Bilinear Bandits with Low-rank Structure

Kwang-Sung Jun [1] **Rebecca Willett** [2] **Stephen Wright** [3] **Robert Nowak** [3]

Abstract

We introduce the bilinear bandit problem with low-rank structure in which an action takes the form of a pair of arms from two different entity types, and the reward is a bilinear function of the known feature vectors of the arms. The unknown in the problem is a d_1 by d_2 matrix $\mathbf{\Theta}^*$ that defines the reward, and has low rank $r \ll \min\{d_1, d_2\}$. Determination of $\mathbf{\Theta}^*$ with this low-rank structure poses a significant challenge in finding the right exploration-exploitation trade-off. In this work, we propose a new two-stage algorithm called "Explore-Subspace-Then-Refine" (ESTR). The first stage is an explicit subspace exploration, while the second stage is a linear bandit algorithm called "almost-low-dimensional OFUL" (LowOFUL) that exploits and further refines the estimated subspace via a regularization technique. We show that the regret of ESTR is $\widetilde{\mathcal{O}}((d_1+d_2)^{3/2}\sqrt{rT})$ where $\widetilde{\mathcal{O}}$ hides logarithmic factors and T is the time horizon, which improves upon the regret of $\widetilde{\mathcal{O}}(d_1 d_2\sqrt{T})$ attained for a naïve linear bandit reduction. We conjecture that the regret bound of ESTR is unimprovable up to polylogarithmic factors, and our preliminary experiment shows that ESTR outperforms a naïve linear bandit reduction.

1 Introduction

Consider a drug discovery application where scientists would like to choose a (drug, protein) pair and measure whether the pair exhibits the desired interaction (Luo et al., 2017). Over many repetitions of this step, one would like to maximize the number of discovered pairs with the desired interaction. Similarly, an online dating service may want to choose a (female, male) pair from the user pool, match them, and receive feedback about whether they like each other or not. For clothing websites, the recommendation system may want to choose a pair of items (top, bottom) for a customer, whose appeal depends in part on whether they match. In these applications, the two types of entities are recommended and evaluated as a unit. Having feature vectors of the entities available,[1] the system must explore and learn what features of the two entities *jointly* predict positive feedback in order to make effective recommendations.

The recommendation system aims to obtain large rewards (the amount of positive feedback) but does not know ahead of time the relationship between the features and the feedback. The system thus faces two conflicting goals: choosing pairs that (i) maximally help estimate the relationship ("exploration") but which may give small rewards and (ii) return relatively large, but possibly suboptimal, rewards ("exploitation"), given the limited information obtained from the feedback collected so far. Such an exploration-exploitation dilemma can be formulated as a multi-armed bandit problem (Lai & Robbins, 1985; Auer et al., 2002). When the feature vectors are available for each arm, one can postulate simple reward structures such as (generalized) linear models to allow a large or even infinite number of arms (Auer, 2002; Dani et al., 2008; Abbasi-Yadkori et al., 2011; Filippi et al., 2010), a paradigm that has received much attention during the past decade, with such applications as online news recommendations (Li et al., 2010). Less is known for the situation we consider here, in which the recommendation (action) involves two different entity types and forms a bilinear structure. The closest work we are aware of is Kveton et al. (2017) whose action structure is the same as ours but without arm feature vectors. Factored bandits (Zimmert & Seldin, 2018) provide a more general view with L entity types rather than two, but they do not utilize arm features nor the low-rank structure. Our problem is different from dueling bandits (Yue et al., 2012a) or bandits with unknown user segment (Bhargava et al., 2017), which choose two arms from *the same* entity set rather than from two *different* entity types. Section 7 below contains detailed comparisons to related work.

