := \begin{cases} l \cdot e_1 & \text{For } i = 1\\ e_1 + m \cdot e_i & \text{Otherwise}\end{cases}$$

Eventually, we will use $l = \frac{1}{\sqrt{d}}, m = 1$ for our final $\mathcal{A}_{\mathsf{hard}}$, but to demonstrate the effect of each scaling factor, we will use l and m throughout this proof.

In this subsection, we will also use following definitions for the brevity.

- $D := d^2$,
- $\pi_i := \pi(a_i)$, and $\hat{\pi} := (\pi_1, \pi_2, \cdots, \pi_D)$ for any $\pi \in \mathcal{P}(\mathcal{A})$
- $\mathsf{Sym}(n)$ be a permutation group of $[n]$.
- For any permutation $\sigma \in \mathsf{Sym}(n)$
  - For any $v \in \mathbb{R}^d$, let $\sigma(v) := (v_{\sigma(1)}, \cdots, v_{\sigma(n)})$
  - For any $\pi \in \mathcal{P}(\mathcal{A})$, $\sigma(\pi(a_i)) := \hat{\pi}_{\sigma(i)}$

for the brevity.

Now, one could check that

$$Q(\pi) = \begin{bmatrix} l^2\pi_1 + \sum_{i=2}^{D} \pi_i & m\hat{\pi}_{2:D}^\top \\ m\hat{\pi}_{2:D} & m^2 \operatorname{diag}(\hat{\pi}_{2:D}) \end{bmatrix} \tag{12}$$

For the notational convenience, let $\hat{q} = (\pi_1^{-1}, \cdots, \pi_D^{-1})$. Then,

$$Q(\pi)^{-1} = \begin{bmatrix} \frac{1}{l^2\pi_1} & -\frac{1}{ml^2\pi_1}\mathbf{1}_{D-1}^\top \\ -\frac{1}{ml^2\pi_1}\mathbf{1}_{D-1} & \frac{1}{m^2}\operatorname{diag}(\hat{q}_{2:D}) + \frac{1}{l^2m^2\pi_1}\mathbf{1}_{D-1}\mathbf{1}_{D-1}^\top \end{bmatrix}$$

C.3.1. CALCULATE $C_{\min}(\mathcal{A}_{\mathsf{HARD}})$

Suppose that Π^C is the set of optimal experimental designs for $C_{\min}$ (which means, the solution of Eq. (10)). Below, we will show that there exists some π^C in Π^C such that $\pi_2^C = \cdots = \pi_D^C$.

Prove that $\exists \pi^C \in \Pi^C$ such that $\pi_2^C = \cdots = \pi_D^C$ Note that $\lambda_{\max}$ and the matrix inversion are both convex functions. Moreover, from the symmetry of the arm set $\mathcal{A}_{\mathsf{hard}}$, for any permutation $\sigma' \in \mathsf{Sym}(D)$ which satisfies $\sigma'(1) = 1$, for all $\pi \in \mathcal{P}(\mathcal{A}_{\mathsf{hard}})$, $\lambda_{\max}(Q(\pi)^{-1}) = \lambda_{\max}(Q(\sigma'(\pi))^{-1})$. Let

$$\sigma_1(i) = \begin{cases} 1 & \text{if } i = 1 \\ 2 & \text{if } i = D \\ i+1 & \text{Otherwise} \end{cases}$$

Now, fix $\pi \in \Pi^C$; Define π^C to be

$$\pi^C(a_i) = \begin{cases} \pi(a_1) & \text{if } i = 1 \\ \frac{\sum_{s=2}^{D} \pi(a_s)}{D-1} & \text{Otherwise} \end{cases}$$

Then,

$$\begin{aligned} \lambda_{\max}(Q(\pi)^{-1}) &= \frac{1}{D-1}\sum_{s=1}^{D-1} \lambda_{\max}(Q(\sigma_1^s(\pi))^{-1}) \\ &\geq \lambda_{\max}(Q(\frac{1}{D-1}\sum_{s=1}^{D-1} \sigma_1^s(\pi))^{-1}) && \text{(Convexity)} \\ &\geq \lambda_{\max}(Q(\pi^C)^{-1}) \\ &\geq \lambda_{\max}(Q(\pi))^{-1}) && \text{(Minimality of } \pi) \end{aligned}$$

Therefore, $\pi^C \in \Pi^C$, and to calculate $C_{\min}(\mathcal{A}_{\mathsf{hard}})$, it suffices to consider only distributions $\pi \in \mathcal{P}(\mathcal{A}_{\mathsf{hard}})$ which satisfies $\pi_2 = \pi_3 = \cdots = \pi_D$. Let $\pi_1 = a$, and $\pi_2 = b$ for brevity.

Then, the characteristic matrix looks like this: if we let I_n be the $n \times n$ dimensional identity matrix,

$$Q(\pi) - \lambda I_D = \begin{bmatrix} l^2a + (D-1)b - \lambda & mb & \cdots & mb \\ mb & & & \\ \vdots & & (m^2b - \lambda)I_{D-1} & \\ mb & & & \end{bmatrix}$$

and by the row operation,

$$\det(Q(\pi)-\lambda I_D) = \det \begin{bmatrix} l^2a+(D-1)b-\lambda-(D-1)\frac{m^2b^2}{m^2b-\lambda} & 0 & \cdots & 0 \\ mb & & & \\ \vdots & & (m^2b-\lambda)I_{D-1} & \\ mb & & & \end{bmatrix}$$

The characteristic polynomial is therefore

$$\det(Q(\pi)-\lambda I) = (m^2b-\lambda)^{D-2}(\lambda^2-((D-1)b+l^2a+m^2b)\lambda+l^2am^2b)$$

We can therefore get two eigenvalues from quadratic equation, and $D-2$ repeated eigenvalue m^2b.

For eigenvalues from the quadratic equation, note that for the quadratic equation of the form $\lambda^2-B\lambda+C=0 (B,C>0)$, the smaller eigenvalue has the order of $\Theta(\frac{C}{B})$ since $B < B+\sqrt{B^2-4C} < 2B$. Therefore, the order of the eigenvalue is $\Theta(\frac{l^2am^2b}{((D-1)b+l^2a+m^2b)})$ and for the inverse, it's of order $\Theta(\frac{B}{C})$.

When $l,m<1$, one can note that the dominating terms are Db and l^2a on the denominator. Therefore,

$$\lambda_{\max}(Q(\pi)^{-1}) = \Theta(\max(\frac{1}{m^2b}, \frac{D}{m^2l^2a}))$$

Using the fact that $a+(D-1)b=1$, one can get we get optimal rate when $\frac{a}{b}=\Theta(\frac{d}{l^2})$ and the final $C_{\min}^{-1}=\Theta(\frac{D}{m^2l^2})$.

C.3.2. Calculate $B_{\min}(\mathcal{A}_{\text{HARD}})$

From Eq. (12),

$$[Q(\pi)^{-1}]_{1:d,1:d} = \begin{bmatrix} \frac{1}{l^2\pi_1} & -\frac{1}{ml^2\pi_1}\mathbf{1}_{d-1}^\top \\ -\frac{1}{ml^2\pi_1}\mathbf{1}_{d-1} & \frac{1}{m^2}\operatorname{diag}(\hat{q}_{2:d})+\frac{1}{l^2\pi_1m^2}\mathbf{1}_{d-1}\mathbf{1}_{d-1}^\top \end{bmatrix}$$

and

$$[Q(\pi)^{-1}]_{d(i-1)+1:di,d(i-1)+1:di} = \frac{1}{m^2}\operatorname{diag}(\hat{q}_{d(i-1)+1:di})+\frac{1}{l^2\pi_1m^2}\mathbf{1}_d\mathbf{1}_d^\top$$

for $i=2,\cdots,d$. Therefore, if we let $G_i(\pi)=\sum_{j=0}^{d-1}\frac{1}{\pi_{dj+i}}$ for $i=2,\cdots,d-1$, and specially $G_1(\pi)=\sum_{j=1}^{d-1}\frac{1}{\pi_{dj+1}}$,

$$\sum_{i=1}^d D_i^{(\text{col})}(\pi) = \sum_{i=1}^d [Q(\pi)^{-1}]_{d(i-1)+1:di,d(i-1)+1:di} = \begin{bmatrix} \frac{1}{l^2\pi_1}+\frac{1}{m^2}G_1 & -\frac{d-2}{ml^2\pi_1}\mathbf{1}_{d-1}^\top \\ -\frac{d-2}{ml^2\pi_1}\mathbf{1}_{d-1} & \frac{1}{m^2}\operatorname{diag}(G_{2:d})+\frac{d-1}{l^2m^2\pi_1}\mathbf{1}_{d-1}\mathbf{1}_{d-1}^\top \end{bmatrix}$$

Suppose that Π^B is the set of optimal experimental designs for $B_{\min}$ (which means, the solution of Eq. (9)). Below, we will show that there exists some π^B in Π^B such that

- $\pi_i^B=\pi_j^B$ for all $i,j \not\equiv 1 \pmod d$
- $\pi_{d+1}^B=\pi_{2d+1}^B=\cdots\pi_{(d-1)d+1}^B$

C.3.3. Proving that $\exists\pi^B\in\Pi^B$ such that $\pi_i^B=\pi_j^B$ for all $i,j\not\equiv 1 \pmod d$

Let $G=\frac{1}{d-1}\sum_{i=2}^d G_i$, and let $\pi\in\Pi^B$. Let σ be the permutation of $[d]$ which is defined as

$$\sigma(n) = \begin{cases} 1 & \text{if } n\equiv 1 \pmod d \\ 2 & \text{if } n\equiv 0 \pmod d \\ n+1 & \text{otherwise} \end{cases}$$

and ρ be the permutation of $[D]$ which is defined as

$$\rho(n) = \begin{cases} n & \text{if } n \equiv 1 \pmod d \\ n-d+2 & \text{if } n \equiv 0 \pmod d \\ n+1 & \text{otherwise} \end{cases}$$

Then,

$$\begin{aligned}
B(Q(\pi)) &\le \frac{1}{d-1}\sum_{s=1}^{d-1} B(Q(\rho^s(\pi)) \\
&= \frac{1}{d-1}\sum_{s=1}^{d-1} \lambda_{max}\left(\begin{bmatrix} \frac{1}{l^2\pi_1}+\frac{1}{m^2}G_1 & -\frac{d-2}{ml^2\pi_1}\mathbf{1}_{d-1}^\top \\ -\frac{d-2}{ml^2\pi_1}\mathbf{1}_{d-1} & \frac{1}{m^2}\operatorname{diag}(\sigma^s(G)_{2:d}) + \frac{d-1}{l^2\pi_1 m^2}\mathbf{1}_{d-1}\mathbf{1}_{d-1}^\top \end{bmatrix}\right) \\
&\hspace{20em}\text{(Property of permutation } \sigma) \\
&\ge \lambda_{max}\left(\frac{1}{d-1}\sum_{s=1}^{d-1}\begin{bmatrix} \frac{1}{l^2\pi_1}+\frac{1}{m^2}G_1 & -\frac{d-2}{ml^2\pi_1}\mathbf{1}_{d-1}^\top \\ -\frac{d-2}{ml^2\pi_1}\mathbf{1}_{d-1} & \frac{1}{m^2}\operatorname{diag}(\sigma^s(G)_{2:d}) + \frac{d-1}{l^2 m^2\pi_1}\mathbf{1}_{d-1}\mathbf{1}_{d-1}^\top \end{bmatrix}\right) \\
&\hspace{20em}\text{(Jensen's inequality and convexity)} \\
&= \lambda_{max}\left(\begin{bmatrix} \frac{1}{l^2\pi_1}+\frac{1}{m^2}G_1 & -\frac{d-2}{ml^2\pi_1}\mathbf{1}_{d-1}^\top \\ -\frac{d-2}{ml^2\pi_1}\mathbf{1}_{d-1} & \frac{1}{m^2}\operatorname{diag}(G\mathbf{1}_{d-1}) + \frac{d-1}{l^2 m^2\pi_1}\mathbf{1}_{d-1}\mathbf{1}_{d-1}^\top \end{bmatrix}\right)
\end{aligned}$$

Plus, note that when $C' > C > 0$,

$$\lambda_{\max}\left(\begin{bmatrix} \frac{1}{l^2\pi_1}+\frac{1}{m^2}G_1 & -\frac{d-2}{ml^2\pi_1}\mathbf{1}_{d-1}^\top \\ -\frac{d-2}{ml^2\pi_1}\mathbf{1}_{d-1} & \frac{1}{m^2}\operatorname{diag}(C'\mathbf{1}_{d-1}) + \frac{d-1}{l^2 m^2\pi_1}\mathbf{1}_{d-1}\mathbf{1}_{d-1}^\top \end{bmatrix}\right) \ge \lambda_{\max}\left(\begin{bmatrix} \frac{1}{l^2\pi_1}+\frac{1}{m^2}G_1 & -\frac{d-2}{ml^2\pi_1}\mathbf{1}_{d-1}^\top \\ -\frac{d-2}{ml^2\pi_1}\mathbf{1}_{d-1} & \frac{1}{m^2}\operatorname{diag}(C\mathbf{1}_{d-1}) + \frac{d-1}{l^2 m^2\pi_1}\mathbf{1}_{d-1}\mathbf{1}_{d-1}^\top \end{bmatrix}\right).$$

Now consider the following distribution π^B

$$\pi^B(a_n) = \begin{cases} \pi(a_n) & \text{if } n \equiv 1 \pmod d \\ \frac{1-\pi_1-\sum_{i=2}^d \pi_i}{D-d} & \text{otherwise} \end{cases}$$

By AM-HM inequality, we have

$$(d-1)G = \sum_{i=2}^d G_i = \sum_{i \not\equiv 1 \pmod d} \frac{1}{\pi_i} \ge \frac{(D-d)^2}{\sum_{i\not\equiv 1 \pmod d}\pi_i} = \frac{(D-d)^2}{1-\pi_1-\sum_{i=2}^d \pi_i} = \frac{d(d-1)}{\pi_2^B}$$

This means $G \ge \frac{d}{\pi_2^B}$, and naturally

$$\begin{aligned}
B(Q(\pi)) &\ge \lambda_{max}\left(\begin{bmatrix} \frac{1}{l^2\pi_1}+\frac{1}{m^2}G_1 & -\frac{d-2}{ml^2\pi_1}\mathbf{1}_{d-1}^\top \\ -\frac{d-2}{ml^2\pi_1}\mathbf{1}_{d-1} & \frac{1}{m^2}\operatorname{diag}(G\mathbf{1}_{d-1}) + \frac{d-1}{l^2 m^2\pi_1}\mathbf{1}_{d-1}\mathbf{1}_{d-1}^\top \end{bmatrix}\right) \\
&\ge \lambda_{max}\left(\begin{bmatrix} \frac{1}{l^2\pi_1}+\frac{1}{m^2}G_1 & -\frac{d-2}{ml^2\pi_1}\mathbf{1}_{d-1}^\top \\ -\frac{d-2}{ml^2\pi_1}\mathbf{1}_{d-1} & \frac{1}{m^2}\operatorname{diag}(\frac{d}{\pi_2^B}\mathbf{1}_{d-1}) + \frac{d-1}{l^2 m^2\pi_1}\mathbf{1}_{d-1}\mathbf{1}_{d-1}^\top \end{bmatrix}\right) \\
&= B(Q(\pi^B))
\end{aligned}$$

By the minimality of π, $B(Q(\pi)) = B(Q(\pi^B))$ and $\pi^B \in \Pi^B$. Therefore we can conclude that one of the optimal allocation π should satisfy $\pi_i = \pi_j$ for all $i, j \not\equiv 1 \pmod d$.

C.4. Proving that $\exists \pi^B \in \Pi^B$ such that $\pi^B_{d+1} = \pi^B_{2d+1} = \ldots = \pi^B_{(d-1)d+1}$ and $\pi^B_i = \pi^B_j$ for all $i, j \not\equiv 1 \pmod d$

Suppose that $\pi \in \Pi^B$ which satisfies $\pi_i = \pi_j$ for all $i, j \not\equiv 1 \pmod d$). We aim to construct a π^B such that in addition to this property, π^B satisfies $\pi^B_{d+1} = \pi^B_{2d+1} = \ldots = \pi^B_{(d-1)d+1}$.

Define π^B as

$$\pi^B_i = \begin{cases} \frac{\sum_{j=2}^d \pi_j}{d-1} & \text{if } i = 2, \cdots, d \\ \pi_i & \text{Otherwise} \end{cases}.$$

Then, from AM-HM we can note that

$$G_1(\pi) = \sum_{i=2}^d \frac{1}{\pi_i} \geq (d-1)^2 \frac{1}{\sum_{i=2}^d \pi_i} = \sum_{i=2}^d \frac{1}{\pi^B_i} := G_1(\pi^B)$$

and therefore, $B(Q(\pi)) \geq B(Q(\pi^B))$ (we need to change only G_1 to $G_1(\pi^B)$ from the above calculation) and therefore $\pi^B \in \Pi^B$.

C.4.1. Calculating $B_{\min}(\mathcal{A})$

From the above observations, to calculate $B_{\min}(\mathcal{A}_{\textsf{hard}})$, it suffices to restrict to those π's of the following form:

- $\pi_1 = a$
- $\pi_{d+1} = \pi_{2d+1} = \cdots = \pi_{(d-1)d+1} = b$
- $\pi_i = \cdots \pi_D = c$ for all $i \not\equiv 1 \pmod d$.
- $a + (d-1)b + (D-d)c = 1$
- $G_2 = \cdots = G_d = G := \frac{d-1}{b}$, $G_1 = \frac{d-1}{c}$.

To compute the maximum eigenvalue, we should solve the following characteristic equation:

$$\det\left(\left(\begin{bmatrix} \frac{1}{l^2 a} + \frac{1}{m^2} G_1 & -\frac{d-2}{ml^2 a}\mathbf{1}_{d-1}^\top \\ -\frac{d-2}{ml^2 a}\mathbf{1}_{d-1} & \frac{1}{m^2}\operatorname{diag}(G\mathbf{1}_{d-1}) + \frac{d-1}{l^2 a m^2}\mathbf{1}_{d-1}\mathbf{1}_{d-1}^\top \end{bmatrix}\right) - \lambda I\right) = 0$$

$$\Leftrightarrow \det\left(\left(\begin{bmatrix} \frac{1}{l^2 a} + \frac{1}{m^2} G_1 - \lambda & -\frac{d-2}{ml^2 a}\mathbf{1}_{d-1}^\top \\ -\frac{d-2}{ml^2 a}\mathbf{1}_{d-1} & \operatorname{diag}((\frac{G}{m^2} - \lambda)\mathbf{1}_{d-1}) + \frac{d-1}{l^2 a m^2}\mathbf{1}_{d-1}\mathbf{1}_{d-1}^\top \end{bmatrix}\right)\right) = 0$$

$$\Leftrightarrow \det\left(\left(\begin{bmatrix} \frac{1}{l^2 a} + \frac{1}{m^2} G_1 - \lambda - (\frac{(d-2)}{ml^2 a})^2 \frac{1}{\frac{G}{m^2} - \lambda + \frac{(d-1)^2}{l^2 a m^2}} & 0 \\ -\frac{d-2}{ml^2 a}\mathbf{1}_{d-1} & \operatorname{diag}((\frac{G}{m^2} - \lambda)\mathbf{1}_{d-1}) + \frac{d-1}{l^2 a m^2}\mathbf{1}_{d-1}\mathbf{1}_{d-1}^\top \end{bmatrix}\right)\right) = 0$$

(Determinant is invariant under row operation)

$$\Leftrightarrow (\frac{1}{l^2 a} + \frac{1}{m^2} G_1 - \lambda - (\frac{(d-2)}{ml^2 a})^2 \frac{1}{\frac{G}{m^2} - \lambda + \frac{(d-1)^2}{l^2 a m^2}}) \cdot \det\left(\operatorname{diag}((\frac{G}{m^2} - \lambda)\mathbf{1}_{d-1}) + \frac{d-1}{l^2 a m^2}\mathbf{1}_{d-1}\mathbf{1}_{d-1}^\top\right) = 0$$

(Determinant cofactor formula)

$$\Leftrightarrow (\frac{1}{l^2 a} + \frac{1}{m^2} G_1 - \lambda - (\frac{(d-2)}{ml^2 a})^2 \frac{1}{\frac{G}{m^2} - \lambda + \frac{(d-1)^2}{l^2 a m^2}}) \cdot (\frac{G}{m^2} - \lambda + \frac{(d-1)^2}{l^2 a m^2}) \cdot (\frac{G}{m^2} - \lambda)^{d-2} = 0$$

$$\Leftrightarrow [(\frac{1}{l^2 a} + \frac{1}{m^2} G_1 - \lambda) \cdot (\frac{G}{m^2} - \lambda + \frac{(d-1)^2}{l^2 a m^2}) - (\frac{(d-2)}{ml^2 a})^2] \cdot (\frac{G}{m^2} - \lambda)^{d-2} = 0$$

From the above characteristic polynomial, we can notice there are $d-2$ repeated eigenvalues of size G, and the remaining two eigenvalues are the solution of the following quadratic equation:

$$[(\frac{1}{l^2 a} + \frac{1}{m^2}G_1 - \lambda)\cdot(\frac{G}{m^2} - \lambda + \frac{(d-1)^2}{l^2 a m^2}) - (\frac{(d-2)}{m l^2 a})^2] = 0$$

After rearrangement, this formula looks like this:

$$\lambda^2 - B\lambda + C = 0$$

where $B = \frac{G_1}{m^2} + \frac{1}{l^2 a} + \frac{G}{m^2} + \frac{(d-1)^2}{m^2 l^2 a}$ and $C = \frac{1}{m^2}[(\frac{G_1}{m^2} + \frac{1}{l^2 a})\cdot(G + \frac{(d-1)^2}{l^2 a}) - \frac{(d-2)^2}{l^4 a^2}] = \frac{1}{m^2}(\frac{G_1 G}{m^2} + \frac{G_1(d-1)^2 + G}{l^2 a} + \frac{2d-3}{l^4 a^2}) > 0$ and we know that the largest solution of the above quadratic equation is of order B, since $C > 0$ and $0 < B^2 - 4C < B^2$ and therefore $B \le \frac{B + \sqrt{B^2 - 4C}}{2} \le 2B$. Now one could note that $B = \Theta(\max(\frac{G_1}{m^2}, \frac{G}{m^2}, \frac{d^2}{m^2 l^2 a}))$, or

$$B = \Theta(\max\left(\frac{d}{m^2 b}, \frac{d}{m^2 c}, \frac{d^2}{m^2 l^2 a}\right))$$

After optimizing the scale, $a = \Theta(\frac{db}{l^2}), c = \Theta(b)$ and from the constraint $a + (d-1)b + (D-d)c = 1$,

$$\frac{1}{b} = \Theta\left(\frac{d}{l^2} + D\right)$$

and $B = \Theta\left(\frac{d^2}{m^2 l^2} + \frac{d^3}{m^2}\right)$ and so as $B_{\min}(\mathcal{A}_{\mathsf{hard}}) = \Theta\left(\frac{d^2}{m^2 l^2} + \frac{d^3}{m^2}\right)$. When applying $l = \frac{1}{\sqrt{d}}$ and $m = 1$, we get $B_{\min}(\mathcal{A}_{\mathsf{hard}}) = \Theta(d^3)$

Recall that we have shown that $C_{\min}^{-1}(\mathcal{A}_{\mathsf{hard}}) = \Theta\left(\frac{d^2}{m^2 l^2}\right)$; with this choice of l and m, $C_{\min}^{-1}(\mathcal{A}_{\mathsf{hard}}) = \Theta(d^3)$. Therefore, for $\mathcal{A}_{\mathsf{hard}}$, $B_{\min}(\mathcal{A}_{\mathsf{hard}}) = \Theta(C_{\min}^{-1}(\mathcal{A}_{\mathsf{hard}}))$.