This paper introduces the bilinear bandit problem with low-rank structure. In each round t, an algorithm chooses a left arm $\mathbf{x}_t$ from $\mathcal{X} \subseteq \mathbb{R}^{d_1}$ and a right arm $\mathbf{z}_t$ from $\mathcal{Z} \subseteq \mathbb{R}^{d_2}$, and

[1]Boston University [2]University of Chicago [3]University of Wisconsin-Madison. Correspondence to: Kwang-Sung Jun <kwangsungjun@gmail.com>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

[1] The feature vectors can be obtained either directly from the entity description (for example, hobbies or age) or by other preprocessing techniques (for example, embedding).

observes a noisy reward of a bilinear form:

$$y_t = \mathbf{x}_t^\top \boldsymbol{\Theta}^* \mathbf{z}_t + \eta_t \ , \tag{1}$$

where $\boldsymbol{\Theta}^* \in \mathbb{R}^{d_1 \times d_2}$ is an unknown parameter and η_t is a σ-sub-Gaussian random variable conditioning on $\mathbf{x}_t$, $\mathbf{z}_t$, and all the observations before (and excluding) time t. Denoting by r the rank of $\boldsymbol{\Theta}^*$, we assume that r is small ($r \ll \min\{d_1, d_2\}$), which means that the reward is governed by a few factors. Such low-rank appears in many recommendation applications (Ma et al., 2008). Our choice of reward model is popular and arguably natural; for example, the same model was used in Luo et al. (2017) for drug discovery.

The goal is to maximize the cumulative reward up to time T. Equivalently, we aim to minimize the cumulative regret:[2]

$$\text{Regret}_T = \sum_{t=1}^{T} \left\{ \max_{\mathbf{x} \in \mathcal{X}, \mathbf{z} \in \mathcal{Z}} \mathbf{x}^\top \boldsymbol{\Theta}^* \mathbf{z} - \mathbf{x}_t^\top \boldsymbol{\Theta}^* \mathbf{z}_t \right\} . \tag{2}$$

A naive approach to this problem is to reduce the bilinear problem to a linear problem, as follows:

$$\mathbf{x}^\top \boldsymbol{\Theta}^* \mathbf{z} = \langle \text{vec}(\mathbf{x}\mathbf{z}^\top), \text{vec}(\boldsymbol{\Theta}^*) \rangle \ . \tag{3}$$

Throughout the paper, we focus on the regime in which the numbers of possible actions $N_1 := |\mathcal{X}| \in \mathbb{N}_+ \cup \{\infty\}$ and $N_2 := |\mathcal{Z}| \in \mathbb{N}_+ \cup \{\infty\}$ are much larger than dimensions d_1 and d_2, respectively.[3] The reduction above allows us to use the standard linear bandit algorithms (see, for example, (Abbasi-Yadkori et al., 2011)) in the $d_1 d_2$-dimensional space and achieve regret of $\widetilde{\mathcal{O}}(d_1 d_2 \sqrt{T})$, where $\widetilde{\mathcal{O}}$ hides logarithmic factors. However, $d_1 d_2$ can be large, making this regret bound take an undesirably large value. Moreover, the regret does not decrease as r gets smaller, since the reduction hinders us from exploiting the low-rank structure.

We address the following challenge: Can we design an algorithm for the bilinear bandit problem that exploits the low-rank structure and enjoys regret strictly smaller than $\widetilde{\mathcal{O}}(d_1 d_2 \sqrt{T})$? We answer the question in the affirmative by proposing *Explore Subspace Then Refine* (ESTR), an approach that achieves a regret bound of $\widetilde{\mathcal{O}}((d_1 + d_2)^{3/2} \sqrt{rT})$. ESTR consists of two stages. In the first stage, we estimate the row and column subspace by randomly sampling from a subset of arms, chosen carefully. In the second stage, we leverage the estimated subspace by invoking an approach called *almost-low-dimensional OFUL* (LowOFUL), a variant of OFUL (Abbasi-Yadkori et al., 2011) that uses regularization to penalize the subspaces that are apparently *not* spanned by the rows and columns (respectively) of $\boldsymbol{\Theta}^*$. We conjecture that our regret upper bound is minimax optimal up to polylogarithmic factors based on the fact that the bilinear model has a much lower expected signal strength than the linear model. We provide a detailed argument on the lower bound in Section 5.

While the idea of having an explicit exploration stage, so-called Explore-Then-Commit (ETC), is not new, the way we exploit the subspace with LowOFUL is novel for two reasons. First, the standard ETC commits to the estimated parameter without refining and is thus known to have $\mathcal{O}(\sqrt{T})$ regret only for "smooth" arm sets such as the unit ball (Rusmevichientong & Tsitsiklis, 2010; Abbasi-Yadkori et al., 2009). This means that the estimate refining is necessary for generic arm sets. Second, after the first stage that outputs a subspace estimate, it is tempting to project all the arms onto the identified subspaces (r dimensions for each row and column space), and naively invoke OFUL in the r^2-dimensional space. However, the subspace mismatch invalidates the upper confidence bound used in OFUL; i.e., the confidence bound does not actually bound the mean reward.