D. Examples of $B_{\min}(\mathcal{A})$ and $\mathcal{C}_{\min}(\mathcal{A})$

D.1. $\mathcal{A}$ is Frobenius norm unit ball

Claim 1. If $\mathcal{A}$ is the unit ball in Frobenius norm: $\mathcal{A} = \{A \in \mathbb{R}^{d_1 \times d_2} : \|A\|_F \le 1\}$, then $\mathcal{C}_{\min}(\mathcal{A}) = \frac{1}{d_1 d_2}$ and $B_{\min}(\mathcal{A}) = d_2 d_2 d$.

Proof. We will prove $\mathcal{C}_{\min}(\mathcal{A}) = \frac{1}{d_1 d_2}$ by proving $\mathcal{C}_{\min}(\mathcal{A}) \le \frac{1}{d_1 d_2}$ and $\mathcal{C}_{\min}(\mathcal{A}) \ge \frac{1}{d_1 d_2}$.

Proving $\mathcal{C}_{\min}(\mathcal{A}) \ge \frac{1}{d_1 d_2}$: Let $\mathcal{B} = \{\text{reshape}(e_k) : k = 1, \cdots, d_1 d_2\}$. Note that $\text{vec}(\mathcal{B})$ is a $d_1 d_2$ dimensional canonical basis, and for any $\pi \in \Delta(\mathcal{B})$, $Q(\pi) = \sum_{i=1}^{d_1 d_2} \pi_i e_i e_i^\top = \text{diag}(\pi_1, \cdots, \pi_{d_1 d_2})$ and $\lambda_{\min}(Q(\pi)) = \min\{\pi_i\}_{i=1}^{d_1 d_2}$. Let π be a uniform distribution over $\mathcal{B}$. Then, $\lambda_{\min}(Q(\pi)) = \frac{1}{d_1 d_2}$ and this fact leads to $\mathcal{C}_{\min}(\mathcal{A}) \ge \frac{1}{d_1 d_2}$.

Proving $\mathcal{C}_{\min}(\mathcal{A}) \le \frac{1}{d_1 d_2}$: Fix any distribution π over $\mathcal{A}$. Therefore, $\text{tr}(\mathbb{E}_{a\sim\pi}[\text{vec}(a)\,\text{vec}(a)^\top]) = \mathbb{E}_{a\sim\pi}\,\text{tr}(\|\,\text{vec}(a)\,\text{vec}(a)^\top)] \le 1$ since for all $a \in \mathcal{A}$, $\|a\|_F \le 1$ and $\text{tr}(\text{vec}(a)\,\text{vec}(a)^\top) = \|\,\text{vec}(a)\|_2^2 = \|a\|_F^2 \le 1$. Therefore, by the minimality of $\lambda_{d_1 d_2}$ we get $\lambda_{d_1 d_2}(Q(\pi)) \le \frac{1}{d_1 d_2}\,\text{tr}(Q(\pi)) = \frac{1}{d_1 d_2}$.

Proving $B_{\min}(\mathcal{A}) \ge d_1 d_2 d$: From the definition of $B(Q)$ (Eq. 4),

$$\begin{aligned} B(Q) &= \max\left(\lambda_{\max}\left(\sum_{i=1}^{d_2} D_i^{(\text{col})}\right), \lambda_{\max}\left(\sum_{j=1}^{d_1} D_j^{(\text{row})}\right)\right) \\ &\ge \max\left(\frac{1}{d_1}\mathsf{tr}\left(\sum_{i=1}^{d_2} D_i^{(\text{col})}\right), \frac{1}{d_2}\mathsf{tr}\left(\sum_{j=1}^{d_1} D_j^{(\text{row})}\right)\right) \qquad (\lambda_{\max}(M) \ge \tfrac{1}{d}\mathsf{tr}(M) \text{ for any matrix } M \in \mathbb{R}^{d\times d}) \end{aligned}$$

$$
\begin{aligned}
&= \max\left(\frac{1}{d_1}\mathsf{tr}\left(Q(\pi)^{-1}\right), \frac{1}{d_2}\mathsf{tr}\left(Q(\pi)^{-1}\right)\right) && \text{(From the definition of } D_i^{(\text{col})} \text{ and } D_i^{(\text{row})}) \\
&= \frac{1}{\min(d_1,d_2)}\mathsf{tr}\left(Q(\pi)^{-1}\right) \\
&\geq \frac{1}{\min(d_1,d_2)}\frac{(d_1d_2)^2}{\mathsf{tr}\left(Q(\pi)\right)} && \text{(AM-HM inequality on the spectrum of } Q(\pi)^{-1})
\end{aligned}
$$

Here, note that $\mathcal{A} \subset \mathcal{B}_{Frob}(1)$, which means

$$
\begin{aligned}
\mathsf{tr}\left(Q(\pi)\right) &= \mathsf{tr}(\mathbb{E}_{a\sim\pi}[\text{vec}(a)\text{vec}(a)^\top]) \\
&= \mathbb{E}_{a\sim\pi}[\mathsf{tr}(\text{vec}(a)\text{vec}(a)^\top)] && \text{(Linearity of expectation)} \\
&= \mathbb{E}_{a\sim\pi}[||a||_F^2] \\
&\leq \mathbb{E}_{a\sim\pi}[1] && (a \in \mathcal{A} \subset \mathcal{B}_{Frob}(1)) \\
&= 1
\end{aligned}
$$

Therefore, $B(Q) \geq \frac{(d_1d_2)^2}{\min(d_1,d_2)} = d_1d_2d$ for any $\pi \in \mathcal{P}(\mathcal{A})$

Proving $B_{\min}(\mathcal{A}) \leq d_1d_2d$: Consider

$$
\pi(a) := \begin{cases} \frac{1}{d_1d_2} & \text{if vec}(a) \in \{e_i : i = 1, \cdots d_1d_2\} \\ 0 & \text{Otherwise} \end{cases}
$$

(Recall that e_i is a canonical basis where only i-th entry is 1 and all other entries are 0.) Obviously $\pi \in \mathcal{P}(\mathcal{A})$. On the other hand, $Q(\pi) = \frac{1}{d_1d_2}I_{d_1d_2}$, which means $Q(\pi)^{-1} = d_1d_2I_{d_1d_2}$ and $B(Q) = d_1d_2d$. Therefore, $B_{\min}(\mathcal{A}) \leq d_1d_2d$ by the minimality of $B_{\min}(\mathcal{A})$. □

D.2. $\mathcal{A}$ is operator norm unit ball

Claim 2. If $\mathcal{A}$ is the unit ball in operator norm: $\mathcal{A} = \{A \in \mathbb{R}^{d_1\times d_2} : \|A\|_{\text{op}} \leq 1\}$, then $\mathcal{C}_{\min}(\mathcal{A}) = \Theta(\frac{1}{\max(d_1,d_2)})$ and $B_{\min}(\mathcal{A}) = \max(d_1,d_2)^2$.

Proof. We will prove that $\mathcal{C}_{\min}(\mathcal{A}) = \Theta(\frac{1}{\max(d_1,d_2)})$ by proving $\mathcal{C}_{\min}(\mathcal{A}) = O(\frac{1}{\max(d_1,d_2)})$ and $\mathcal{C}_{\min}(\mathcal{A}) = \Omega(\frac{1}{\max(d_1,d_2)})$. WLOG $d_2 \geq d_1$.

Proving $\mathcal{C}_{\min}(\mathcal{A}) \geq \frac{1}{\max(d_1,d_2)}$: Without loss of generality, assume that $d_2 \geq d_1$; we will show that $C_{\min}(\mathcal{A}) \geq \frac{1}{d_2}$. Consider a distribution $\pi \in \Delta(\mathcal{A})$ which draws a matrix $A \in \mathcal{A}$ by following process:

- Let $U \sim \sigma(d_1)$ and $V = [v_1; v_2; \ldots; v_{d_2}] \sim \sigma(d_2)$ where $\sigma(d)$ denotes the Haar measure over $d \times d$ orthogonal matrices and $[v_1; v_2; \ldots; v_{d_2}]$ is a concatenation of d_2 vectors.
- Let $\Sigma = \begin{bmatrix} I_{d_1} & 0_{d_1\times(d_2-d_1)} \end{bmatrix}$ where I_d denotes d-dimensional identity matrix and $0_{a\times b}$ denotes $a \times b$ dimensional zero matrix.
- Let $A = U\Sigma V^\top = U[v_1; \cdots; v_{d_1}]^\top$. Since U and V are all orthogonal matrices, we have $\|A\|_{\text{op}} = 1$.

Note that A has the same distribution as $[v_1; \cdots; v_{d_1}]^\top$. This is because $AA^\top = UU^\top = I_{d_1}$ so those rows are mutually orthonormal, and for any v_j where $j > d_1$, $Av_j = U\Sigma[v_1; \cdots; v_{d_1}]^\top v_j = U\Sigma 0_{d_1\times 1} = 0_{d_1\times 1}$ which implies that all rows in A and $v_{d_1+1}, \cdots, v_{d_2}$ forms an orthogonal basis. Therefore we can conclude

$$
\begin{bmatrix} v_1; \ldots; v_{d_2} \end{bmatrix} \cdot \begin{bmatrix} U^\top & 0 \\ 0 & I \end{bmatrix} \overset{d}{=} \begin{bmatrix} v_1; \ldots; v_{d_2} \end{bmatrix}
$$

and $A \overset{d}{=} [v_1; \cdots; v_{d_1}]^\top$. Now we should check the covariance matrix of A, $\mathbb{E}\left[\text{vec}(A)\,\text{vec}(A)^\top\right]$. As mentioned in Appendix C.1, there exists a permutation matrix $P \in \mathbb{R}^{d_1d_2\times d_1d_2}$ such that $P\,\text{vec}(A) = \text{vec}(A^\top)$ and $\mathbb{E}\left[\text{vec}(A)\,\text{vec}(A)^\top\right] =$

$P^\top \mathbb{E}\left[\mathrm{vec}(A^\top)\,\mathrm{vec}(A^\top)^\top\right] P$. In our case it is easier to compute $\mathbb{E}\left[\mathrm{vec}(A^\top)\,\mathrm{vec}(A^\top)^\top\right]$. Since $A \overset{d}{=} [v_1; \cdots; v_{d_1}]$,

$$\mathbb{E}\left[\mathrm{vec}(A^\top)\,\mathrm{vec}(A^\top)^\top\right] = \mathbb{E}\begin{bmatrix} V_{1,1} & \cdots & V_{1,d_1} \\ \vdots & \ddots & \vdots \\ V_{d_1,1} & \cdots & V_{d_1,d_1} \end{bmatrix}$$

where $V_{ij} = v_i v_j^\top$. We can easily note that

$$\mathbb{E}\left[V_{ij}\right] = \begin{cases} 0_{d_2\times d_2} & i \neq j \\ \frac{1}{d_2} I_{d_2} & i = j \end{cases}$$

and therefore $\mathbb{E}\left[\mathrm{vec}(A^\top)\mathrm{vec}(A^\top)^\top\right] = \frac{1}{d_2} I_{d_1d_2\times d_1d_2}$. As a result,

$$\mathbb{E}\left[\mathrm{vec}(A)\mathrm{vec}(A)^\top\right] = P^\top\left(\frac{1}{d_2} I_{d_1d_2\times d_1d_2}\right) P = \frac{1}{d_2} P^\top P = \frac{1}{d_2} I_{d_1d_2\times d_1d_2}.$$

This implies that $\mathcal{C}_{\min}(\mathcal{A}) \geq \frac{1}{\max(d_1,d_2)}$.

Proving $\mathcal{C}_{\min}(\mathcal{A}) \leq O(\frac{1}{\max(d_1,d_2)})$: We know that nuclear norm is a convex function. Therefore, $\|\mathbb{E}_{a\sim\pi}[\mathrm{vec}(a)\,\mathrm{vec}(a)^\top]\|_* \leq \mathbb{E}_{a\sim\pi}[\|\mathrm{vec}(a)\,\mathrm{vec}(a)^\top\|_*] \leq d_1 + d_2$ since for all $a \in \mathcal{A}$, $\|a\|_{\mathrm{op}} \leq 1$ means $\|a\|_F \leq \sqrt{\min(d_1,d_2)}$, and $\|\mathrm{vec}(a)\,\mathrm{vec}(a)^\top\|_* = \|\mathrm{vec}(a)\,\mathrm{vec}(a)^\top\|_{\mathrm{op}} = \|\mathrm{vec}(a)\|^2 = \|a\|_F^2 \leq \min(d_1,d_2)$. Therefore, by the minimality of $\lambda_{d_1d_2}$ we get $\lambda_{d_1d_2}(Q(\pi)) \leq \frac{1}{d_1d_2}\|Q(\pi)\|_* = \frac{1}{\max(d_1,d_2)}$.

Proving $B_{\min}(\mathcal{A}) \geq \max(d_1,d_2)^2$: From the definition of $B(Q)$ (Eq. 4),

$$\begin{aligned}
B(Q) &= \max\left(\lambda_{\max}\left(\sum_{i=1}^{d_2} D_i^{(\mathrm{col})}\right), \lambda_{\max}\left(\sum_{j=1}^{d_1} D_j^{(\mathrm{row})}\right)\right) \\
&\geq \max\frac{1}{d_1}\left(\mathrm{tr}\left(\sum_{i=1}^{d_2} D_i^{(\mathrm{col})}\right), \frac{1}{d_2}\mathrm{tr}\left(\sum_{j=1}^{d_1} D_j^{(\mathrm{row})}\right)\right) && (\lambda_{\max}(M) \geq \tfrac{1}{d}\mathrm{tr}(M) \text{ for any matrix } M \in \mathbb{R}^{d\times d}) \\
&= \max\left(\frac{1}{d_1}\mathrm{tr}\left(Q(\pi)^{-1}\right), \frac{1}{d_2}\mathrm{tr}\left(Q(\pi)^{-1}\right)\right) && (\text{From the definition of } D_i^{(\mathrm{col})} \text{ and } D_i^{(\mathrm{row})}) \\
&= \frac{1}{\min(d_1,d_2)}\mathrm{tr}\left(Q(\pi)^{-1}\right) \\
&\geq \frac{1}{\min(d_1,d_2)}\frac{(d_1d_2)^2}{\mathrm{tr}\left(Q(\pi)\right)} && (\text{AM-HM inequality on the spectrum of } Q(\pi)^{-1})
\end{aligned}$$

Here, note that $\mathcal{A} = \mathcal{B}_{\mathrm{op}}(1) \subset \mathcal{B}_{Frob}(\sqrt{\min(d_1,d_2)})$. Then,

$$\begin{aligned}
\mathrm{tr}\left(Q(\pi)\right) &= \mathrm{tr}(\mathbb{E}_{a\sim\pi}[\mathrm{vec}(a)\mathrm{vec}(a)^\top]) \\
&= \mathbb{E}_{a\sim\pi}[\mathrm{tr}(\mathrm{vec}(a)\mathrm{vec}(a)^\top)] && (\text{Linearity of expectation}) \\
&= \mathbb{E}_{a\sim\pi}[||a||_F^2] \\
&\leq \mathbb{E}_{a\sim\pi}[\min(d_1,d_2)] && (a \in \mathcal{A} \subset \mathcal{B}_{Frob}(\sqrt{\min(d_1,d_2)})) \\
&= \min(d_1,d_2)
\end{aligned}$$

Therefore, $B(Q) \geq \frac{(d_1d_2)^2}{\min(d_1,d_2)^2} = \max(d_1,d_2)^2$ for any $\pi \in \mathcal{P}(\mathcal{A})$

Proving $B_{\min}(\mathcal{A}) \leq \max(d_1,d_2)^2$: From Lemma 3.6, $B_{\min}(\mathcal{A}) \leq \frac{\max(d_1,d_2)}{C_{\min}(\mathcal{A})} \leq \max(d_1,d_2)^2$.

□

E. Proof of Theorem 4.1

Proof. First, if $T \le \frac{\sigma^2 r^2 B_{\min}(\mathcal{A})}{S_*^2}$, we have $TS_* \le (\sigma^2 S_* r^2 B_{\min}(\mathcal{A}) T^2)^{1/3})$, therefore

$$\mathsf{Reg}(T) \le TS_* \le \tilde{O}((\sigma^2 S_* r^2 B_{\min}(\mathcal{A}) T^2)^{1/3})$$

trivially holds.

Therefore, throughout the reset of the proof we focus on the case when $T \ge \frac{\sigma^2 r^2 B_{\min}(\mathcal{A})}{S_*^2}$. In this case, $n_0 = \left(\frac{\sigma^2 r^2 B_{\min}(\mathcal{A}) T^2}{S_*^2}\right)^{1/3} \le T$, and by our assumption that $T \ge r^2 B_{\min}(\mathcal{A})(\frac{\sigma + S_*}{\sigma})^4$, we have $n_0 \ge r^2 B_{\min}(\mathcal{A})(\frac{\sigma + R_{\max}}{\sigma})^2$. This range of n_0 satisfies the condition of Theorem 3.2, which gives the following recovery bound on $\hat{\Theta}$ with probability $1-\delta$:

$$\|\hat{\Theta} - \Theta^*\|_* \le 2r\|\hat{\Theta} - \Theta^*\|_{\mathrm{op}} \le 2r\sigma\sqrt{\frac{\left(B_{\min}(\mathcal{A}) \ln \frac{2(d_1+d_2)}{\delta}\right)}{n_0}}$$

For the rest of the rounds, we can bound the instantaneous regret of the exploitation as follows:

$$\begin{aligned}
\langle \Theta^*, A^* - A_t \rangle &= \langle \Theta^* - \hat{\Theta}, A^* \rangle + \langle \hat{\Theta}, A^* \rangle - \langle \Theta^*, A_t \rangle \\
&\le \langle \Theta^* - \hat{\Theta}, A^* \rangle + \langle \hat{\Theta} - \Theta^*, A_t \rangle && \text{(Definition of } A_t\text{)} \\
&\le \|\Theta^* - \hat{\Theta}\|_* (\|A^*\|_{\mathrm{op}} + \|A_t\|_{\mathrm{op}}) && \text{(Holder's inequality)} \\
&\le 2\sigma r \sqrt{\left(2\frac{B_{\min}(\mathcal{A})}{n_0} \ln \frac{2(d_1+d_2)}{\delta}\right)} \times 2
\end{aligned}$$

Therefore, we can conclude the upper bound of the total regret bound as follows:

$$\begin{aligned}
\mathsf{Reg}(T) &= \sum_{t=1}^{T} \langle \Theta^*, A^* - A_t \rangle \\
&\le n_0 S_* + T \cdot 8\sigma r \sqrt{\frac{\left(B_{\min}(\mathcal{A}) \ln \frac{2(d_1+d_2)}{\delta}\right)}{n_0}}
\end{aligned}$$

The final regret bound of Theorem 4.1 follows by plugging in the setting of $n_0 = \left(\frac{\sigma^2 r^2 B_{\min}(\mathcal{A}) T^2}{S_*^2}\right)^{1/3}$. □

F. Results of (Jun et al., 2019)

F.1. LowOFUL Algorithm

Algorithm 5 LowOFUL (Jun et al., 2019)

Input: time horizon T', arm set $\mathcal{A}'$, lower dimension k, regularization parameter λ_1, failure rate δ, positive constants $B, B_\perp, \lambda, \lambda_\perp$

Set $\Lambda = \mathrm{diag}(\lambda, \cdots, \lambda, \lambda_\perp, \cdots, \lambda_\perp)$ where λ occupies the first k diagonal entries, and set $V_0 = \Lambda$, $\theta_0 = \mathrm{vec}\,(0_{d_1 \times d_2})$.

for $t = 1, \cdots, T'$ **do**

  $\sqrt{\beta_t} = \sigma\sqrt{\log \frac{|V_{t-1}|}{|\Lambda|\delta^2}} + \sqrt{\lambda}B + \sqrt{\lambda_\perp}B_\perp$

  $C_t = \{\theta : \|\theta - \hat{\theta}_{t-1}\|_{V_{t-1}} \le \sqrt{\beta_t}\}$

  Compute $a_t = \arg\max_{a \in \mathcal{A}'} \max_{\theta \in C_t} \langle a, \theta \rangle$

  Pull arm a_t and receive reward y_t.

  Update $V_t = V_{t-1} + a_t a_t^\top$, $A = [a_1; \cdots; a_t]$, $\mathbf{y} = [y_1, \cdots, y_t]$ $\theta_t = V_t^{-1} A\mathbf{y}$

end for

F.2. Covariance matrix of (Jun et al., 2019)

Consider π to be the uniform distribution of $\mathcal{X}_0 \times \mathcal{Z}_0$, where $\mathcal{X}_0 = \{X_1, \ldots, X_{d_1}\}$ and $\mathcal{Z}_0 = \{Z_0, \ldots, Z_{d_2}\}$ are sets of linearly independent vectors in $\mathcal{X}$ and $\mathcal{Z}$, respectively. This is exactly how (Jun et al., 2019) have sampled. They achieved the regret bound of on $\tilde{O}(\|X^{-1}\|_{op}\|Z^{-1}\|_{op} d^{3/2}\sqrt{rT})$ where $X := [X_1; \cdots; X_{d_1}]$ and $Z := X := [Z_1; \cdots; Z_{d_2}]$. In this section, we show that this $\|X^{-1}\|_{op}\|Z^{-1}\|_{op}$ is actually $\sqrt{\frac{1}{d_1 d_2 \lambda_{\min}(Q(\pi))}}$, and therefore must be larger than or equal to $\sqrt{\frac{1}{d_1 d_2 C_{\min}(\mathcal{A})}}$.

$$\begin{aligned}
d_1 d_2 Q &= \sum_{i=1}^{d_1}\sum_{j=1}^{d_2} (X_i \otimes Z_j)(X_i \otimes Z_j)^\top \\
&= \sum_{i=1}^{d_1}\sum_{j=1}^{d_2} (X_i \otimes Z_j)(X_i^\top \otimes Z_j^\top) \\
&= \sum_{i=1}^{d_1}\sum_{j=1}^{d_2} (X_i X_i^\top) \otimes (Z_j Z_j^\top) \\
&= \sum_{i=1}^{d_1} (X_i X_i^\top) \otimes (ZZ^\top) \\
&= (XX^\top) \otimes (ZZ^\top)
\end{aligned}$$

Therefore,

$$\begin{aligned}
\frac{1}{d_1 d_2}||Q^{-1}||_{\text{op}} &= ||[(XX^\top) \otimes (ZZ^\top)]^{-1}||_{\text{op}} \\
&= ||[(XX^\top)^{-1} \otimes (ZZ^\top)^{-1}]||_{\text{op}} \\
&= ||[XX^\top]^{-1}||_{\text{op}} ||[ZZ^\top]^{-1}||_{\text{op}} \\
&= ||X^{-1}||_{\text{op}}^2 ||Z^{-1}||_{\text{op}}^2
\end{aligned}$$

G. Proof of Theorem 4.2

Proof. Let's divide the regret of Algorithm 4 into two terms. Let R_1 be the regret occured by the procedure before calling Algorithm 5, and let R_2 be the regret occured by invoking LowOFUL.