Attempts to correct the confidence bound so that it is faithful are not trivial, and we are unaware of a solution that leads to improved regret bounds. Departing from completely committing to the identified subspaces, LowOFUL works with the full $d_1 d_2$-dimensional space, but penalizes the subspace that is *complementary to the estimated subspace*, thus continuing to *refine* the subspace. We calibrate the amount of regularization to be a function of the subspace estimation error; this is the key to achieving our final regret bound.

We remark that our bandit problem can be modified slightly for the setting in which the arm $\mathbf{z}_t$ is considered as a context, obtained from the environment. This situation arises, for example, in recommendation systems where $\mathcal{Z}$ is the set of users represented by indicator vectors (i.e., $d_2 = N_2$) and $\mathcal{X}$ is the set of items. Such a setting is similar to Cesa-Bianchi et al. (2013), but we assume that $\boldsymbol{\Theta}^*$ is low-rank rather than knowing the graph information. Furthermore, when the user information is available, one can take $\mathcal{Z}$ as the set of user feature vectors.

The paper is structured as follows. In Section 2, we define the problem formally and provide a sketch of the main contribution. Sections 3 and 4 describe the details of stages 1 and 2 of ESTR, respectively. We elaborate our conjecture on the regret lower bound in Section 5. After presenting our preliminary experimental results in Section 6, we discuss related work in Section 7 and propose future research directions in Section 8.

[2] This regret definition is actually called *pseudo* regret; we refer to Bubeck & Cesa-Bianchi (2012, Section 1) for detail.

[3] Otherwise, one can reduce the problem to the standard K-armed bandit problem and enjoy regret of $\widetilde{\mathcal{O}}(\sqrt{N_1 N_2 T})$. With SupLinRel (Auer, 2002), one may also achieve $\widetilde{\mathcal{O}}(\sqrt{d_1 d_2 T \log(N_1 N_2)})$, but this approach wastes a lot of samples and does not allow an infinite number of arms.

Input: time horizon T, the exploration length T_1, the rank r of $\mathbf{\Theta}^*$, and the spectral bounds S_F, S_2, and S_r of $\mathbf{\Theta}^*$.

Stage 1 (Section 3)

- Solve (approximately)
$$\arg \max_{\text{distinct } \mathbf{x}^{(1)},\dots,\mathbf{x}^{(d_1)} \in \mathcal{X}} \left(\text{the smallest eigenvalue of } \left[\mathbf{x}^{(1)},\dots,\mathbf{x}^{(d_1)}\right]\right) \tag{4}$$
and define $\mathbf{X} = \{\mathbf{x}^{(1)},\cdots,\mathbf{x}^{(d_1)}\}$. Define $\mathbf{Z}$ similarly.
- For T_1 rounds, choose a pair of arms from $\mathbf{X} \times \mathbf{Z}$, pulling each pair the same number of times to the extent possible. That is, choose each pair $\lfloor \frac{T_1}{d_1 d_2} \rfloor$ times, then choose $T_1 - d_1 d_2 \lfloor \frac{T_1}{d_1 d_2} \rfloor$ pairs uniformly at random without replacement.
- Let $\widetilde{\mathbf{K}}$ be a matrix such that $\widetilde{K}_{ij}$ is the average reward of pulling the arm $(\mathbf{x}^{(i)}, \mathbf{z}^{(j)})$. Invoke a noisy matrix recovery algorithm (e.g., OptSpace (Keshavan et al., 2010)) with $\widetilde{\mathbf{K}}$ and the rank r to obtain an estimate $\widehat{\mathbf{K}}$.
- Let $\widehat{\mathbf{\Theta}} = \mathbf{X}^{-1}\widehat{\mathbf{K}}(\mathbf{Z}^\top)^{-1}$ where $\mathbf{X} = [(\mathbf{x}^{(1)})^\top; \cdots; (\mathbf{x}^{(d_1)})^\top] \in \mathbb{R}^{d_1 \times d_1}$ (abusing notation) and $\mathbf{Z}$ is defined similarly.
- Let $\widehat{\mathbf{\Theta}} = \widehat{\mathbf{U}}\widehat{\mathbf{S}}\widehat{\mathbf{V}}^\top$ be the SVD of $\widehat{\mathbf{\Theta}}$. Let $\widehat{\mathbf{U}}_\perp$ and $\widehat{\mathbf{V}}_\perp$ be orthonormal bases of the complementary subspaces of $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{V}}$, respectively.
- Let $\gamma(T_1)$ be the subspace angle error bound such that, with high probability,
$$\|\widehat{\mathbf{U}}_\perp^\top \mathbf{U}^*\|_F \|\widehat{\mathbf{V}}_\perp^\top \mathbf{V}^*\|_F \le \gamma(T_1) \tag{5}$$
where $\mathbf{\Theta}^* = \mathbf{U}^*\mathbf{S}^*\mathbf{V}^{*\top}$ is the SVD of $\mathbf{\Theta}^*$.

Stage 2 (Section 4)

- Rotate the arm sets: $\mathcal{X}' = \left\{[\widehat{\mathbf{U}}\widehat{\mathbf{U}}_\perp]^\top \mathbf{x} : \mathbf{x} \in \mathcal{X}\right\}$ and $\mathcal{Z}' = \left\{[\widehat{\mathbf{V}}\widehat{\mathbf{V}}_\perp]^\top \mathbf{z} : \mathbf{z} \in \mathcal{Z}\right\}$.
- Define a vectorized arm set so that the last $(d_1 - r)\cdot(d_2 - r)$ components are from the complementary subspaces:
$$\mathcal{A} = \left\{[\mathrm{vec}(\mathbf{x}_{1:r}\mathbf{z}_{1:r}^\top); \mathrm{vec}(\mathbf{x}_{r+1:d_1}\mathbf{z}_{1:r}^\top); \mathrm{vec}(\mathbf{x}_{1:r}\mathbf{z}_{r+1:d_2}^\top); \mathrm{vec}(\mathbf{x}_{r+1:d_1}\mathbf{z}_{r+1:d_2}^\top)] \in \mathbb{R}^{d_1 d_2} : \mathbf{x} \in \mathcal{X}',\ \mathbf{z} \in \mathcal{Z}'\right\}\ .$$
- For $T_2 = T - T_1$ rounds, invoke LowOFUL with the arm set $\mathcal{A}$, the low dimension $k = (d_1 + d_2)r - r^2$, and $\gamma(T_1)$.

Figure 1. A sketch of Explore Subspace Then Refine (ESTR)

2 Preliminaries

We define the problem formally as follows. Let $\mathcal{X} \subseteq \mathbb{R}^{d_1}$ and $\mathcal{Z} \subseteq \mathbb{R}^{d_2}$ be the left and right arm space, respectively. Define $N_1 = |\mathcal{X}|$ and $N_2 = |\mathcal{Z}|$. (Either or both can be infinite.) We assume that both the left and right arms have Euclidean norm at most 1: $\|\mathbf{x}\|_2 \le 1$ and $\|\mathbf{z}\|_2 \le 1$ for all $\mathbf{x} \in \mathcal{X}$ and $\mathbf{z} \in \mathcal{Z}$. Without loss of generality, we assume $\mathcal{X}$ ($\mathcal{Z}$) spans the whole d_1 (d_2) dimensional space (respectively) since, if not, one can project the arm set to a lower-dimensional space that is now fully spanned.[4] We assume $d_2 = \Theta(d_1)$ and define $d = \max\{d_1, d_2\}$. If A is a positive integer, we use notation $[A] = \{1, 2, \dots, A\}$. We denote by $\mathbf{v}_{i:j}$ the $(j - i + 1)$-dimensional vector taking values from the coordinates from i to j from $\mathbf{v}$. Similarly, we define $\mathbf{M}_{i:j,k:\ell} \in \mathbb{R}^{(j-i+1)\times(\ell-k+1)}$ to be a submatrix taking values from $\mathbf{M}$ with the row indices from i to j and the column indices from k to ℓ. We denote by v_i the i-th component of the vector $\mathbf{v}$ and by M_{ij} the entry of a matrix $\mathbf{M}$ located at the i-th row and j-th column. Denote by $\Sigma_k(\mathbf{M})$ the k-th largest singular value, and define $\Sigma_{\max}(\mathbf{M}) = \Sigma_1(\mathbf{M})$. Let $\Sigma_{\min}(\mathbf{M})$ be the smallest nonzero singular value of $\mathbf{M}$. $|\mathbf{M}|$ denotes the determinant of a matrix $\mathbf{M}$.