Part 1: Bounding R_1. For R_1, since each instantaneous regret is bounded as follows:

$$\langle \Theta^*, A^* - A_t \rangle \le \|\Theta^*\|_* \|A^* - A_t\|_{\text{op}} \le \|\Theta^*\|_* (\|A^*\|_{\text{op}} + \|A_t\|_{\text{op}}) \le 2\|\Theta^*\|_*$$

Therefore, we can safely bound $R_1 \le n_0 \|\Theta^*\|_* \le n_0 S_*$.

Part 2: bounding subspace estimation error. From the analysis on Section 3, we have $\|\hat{\Theta} - \Theta^*\|_{\text{op}} \le \sqrt{\frac{B_{\min}(\mathcal{A})\sigma^2}{n_0}}$. Here, we will use the following operator norm version of Wedin's Theorem (Stewart & Sun, 1990, Theorem 4.4); this is sometimes tighter than the Frobenius norm version of Wedin's Theorem (Stewart & Sun, 1990, Theorem 4.1).

Theorem G.1 (Wedin Theorem)**.** *Let M and M^* be two $d_1 \times d_2$ matrices with the following SVD:*

$$M = \begin{bmatrix} U_1 & U_\perp \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0_{r\times(d_2-r)} \\ 0_{(d_1-r)\times r} & \Sigma_2 \end{bmatrix} \begin{bmatrix} V_1^\top \\ V_\perp^\top \end{bmatrix}$$

$$M^* = \begin{bmatrix} U_1^* & U_\perp^* \end{bmatrix} \begin{bmatrix} \tilde{\Sigma}_1 & 0_{r\times(d_2-r)} \\ 0_{(d_1-r)\times r} & \tilde{\Sigma}_2 \end{bmatrix} \begin{bmatrix} (V_1^*)^\top \\ (V_\perp^*)^\top \end{bmatrix}$$

Where

- $\Sigma_1, \tilde{\Sigma}_1$ *represents top-r singular values for M and M^**

- $\Sigma_2, \tilde{\Sigma}_2$ *represents the rest of singular values for* M *and* M^*
- (U_1, V_1), (U_1^*, V_1^*) *be the corresponding singular vectors for* Σ_1 *and* $\tilde{\Sigma}_1$
- $(U_\perp, V_\perp)$, $(U_\perp^*, V_\perp^*)$ *be the corresponding singular vectors for* Σ_2 *and* $\tilde{\Sigma}_2$

respectively. Suppose that there are numbers $\alpha, \delta > 0$ *such that*

$$\lambda_r(\Sigma_1) \geq \alpha + \delta, \textit{ and } \lambda_{\max}(\tilde{\Sigma}_2) \leq \alpha$$

Then,

$$\max\{\|U_\perp^\top U_1^*\|_{\text{op}}^2, \|V_\perp^\top V_1^*\|_{\text{op}}^2\} \leq \frac{\max\{\|(U_1^*)^\top (M - M^*)\|_{\text{op}}^2, \|(M - M^*)V_1^*\|_{\text{op}}^2\}}{\delta^2}$$

Now, substitute parameters as follows: suppose the SVD of $\Theta^* = U^* \Sigma^* V^*$

$$\begin{aligned}
M &= \hat{\Theta}, M^* = \Theta^* \\
U_1 &= U_1, U^* = U^* \\
V_1 &= V_1, V^* = V^* \\
\Sigma_1 &= \tilde{\Sigma}_1, \Sigma_1^* = \Sigma_1^*
\end{aligned}$$

Plus, note that

$$\|\Theta_1 - \Theta^*\|_{\text{op}}^2 = \|\Theta_1 - \Theta^*\|_{\text{op}}^2 \|V_1^*\|_{\text{op}}^2 \geq \|(\Theta_1 - \Theta^*)V_1^*\|_{\text{op}}^2$$

and similarly, $\|\Theta_1 - \Theta^*\|_F^2 \geq \|U_1^*(\Theta_1 - \Theta^*)\|_F^2$. Now Wedin's theorem implies that

$$\max(\|U_\perp^\top U_1^*\|_{\text{op}}^2, \|V_\perp^\top V_1^*\|_{\text{op}}^2) \leq \frac{\|\Theta - \Theta^*\|_{\text{op}}^2}{\delta^2}$$

We can check that by the assumption that $T \geq \frac{16 B_{\min}(\mathcal{A})\sigma^4}{d^{0.5} S_r^2}$, $n_0 \geq \frac{4 B_{\min}(\mathcal{A})\sigma^2}{S_r^2}$. Thus by Weyl's Theorem, $\lambda_r(\hat{\Theta}) \geq \lambda_r(\Theta^*) - \sqrt{\frac{B_{\min}(\mathcal{A})\sigma^2}{n_0}} \geq \frac{S_r}{2}$, therefore, choosing $\delta = \frac{S_r}{2}, \alpha = 0$ satisfies the condition, since the rank of $\hat{\Theta}$ is r and therefore $\lambda_j(\Sigma_2) = 0$ for all $j = r+1, \ldots, \min(d_1, d_2)$. With the result of Theorem 3.4, we can conclude

$$\|\hat{U}_\perp^\top U^*\|_{\text{op}} \leq \frac{1}{S_r}\sqrt{\frac{B_{\min}(\mathcal{A})\sigma^2}{n_0}}, \|\hat{V}_\perp^\top V^*\|_{\text{op}} \leq \frac{1}{S_r}\sqrt{\frac{B_{\min}(\mathcal{A})\sigma^2}{n_0}}$$

Therefore, $\|\hat{U}_\perp^\top \Theta \hat{V}_\perp\|_F \leq \|\hat{U}_\perp^\top U\|_{\text{op}} \cdot \|\Sigma\|_F \cdot \|V^\top \hat{V}_\perp\|_{\text{op}} \leq \frac{B_{\min}(\mathcal{A})\sigma^2 \|\Theta^*\|_F}{n_0 S_r^2} \leq \frac{B_{\min}(\mathcal{A})\sigma^2 S_*}{n_0 S_r^2} =: B_\perp$

Part 3: bounding R_2. Recall that we set $\lambda_\perp = \frac{T}{r}, B_\perp = \frac{\sigma^2 B_{\min}(\mathcal{A}) S_*}{n_0 S_r^2}$ in low-OFUL

Let reg_t be the instantaneous pseudo-regret at time step t: $reg_t = \langle \Theta^*, A^* - A_t \rangle = \langle \text{vec}(\Theta^*), \text{vec}(A^*) - \text{vec}(A_t) \rangle$ where $A^* = \arg\max_{A \in \mathcal{A}} \langle \Theta^*, A \rangle$. From the fact that $\Theta^* \in C_t$ and using Cauchy-Schwarz inequality, we have

$$\begin{aligned}
reg_t &= \langle \text{vec}(\Theta^*), \text{vec}(A^*) - \text{vec}(A_t) \rangle \\
&\leq \max_{\Theta \in C_{t-1}} \langle \text{vec}(\Theta) - \text{vec}(\Theta^*), \text{vec}(A_t) \rangle && \text{(Definition of } A_t\text{)} \\
&\leq \max_{\Theta \in C_{t-1}} \|\text{vec}(\Theta) - \text{vec}(\Theta^*)\|_{V_{t-1}} \|\text{vec}(A_t)\|_{V_{t-1}^{-1}} && (13) \\
&\leq 2\sqrt{\beta_t} \|\text{vec}(A_t)\|_{V_{t-1}^{-1}} && \text{(Definition of } C_t\text{)} \\
&\leq \sqrt{\beta_T} \|\text{vec}(A_t)\|_{V_{t-1}^{-1}} && (14)
\end{aligned}$$

Now, define $H_T := \{t \in [T] : t > n_0, \|A_t\|_{V_{t-1}^{-1}} > 1\}$ and $\bar{H}_T := \{t \in [T] : t > n_0, \|A_t\|_{V_{t-1}^{-1}} \leq 1\}$. Then,

$$R_2 = \sum_{t=n_0+1}^{T} reg_t$$

$$= \sum_{t=n_0+1}^{T} reg_t \mathbb{1}\{t \in \bar{H}_T\} + \sum_{t=n_0+1}^{T} reg_t \mathbb{1}\{t \in H_T\}$$

$$= \sum_{t=n_0+1}^{T} reg_t \mathbb{1}\{t \in \bar{H}_T\} + 2S_*|H_T| \qquad (reg_t \le 2S_*)$$

$$\le \sqrt{\|\bar{H}_T| \sum_{t\in\bar{H}_T} reg_t^2} + 2S_*|H_T| \qquad \text{(Cauchy-Schwarz)}$$

$$\le \sqrt{\|\bar{H}_T|\beta_T \sum_{t\in\bar{H}_T} \|\operatorname{vec}(A_t)\|_{V_{t-1}^{-1}}} + 2S_*|H_T| \qquad \text{(Eq. (14))}$$

$$\le \sqrt{\|\bar{H}_T|\beta_T \sum_{t=n_0+1}^{T} \min(1, \|\operatorname{vec}(A_t)\|_{V_{t-1}^{-1}}) + 2S_*|H_T|} \tag{15}$$

Now for the first term of Eq. (15), we can use the elliptic potential lemma (Abbasi-Yadkori et al., 2011; Lattimore & Szepesvári, 2020):

Lemma G.2 ((Lattimore & Szepesvári, 2020), Lemma 19.4)**.** $\sum_{t=1}^{n} \min(1, \|\operatorname{vec}(A_t)\|_{V_{t-1}^{-1}}) \le 2\log\frac{|V_T|}{|\Lambda|}$

For the second term, $S_*|H_T|$, we can use the slight modification of the elliptical potential count lemma in (Gales et al., 2022):

Lemma G.3 (Modification of Lemma 7, (Gales et al., 2022))**.** $|H_T| \le \frac{2d}{\log 2}\log(\frac{2d}{\log 2})$

Overall, we have

$$R_2 \le 4\sqrt{\beta_T}\sqrt{\log\frac{|V_T|}{|\Lambda|}}\sqrt{T} + \frac{4d}{\log 2} S_* d \log(\frac{2d}{\log 2}) \tag{16}$$

where $\sqrt{\beta_t} = B\sqrt{\lambda} + B_\perp\sqrt{\lambda_\perp} + \sigma\sqrt{\log\frac{|V_t|}{|\Lambda|}}$.

Now the minor difference comes from the computation of $\log\frac{|V_T|}{|\Lambda|}$, simply because we have different bounds on the l_2 norm of the actions (note that for all $a \in \mathcal{A}'$, $\|a\|_2 = \|\text{reshape}(a)\|_F \le \sqrt{d}\|\text{reshape}(a)\|_{\text{op}} \le \sqrt{d}$.)

Lemma G.4 (Modification of Valko et al. (2014), Lemma 5)**.** *For any T, let $\Lambda = diag([\lambda_1, \cdots, \lambda_p])$. Then,*

$$\log\frac{|V_T|}{|\Lambda|} \le \max\{\sum_{i=1}^{p} \log(1 + \frac{dt_i}{\lambda_i})\}$$

where the maximum is taken over all possible positive real numbers $t_1, \cdots, t_p$ such that $\sum_{i=1}^{p} t_i = T$.

Note that in comparison with (Valko et al., 2014) (which originally assumes $\|a_t\|_2 \le 1$ for all t), we added a factor of d inside the log because $V_T = \sum_{t=1}^{T} a_t a_t^\top$ and each $\|a_t\|_2 \le \sqrt{d}$. Detailed proof is in Appendix G.1.1

The only difference from the original lemma is that our Frobenius norm of $\|a\|_F$ is bounded by $\sqrt{d}$, so we need to compensate that scale difference inside the logarithm. Using our $\Lambda = \text{diag}(\lambda, \cdots, \lambda, \lambda_\perp, \cdots, \lambda_\perp)$ with Lemma G.4 we can induce the following result:

Lemma G.5 (Jun et al. (2019), Lemma 3)**.** *If $\lambda_\perp = \frac{T}{r\log(1+\frac{dT}{\lambda})}$, then*

$$\log\frac{|V_T|}{|\Lambda|} \le 2k\log(1 + \frac{dT}{\lambda})$$

Proof.

$$\log\frac{|V_T|}{|\Lambda|} \le \max\{\sum_{i=1}^{p} \log(1 + \frac{dt_i}{\lambda_i})\}$$

$$\leq k\log(1+\frac{dT}{\lambda})+\sum_{i=k+1}^{p}\log(1+\frac{dt_i}{\lambda_\perp})$$

$$\sum_{i=k+1}^{p}\log(1+\frac{dt_i}{\lambda_i})\leq\sum_{i=k+1}^{p}(\frac{dt_i}{\lambda_\perp})\leq\frac{dT}{\lambda_\perp}\leq k\log(1+\frac{dT}{\lambda})$$

□

One can note that the additional d factor from Lemma G.4 leads $\lambda_\perp$ should have order $\frac{T}{r}$, not like $\frac{T}{k}$ in Jun et al. (2019). Combining Lemma G.5 with Eq. (16), regret occured by the LowOFUL algorithm is

$$\begin{aligned}R_2 &\leq \tilde{O}((\sigma k\sqrt{T}+B\sqrt{k\lambda T}+B_\perp\sqrt{k\lambda_\perp T}))\\ &\leq \tilde{O}(\sigma rd\sqrt{T}+T\sqrt{d}B_\perp)\\ &\leq \tilde{O}(\sigma rd\sqrt{T}+T\frac{\sigma^2 d^{0.5}B_{\min}(\mathcal{A})S_*}{n_0 S_r^2})\end{aligned}$$

Part 4: putting it together. Therefore, the total regret of ESTR can be bounded by

$$\begin{aligned}\mathsf{Reg}_T = R_1+R_2 \leq& \tilde{O}\left(n_0 S_* + \sigma rd\sqrt{T}+\frac{Td^{0.5}B_{\min}(\mathcal{A})\sigma^2 S_*}{n_2 S_r^2}\right)\\ \leq& \tilde{O}\left(\sigma rd\sqrt{T}+\sigma\sqrt{S_*^2\frac{d^{0.5}B_{\min}(\mathcal{A})}{S_r^2}T}\right)\end{aligned}$$

with the setting of $n_0=\sqrt{\frac{d^{0.5}B_{\min}(\mathcal{A})}{S_r^2}T}$ in the algorithm. □

G.1. Proof of Lemmas we have used in this section

G.1.1. Proof of Lemma G.4

Proof. We need the following lemma of Valko et al. (2014)

Lemma G.6 (Modification of Valko et al. (2014), Lemma 4)**.** *Let $\Lambda = diag(\lambda_1,...,\lambda_p)$ be any diagonal matrix with strictly positive entries. Then for any vectors $(a_t)_{1\leq t\leq T}$ such that $\|a_t\|_2\leq C$ for some constant C for all $1\leq t\leq T$, we have that the determinant $|V_T|$ is maximized when all a_t are aligned with the axes.*

The proof of Lemma G.6 is exactly the same as Valko et al. (2014), Lemma 4. Now, in our case, for each $1\leq t\leq T$, $x_t=\text{vec}\,X_t$ and $\|x_t\|_2\leq\|X_t\|_F\leq\sqrt{d}\|X_t\|_{op}\leq\sqrt{d}$. Now,

$$\begin{aligned}|V_T| &= |\Lambda+\sum_{t=1}^{T}x_t x_t^\top|\\ &\leq \max_{(a_i)_{i=1}^t:\|a_i\|_2\leq\sqrt{d}}|\Lambda+\sum_{t=1}^{T}a_t a_t^\top|\\ &= \max_{(a_i)_{i=1}^t:a_i\in\{\sqrt{d}e_1,\cdots,\sqrt{d}e_p\}}|\Lambda+\sum_{t=1}^{T}a_t a_t^\top| \qquad \text{(Lemma G.6)}\\ &\leq \max_{(t_i)_{i=1}^t:t_i\geq 0,\sum_{i=1}^t t_i=T}(\lambda_i+dt_i)\end{aligned}$$

and dividing $|V_T|$ by $|\Lambda|$ and taking logarithm leads the result of Lemma G.4. □

H. Additional discussions of related works

H.1. Discussion of Huang et al. (2021)

The result of Huang et al. (2021) is mainly based on the noisy power method (Hardt & Price, 2014). After using noisy power method to estimate $\hat{\Theta}$ such that $\|\hat{\Theta} - \Theta^*\|_F \le \varepsilon \|\Theta\|_F$, they use the fact that their arm set is a sphere and therefore the empirical best arm (greedy) is explicitly $\hat{A} = \hat{\Theta}/\|\hat{\Theta}\|_F$ and the true best arm $A^* = \Theta^*/\|\Theta^*\|_F$.

$$\|\hat{\Theta} - \Theta^*\|_F \le \varepsilon \|\Theta^*\|_F$$
$$\iff \|\frac{\hat{\Theta}}{\|\Theta^*\|_F} - A^*\|_F \le \varepsilon$$

and by trigonometry, one can deduce that $\|\hat{A} - A^*\|_F \le \varepsilon$. See Huang et al. (2021, Appendix B.2) for details.

They then use the fact $\hat{A} = \hat{\Theta}/\|\hat{\Theta}\|_F$ and $A^* = \Theta^*/\|\Theta^*\|_F$ to achieve the instantaneous regret bound of ε^2 as follows:

$$\langle \Theta^*, A^* \rangle - \langle \Theta^*, \hat{A} \rangle = \frac{\langle \Theta^*, A^* \rangle}{2}(2 - \langle \frac{\Theta^*}{\|\Theta^*\|_F}, \hat{A} \rangle) = \frac{\|\Theta^*\|_F}{2} \left\| \frac{\Theta^*}{\|\Theta^*\|_F} - \hat{A} \right\|_F^2 \le \|\Theta\|_F \varepsilon^2 \tag{17}$$

This small ε^2 error guarantee (as opposed to, say, ε described below) is crucial for obtaining their regret bound.

To summarize, a key property Huang et al. (2021) used was the fact when Θ^* and $\hat{\Theta}$ are close enough, then A^* and $\hat{A}$ is also close enough in their setting. This is true when the arm set $\mathcal{A}$ has a smooth curvature. However, without curvature on the arm set, the greedy arm $\hat{A} = \arg\max_{A \in \mathcal{A}} \langle \hat{\Theta}, A \rangle$ can only be guaranteed such that

$$\langle \Theta^*, A^* \rangle - \langle \Theta^*, \hat{A} \rangle \le 2 \max_{a \in \mathcal{A}} |\langle \hat{\Theta} - \Theta^*, A \rangle| \le O(\varepsilon)$$

Here's one example that shows the importance of the Frobenius norm unit ball arm set for their anaylsis. Suppose that arm set $\mathcal{A} = \mathcal{B} \cup \{\text{diag}(1, 1, 0, \cdots, 0)\}$, where $\mathcal{B} = \{M \in \mathbb{R}^{d \times d} : \|M\|_F \le 1\}$. Consider $\Theta^* = \text{diag}(1, \varepsilon, 0, \cdots, 0)$ for some small ε. Suppose that we run the algorithm of Huang et al. (2021) using $\mathcal{B}$. Then, for an arbitrary estimation error ε_h, the for the estimator using Huang et al. (2021), $\hat{\Theta}_h$, we have guarantee $\|\hat{\Theta}_h - \Theta^*\| \le \varepsilon_h \|\Theta^*\|_F$ when $n_0 = \tilde{O}(d^2 r \lambda_r^{-2} \varepsilon_h^{-2})$ is number of total exploration steps (From Theorem 3.8 of (Huang et al., 2021)). As we have stated above, Huang et al. (2021) converted this to a bound of $\|\hat{A} - A^*\|_F$ when the arm set was $\mathcal{B}$. However, in the case when the arm set is $\mathcal{A}$, $\hat{A}$ and A^* can be close enough only when $(\hat{\Theta}_h)_{22}$ is positive. If not, then we have $\langle \Theta^*, \text{diag}(1, 1, 0, \cdots, 0) \rangle - \max_{A \in \mathcal{B}} \langle \Theta^*, A \rangle = \Omega(\varepsilon)$ and this incurs εT exploitation regret. To guarantee $\varepsilon_h \le \varepsilon$, we need to spend $\tilde{O}(d^2 r \lambda_r^{-2} \varepsilon^{-2})$ samples for exploration. Thus, with this analysis, the best regret upper bound we can hope for is

$$\min(\varepsilon T, \frac{d^2 r}{\lambda_r^2 \varepsilon^2}).$$

Choosing ε that maximizes this leads to $\tilde{O}((d^2 r T^2 \lambda_r^2)^{1/3})$ regret upper bound. This is much worse than their previous bound $\tilde{O}(\sqrt{d^2 r T}/\lambda_r)$.

H.2. Discussion of Kang et al. (2022)

The result of Kang et al. (2022) is directly associated with a sampling distribution constant called M, which was treated as a constant unrelated to dimensionality in the paper. However, we explain here that M, has hidden dependence on the dimensionality.

To see this, consider the reward model $y_t = \langle \Theta^*, X_t \rangle + \eta_t$ where $\eta_t \sim N(0, \sigma^2)$. It lies in the (conditional) canonical exponential family:

$$p_{\Theta^*}(y_t \mid X_t) = \exp\left(\frac{y_t \beta - b(\beta)}{\phi} + c(y_t, \phi) \right),$$

where $\beta = \langle \Theta^*, X_t \rangle$, $b(\beta) = \frac{1}{2}\beta^2$, $\phi = \sigma^2$, $c(y_t, \phi) = \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{y_t^2}{2\sigma^2}$. The inverse link function is $\mu(\beta) = \nabla b(\beta) = \beta$.

Consider arm set $\mathcal{A} = \left\{X \in \mathbb{R}^{d_1 \times d_2} : \|X\|_F \le 1\right\}$. We consider $\mathcal{D} = N(0, \frac{c}{d_1 d_2} I_{d_1 d_2})$ (with $c = O(\frac{1}{\ln T})$), so that T

arms drawn iid from $\mathcal{D}$ all lie in $\mathcal{X}$ with probability $1 - O(\frac{1}{T})$. With this distribution,

$$p(X) \propto \exp(-\frac{\|X\|_F^2}{\frac{2c}{d_1 d_2}}) \implies \ln p(X) = -\frac{d_1 d_2 \|X\|_F^2}{2c} + \text{constant},$$

Therefore, the associated score function $S(X) = \nabla \ln p(X) = -\frac{d_1 d_2}{c} X$.

Now, checking (Kang et al., 2022, Assumption 3.3), we have for all i, j,

$$\mathbb{E}\left[S(X)_{i,j}^2\right] = \frac{(d_1 d_2)^2}{c^2} \mathbb{E}\left[X_{i,j}^2\right] = \frac{d_1 d_2}{c}$$

As M is chosen such that for all i, j, $\mathbb{E}\left[S(X)_{i,j}^2\right] \le M$, M has to be at least $\frac{d_1 d_2}{c} \ge d_1 d_2$.

Plugging this into (Kang et al., 2022, Theorem 4.1), and note that $\mu^* = \mathbb{E}\left[\mu'(\langle X, \Theta^* \rangle)\right] = 1$, we have that given T_1 iid samples from $\mathcal{D}$, the estimator $\hat{\Theta}$ has a Frobenius recovery error bound of:

$$\|\hat{\Theta} - \Theta^*\|_F^2 \le \tilde{O}((\sigma^2 + S_f^2) \frac{(d_1 d_2)^2 r}{T_1}).$$

When $d_1 = \Theta(d_2)$, this guarantee is in fact worse than the recovery bound obtained by the $\tilde{O}(\sigma^2 \frac{d^3 r}{T_1})$ bound provided by nuclear norm penalty method (Lu et al., 2021, Theorem 16). We also provide an different argument on the necessity of hidden dimensionality dependence of M or S_f. Suppose not; then $\hat{\Theta}$, based on iid measurements from $\mathcal{D}$, achieves $\|\hat{\Theta} - \Theta^*\|_F^2 \le \tilde{O}(\sigma^2 \frac{d_1 d_2 r}{T_1})$. However, the information-theoretic lower bound (Koltchinskii et al., 2011, Theorem 5) implies that for any estimator using samples drawn iid from $\mathcal{D}$, there exists some Θ such that $\max_{i,j} |\Theta_{i,j}| \le \sigma$, and $\|\hat{\Theta} - \Theta^*\|_{\mathcal{D}}^2 \ge \Omega(\sigma^2 \frac{r(d_1+d_2)}{T})$. The latter implies that $\|\hat{\Theta} - \Theta^*\|_F^2 \ge \Omega(\sigma^2 \frac{r(d_1+d_2) d_1 d_2}{T})$.

For Kang et al. (2022), they also stated their result based on the Frobenius norm bounded arm set: $\mathcal{A} \subset \{A \in \mathbb{R}^{d_1 \times d_2} : \|A\|_F \le 1\}$. When we change the Frobenius norm bound to operator norm bound, their estimation bound (Kang et al. (2022, Theorem 4.1)) does not change much, but their regret analysis on ESTS needs additional $d^{0.25}$ factor.This additional dimensional dependence also applies for all ESTR-based algorithms Jun et al. (2019); Lu et al. (2021) and it is because of the log-determinant term computation - check Lemma G.5 and Lemma G.4 to see details of why additional d appears.

H.3. Justifying (Lu et al., 2021) bound in Table 1

In this section, we show that the regret bound of LowESTR (Lu et al., 2021) (originally proposed for the setting of $\mathcal{A} \subset B_{\text{Frob}}(1)$), when applied to our setting ($\mathcal{A} \subset B_{\text{op}}(1)$), gives a regret bound of $\tilde{O}(d^{1/4}\sqrt{r \frac{(S_* + \sigma)^2}{\lambda_{\min}(Q(\pi))^2} T} \left(\frac{S_*}{\lambda_r}\right))$. First, with the new assumption on the arm set $\mathcal{A}$, it is necessary to set $\lambda_\perp = \frac{T}{r}$ instead of $\frac{T}{rd}$ in (Jun et al., 2019; Lu et al., 2021) to ensure that $\log \frac{|V_T|}{|\Lambda|} \le \tilde{O}(rd)$.

Therefore, the total regret bound of LowESTR is

$$\tilde{O}\left(S_* n_0 + \sigma k \sqrt{T} + B\sqrt{k\lambda T} + B_\perp \sqrt{k \lambda_\perp T}\right)$$

Next, as mentioned in Remark 3, our LPA-ESTR also achieves an improved regret guarantee over LowESTR ((Lu et al., 2021)) not only w.r.t. d but also w.r.t. rank r too.

The main reason is that the LowPopArt provides operator norm-based recovery bound as discussed in Theorem 3.2. This allows us to use the operator norm version of Wedin Theorem (See Section G), which means we obtained the bound of $\|U_\perp^\top U^*\|_{\text{op}}$ and $\|V_\perp^\top V^*\|_{\text{op}}$. From this bound, we used the fact that $\|AB\|_F \le \|A\|_{\text{op}} \|B\|_F$ to derive the following relationship:

$$\|\hat{U}_\perp^\top \Theta \hat{V}_\perp\|_F \le \|\hat{U}_\perp^\top U\|_{\text{op}} \cdot \|\Sigma\|_F \cdot \|V^\top \hat{V}_\perp\|_{\text{op}} \le \frac{B_{\min}(\mathcal{A}) \sigma^2 \|\Theta^*\|_F}{n_0 \lambda_r(\Theta^*)^2} \quad \text{(This is LowPopArt version.)}$$

Remember that there's no r term on the RHS. On the other hand, (Lu et al., 2021) used the Frobenius norm version of the Wedin Theorem, since they mainly used the Frobenius norm bound of the nuclear norm regularized least square.

Theorem H.1 (Lemma 23 and Appendix E.2 of (Lu et al., 2021))**.** *For the nuclear norm regularized least square estimator*

$\hat{\Theta}_{nuc}$, *we have*

$$\|\hat{\Theta}_{nuc} - \Theta^*\|_F^2 \leq 4.5\frac{\lambda_n^2}{\kappa^2} r \approx \frac{\sigma^2}{n\lambda_{\min}(Q(\pi))^2} \cdot r$$

where κ is the restricted strong convexity constant (in (Lu et al., 2021) it is $\lambda_{\min}(Q(\pi))$), and λ_n is a constant which satisfies $\|\frac{1}{n}\sum_{t=1}^n \eta_t X_t\|_{\text{op}} \leq \frac{\lambda_n}{2}$ (it is $O(\sqrt{\frac{\sigma}{n}})$; by (Koltchinskii et al., 2011), Propositioin 2).

Under this result, they are forced to use the Frobenius version of Wedin's Theorem and trivially bound $\|U_\perp^\top U^*\|_{\text{op}}$ by $\|U_\perp^\top U^*\|_F$ (marked as (opF) in Eq. (18)). This leads to the following looser estimation:

$$\|\hat{U}_\perp^\top \Theta \hat{V}_\perp\|_F \leq \|\hat{U}_\perp^\top U\|_{\text{op}} \cdot \|\Sigma\|_F \cdot \|V^\top \hat{V}_\perp\|_{\text{op}} \overset{(opF)}{\leq} \|\hat{U}_\perp^\top U\|_F \cdot \|\Sigma\|_F \cdot \|V^\top \hat{V}_\perp\|_F \leq \frac{\sigma^2\|\Theta^*\|_F}{\lambda_{\min}(Q(\pi))^2 \cdot n_0\lambda_r(\Theta^*)^2} \cdot r \tag{18}$$

Note that there's r term on RHS now. Since $\frac{1}{\lambda_{\min}(Q(\pi))} \geq \frac{1}{C_{\min}} \geq \frac{B_{\min}(\mathcal{A})}{d} \geq d$ by Lemma 3.5, LowPopArt version bound is much tighter than Eq. (18) in all manners.

Now from the construction, $B_\perp \leq \frac{\sigma^2\|\Theta^*\|_F}{\lambda_{\min}(Q(\pi))^2 \cdot n_0\lambda_r(\Theta^*)^2} \cdot r$.

Therefore, the total regret of LowESTR can be bounded by

$$\begin{aligned} Reg(T) \leq& \tilde{O}\left(n_0 S_* + \sigma r d\sqrt{T} + \frac{T d^{0.5}\sigma^2 S_*}{n_0\lambda_{\min}(Q(\pi))^2\lambda_r(\Theta^*)^2} \cdot r\right) \\ \leq& \tilde{O}\left(\sigma r d\sqrt{T} + \sigma\sqrt{S_*^2 \frac{d^{0.5}\sigma^2}{\lambda_{\min}(Q(\pi))^2\lambda_r(\Theta^*)^2} T \cdot r}\right) \end{aligned}$$

with the optimal tuning of n_0.

I. Comparison between our algorithm and (Koltchinskii et al., 2011)

Suppose we are given $(A_i, y_i)_{i=1}^n$ iid samples such that $A_i \sim \Pi$ and Π is supported on $\{A : \|A\|_{\text{op}} \leq 1\}$, and for every i, $y_i = \langle \Theta^*, A_i \rangle + \eta_i$, where η_i's are independent zero-mean σ-subgaussian noise. (Koltchinskii et al., 2011) considers a nuclear-norm penalized estimator, defined as follows:

$$\hat{\Theta} = \arg\min_{\Theta} \|\Theta\|_{L_2(\Pi)}^2 - \langle \frac{2}{n}\sum_{i=1}^n y_i A_i, \Theta \rangle + \lambda\|\Theta\|_*, \tag{19}$$

where $\|B\|_{L_2(\Pi)} = \sqrt{\mathbb{E}_{A\sim\Pi}\langle A, B \rangle^2}$.

Theorem I.1 (Adapted from (Koltchinskii et al., 2011), Corollary 1)**.** *Given the setting above, and suppose additionally that:*

- *there exists $C > 0$ such that for all B, $\|B\|_{L_2(\Pi)}^2 \geq C\|B\|_F^2$,*
- *rank-r matrix Θ_0 is such that $\|\frac{1}{n}\sum_{i=1}^n A_i y_i - \mathbb{E}_{A\sim\Pi}\left[\langle \Theta_0, A \rangle A\right]\|_{\text{op}} \leq \frac{\lambda}{2}$.*

Then, there exists some absolute constant $c > 0$ such that

$$\|\hat{\Theta} - \Theta_0\|_F \leq c\frac{\sqrt{r}\lambda}{C}, \quad \|\hat{\Theta} - \Theta_0\|_F \leq c\frac{r\lambda}{C}.$$

Now the Lemma I.2 below states that Θ^* satisfies the condition of Θ_0 in Theorem I.1.

Lemma I.2. *Suppose $n \geq O(\ln\frac{d}{\delta})$. Then with probability $1 - \delta$,*

$$\left\|\frac{1}{n}\sum_{i=1}^n A_i y_i - \mathbb{E}_{A\sim\Pi}\left[\langle \Theta^*, A \rangle A\right]\right\|_{\text{op}} \leq O\left((S_* + \sigma)\sqrt{\frac{\ln\frac{d}{\delta}}{n}}\right)$$

Proof. Let $Z_i = A_i y_i - \mathbb{E}_{A\sim\Pi}\left[\langle\Theta^*, A\rangle A\right]$. We first upper bound $\|Z_i\|_{\text{op}}$'s ψ_2-Orlicz norm; to this end, first note that

$$\left\|\|A_i y_i\|_{\text{op}}\right\|_{\psi_2} \le \left\||y_i|\right\|_{\psi_2} \le \left\||\langle\Theta, A_i\rangle|\right\|_{\psi_2} + \left\||\eta_i|\right\|_{\psi_2} \le S_* + \sigma$$

Therefore, $\mathbb{E}_{A\sim\Pi}\left[\langle\Theta^*, A\rangle A\right] = \mathbb{E}[A_i y_i]$ also satisfies that

$$\left\|\|\,\mathbb{E}_{A\sim\Pi}\left[\langle\Theta^*, A\rangle A\right]\|_{\text{op}}\right\|_{\psi_2} \le S_* + \sigma,$$

hence $\|Z_i\|_{\psi_2} \le \left\|\|A_i y_i\|_{\text{op}}\right\|_{\psi_2} + \left\|\|\,\mathbb{E}_{A\sim\Pi}\left[\langle\Theta^*, A\rangle A\right]\|_{\text{op}}\right\|_{\psi_2} \le 2(S_* + \sigma)$.

Meanwhile,

$$\|\,\mathbb{E}[Z_i Z_i^\top]\|_{\text{op}} \le \|\,\mathbb{E}[A_i A_i^\top y_i^2]\|_{\text{op}} = \|\,\mathbb{E}[A_i A_i^\top(\langle\Theta^*, A_i\rangle^2 + \sigma^2)]\|_{\text{op}} \le S_*^2 + \sigma^2,$$

likewise,

$$\|\,\mathbb{E}[Z_i^\top Z_i]\|_{\text{op}} \le S_*^2 + \sigma^2.$$

Therefore, applying Proposition 2 of (Koltchinskii et al., 2011)[4] on $Z_1, \ldots, Z_n$, with $\sigma_Z = S_* + \sigma$, $\alpha = 2$, and $U_Z^{(\alpha)} = 2(S_* + \sigma)$, $t = \ln\frac{1}{\delta}$ gives that with probability $1-\delta$,

$$\left\|\frac{1}{n}\sum_{i=1}^n Z_i\right\|_{\text{op}} = O\left(\sigma_Z\sqrt{\frac{\ln\frac{d}{\delta}}{n}} + U_Z^{(\alpha)}\sqrt{\ln\frac{U_Z^{(\alpha)}}{\sigma_Z}}\frac{\ln\frac{d}{\delta}}{n}\right) \le O\left((S_* + \sigma)\sqrt{\frac{\ln\frac{d}{\delta}}{n}}\right).$$

□

Applying the theorem to $(A_i, y_i)_{i=1}^n$ with Θ_0 set to be Θ^*, where $A_i \sim \pi^*$ as defined in (10), we can choose $C = C_{\min}(\mathcal{A})$. On the other hand, Lemma I.2 below shows that choosing $\lambda = O\left((S_* + \sigma)\sqrt{\frac{\ln\frac{d}{\delta}}{n}}\right)$, with probability $1-\delta$, $\|\frac{1}{n}\sum_{i=1}^n A_i y_i - \mathbb{E}_{A\sim\Pi}\left[\langle\Theta_0, A\rangle A\right]\|_{\text{op}} \le \frac{\lambda}{2}$. Therefore, we conclude that with the above setting of λ and $\Pi = \pi^*$, the nuclear norm penalized estimator $\hat{\Theta}$ defined in Eq. (19) with satisfies that

$$\|\hat{\Theta} - \Theta^*\|_F \le \tilde{O}\left(\frac{S_* + \sigma}{C_{\min}(\mathcal{A})}\sqrt{\frac{r}{n}}\right), \quad \|\hat{\Theta} - \Theta^*\|_* \le \tilde{O}\left(\frac{S_* + \sigma}{C_{\min}(\mathcal{A})}\sqrt{\frac{r^2}{n}}\right).$$

J. Experimental details settings

J.1. Experiment settings

Common settings

- Computation resource: Apple M2 Pro, 16GB memory.
- Error bar: 1-standard deviation for the shadowed area.
- We attached our code as supplementary material and will upload a public link when this paper is accepted. Please read **README.md** file before running.

J.1.1. Figure 2 left

- Dimension $d_1 = d_2 = 3$
- Time steps: from 1000 to 10000, increased by 1000
- $\Theta^* = uv^\top$, where u and v are drawn from $\mathbb{S}^{d_1-1}$ and $\mathbb{S}^{d_2-1}$, respectively ($\mathbb{S}^{d-1}$ is the d-dimensional unit sphere.)
- Action set $\mathcal{A}$ is drawn uniformly at random from the $\mathcal{B}_{Frob}(1)$. $|\mathcal{A}| = 150$.
- Noise $\eta_t \sim N(0,1)$, which means $\sigma^2 = 1$.
- Repeated the experiment 60 times

[4]The original proposition statement is stated for the setting of $\sigma_Z^2 = \max(\mathbb{E}[Z_i Z_i^\top], \mathbb{E}[Z_i^\top Z_i])$ exactly; it can be checked that the proposition continues to hold when $\sigma_Z^2 \ge \max(\mathbb{E}[Z_i Z_i^\top], \mathbb{E}[Z_i^\top Z_i])$.

J.1.2. FIGURE 2 RIGHT

- Dimension $d_1 = d_2 = 3$
- Time steps: from 10000 to 100000, increased by 10000
- $\Theta^* = uv^\top$, where u and v are drawn from $\mathbb{S}^{d_1-1}$ and $\mathbb{S}^{d_2-1}$, respectively ($\mathbb{S}^{d-1}$ is the d-dimensional unit sphere.)
- Action set $\mathcal{A}$ is $\mathcal{A}_{hard}$, which is defined as follows:

$$a_i = \begin{cases} \mathsf{reshape}(\frac{1}{\sqrt{3}}e_1) & \text{if } i = 1 \\ \mathsf{reshape}(e_1 + \frac{1}{\sqrt{3}}e_i) & \text{if } i = 2, 3, \cdots, d_1 d_2 \end{cases}$$

- Noise $\eta_t \sim N(0, 1)$, which means $\sigma^2 = 1$.
- Repeated the experiment 60 times

J.1.3. FIGURE 3 LEFT

- Dimension $d_1 = d_2 = 5$
- Time steps: 100000
- $\Theta^* = uv^\top$, where u and v are drawn from $\mathbb{S}^{d_1-1}$ and $\mathbb{S}^{d_2-1}$, respectively ($\mathbb{S}^{d-1}$ is the d-dimensional unit sphere.)
- Action set $\mathcal{A}$ is drawn uniformly at random from the $\mathcal{B}_{Frob}(1)$. $|\mathcal{A}| = 100$.
- Noise $\eta_t \sim N(0, 1)$, which means $\sigma^2 = 1$.
- Repeated the experiment 60 times

J.1.4. FIGURE 3 RIGHT

- Dimension $d_1 = d_2 = 6$
- Time steps: 100000
- $\Theta^* = uv^\top$, where u and v are drawn from $\mathbb{S}^{d_1-1}$ and $\mathbb{S}^{d_2-1}$, respectively ($\mathbb{S}^{d-1}$ is the d-dimensional unit sphere.)
- Action set $\mathcal{A}$ is in bilinear setting. Which means, $\mathcal{A} = \{xz^\top : x \in \mathcal{X}, z \in \mathcal{Z}\}$ where $\mathcal{X}$ and $\mathcal{Z}$ are drawn uniformly at random from the $\mathbb{S}^{d_1-1}$ and $\mathbb{S}^{d_2-1}$, respectively. $|\mathcal{X}| = 4d_1 = 24$, $|\mathcal{Z}| = 4d_2 = 24$.
- Noise $\eta_t \sim N(0, 1)$, which means $\sigma^2 = 1$.
- Repeated the experiment 60 times

J.2. Algorithm for Left figures of Figure 3

Algorithm 6 Nuc-ETC (Nuclear norm regularized least square based Explore then commit)

Input: time horizon T, arm set $\mathcal{A}$, exploration lengths n_0^*, regularization parameter λ
Solve the optimization problem in Eq. (10) and denote the solution as π^*
for $t = 1, \ldots, n_0^*$ **do**
  Independently pull the arm A_t according to π^* and receives the reward Y_t
end for
$\hat{\Theta}_* := \arg\min_{\Theta \in \mathbb{R}^{d_1 \times d_2}} \frac{1}{2} \sum_{t=1}^{n_0^*} \left(\langle \Theta, A_t \rangle - Y_t \right)^2 + \lambda \|\Theta\|_*$
for $t = n_0^* + 1, \ldots, T$ **do**
  Pull the arm $X_t = \arg\max_{A \in \mathcal{A}} \langle \hat{\Theta}_*, A \rangle$
end for

J.2.1. THEORETICAL ANALYSIS OF THE EXPLORATION LENGTH n_0^*

As discussed in Appendix I, we have the following guarantee for the nuclear norm error bound of the nuclear norm regularized least square estimator:

$$\|\hat{\Theta} - \Theta^*\|_* \le \tilde{O}\left(\frac{S_* + \sigma}{C_{\min}(\mathcal{A})}\right) \sqrt{\frac{r^2}{n_0^*}}$$

Also, we have the following upper bound of the instantaneous regret after n_0^*:

$$\begin{aligned}\langle \Theta^*, A^* - A_t\rangle &= \langle \Theta^* - \hat{\Theta}, A^*\rangle + \langle \hat{\Theta}, A^*\rangle - \langle \Theta^*, A_t\rangle \\ &\leq \langle \Theta^* - \hat{\Theta}, A^*\rangle + \langle \hat{\Theta} - \Theta^*, A_t\rangle && \text{(Definition of } A_t) \\ &\leq \|\Theta^* - \hat{\Theta}\|_*(\|A^*\|_{\mathrm{op}} + \|A_t\|_{\mathrm{op}}) && \text{(Holder's inequality)} \\ &\leq 2\|\Theta^* - \hat{\Theta}\|_*\end{aligned}$$

Overall,the regret is

$$\begin{aligned}\mathsf{Reg}_T = \sum_{t=1}^{T} \langle \Theta^*, A^* - A_t\rangle &= \sum_{t=1}^{n_0^*} \langle \Theta^*, A^* - A_t\rangle + \sum_{t=n_0^*+1}^{T} \langle \Theta^*, A^* - A_t\rangle \\ &\leq S_* n_0^* + \tilde{O}\left(\frac{S_* + \sigma}{C_{\min}(\mathcal{A})}\right)\sqrt{\frac{r^2}{n_0^*}} \cdot (T - n_0^*)\end{aligned}$$

and the n_0^* which optimizes above value is $n_0^* = \left(\sigma^2 r^2 T^2 C_{\min}(\mathcal{A})^{-2} S_*^{-2}\right)^{1/3}$

J.3. Computational efficiency of Algorithm 1

For estimation only (Algorithm 1), we need $O(d_1^3 d_2^3)$ for matrix inversion (Eq. (2)), $O(n_0(d_1 d_2)^2)$ for estimators in Line 2, and $O(d_1^2 d_2)$ for SVD in Line 3 and 4, and no more computation is needed. On the other hand, (Koltchinskii et al., 2011) and other popular tools require optimizations that have several iterations dependent on the precision requirement of the optimization. For (Koltchinskii et al., 2011), it requires $O(n_0 d_1 d_2)$ for each iteration. In our experiment, both were very fast (ours: 0.3 sec, (Koltchinskii et al., 2011): 0.1 sec). For the experimental design part, no prior work explicitly studied on experimental design in the low-rank setting as far as we know. One natural approach is to optimize the conditions of the covariance matrix such as RIP, but there is no known computationally efficient way to directly compute these quantities (See the last part of the second contribution in Section 1). Other naive approaches are A/D/E/G/V-optimalities that are used in linear experimental design. They can be optimized by traditional optimization solvers like CVXPY or MOSEK. Our algorithm could also be done in the same way since our optimization problem is also convex. (in our experiment, ours: 0.046 sec, E-optimality: 0.039 sec).