[4] In this case, we effectively work with a projected version of $\mathbf{\Theta}^*$, and its rank may become smaller as well.

The protocol of the bilinear bandit problem is as follows. At time t, the algorithm chooses a pair of arms $(\mathbf{x}_t, \mathbf{z}_t) \in \mathcal{X} \times \mathcal{Z}$ and receives a noisy reward y_t according to (1). We make the standard assumptions in linear bandits: the Frobenius and operator norms of $\mathbf{\Theta}^*$ are bounded by known constants, $\|\mathbf{\Theta}^*\|_F \le S_F$ and $\|\mathbf{\Theta}^*\|_2 \le S_2$,[5] and the sub-Gaussian scale σ of η_t is known to the algorithm. We denote by s_i^* the i-th largest singular value of $\mathbf{\Theta}^*$. We assume that the rank r of the matrix is known and that $s_r^* \ge S_r$ for some known $S_r > 0$. [6]

The main contribution of this paper is the first nontrivial upper bound on the achievable regret for the bilinear bandit problem. In this section, we provide a sketch of the overall result and the key insight. For simplicity, we omit constants and variables other than d, r, and T. Our proposed ESTR algorithm enjoys the following regret bound, which strictly improves the naive linear bandit reduction when $r \ll d$.

Theorem 1 (An informal version of Corollary 2)**.** *Under mild assumptions, the regret of ESTR is $\widetilde{\mathcal{O}}(d^{3/2}\sqrt{rT})$ with high probability.*

[5] When S_2 is not known, one can set $S_2 = S_F$. In some applications, S_2 is known. For example, the binary model $y_t \sim \text{Bernoulli}((\mathbf{x}_t^\top \mathbf{\Theta}^* \mathbf{z}_t) + 1)/2)$, we can evidently set $S_2 = 1$.

[6] In practice, one can perform rank estimation after the first stage (see, for example, Keshavan et al. (2010)).

We conjecture that the regret bound above is minimax optimal up to polylogarithmic factors since the expected signal strength in the bilinear model is much weaker than the linear model. We elaborate on this argument in Section 5.

We describe ESTR in Figure 1. The algorithm proceeds in two stages. In the first stage, we estimate the column and row subspace of $\mathbf{\Theta}^*$ from noisy rank-one measurements using a matrix recovery algorithm. Specifically, we first identify d_1 and d_2 arms from the set $\mathcal{X}$ and $\mathcal{Z}$ in such a way that the smallest singular values of the matrices formed from these arms are maximized approximately (see (4)), which is a form of submatrix selection problem (details in Section 3). We emphasize that finding the exact solution is not necessary here since Theorem 1 has a mild dependency on the smallest eigenvalue found when approximating (4). We then use the popular matrix recovery algorithm, OptSpace (Keshavan et al., 2010) to estimate $\mathbf{\Theta}^*$. The $\sin\Theta$ theorem of Wedin (Stewart & Sun, 1990) is used to convert the matrix recovery error bound from OptSpace to the desired subspace angle guarantee (5) with $\gamma(T_1) = \mathcal{O}\left(\frac{d^3 r}{T_1}\right)$. The regret incurred in stage 1 is bounded trivially by $T_1 \|\mathbf{\Theta}^*\|_2$.

In the second stage, we transform the problem into a $d_1 d_2$-dimensional linear bandit problem and invoke LowOFUL that we introduce in Section 4. This technique projects the arms onto both the estimated subspace and its complementary subspace and uses $\gamma(T_1)$ to penalize weights in the complementary subspaces $\widehat{\mathbf{U}}_\perp$ and $\widehat{\mathbf{V}}_\perp$. LowOFUL enjoys regret bound $\widetilde{\mathcal{O}}((dr + \sqrt{T}\gamma(T_1))\sqrt{T - T_1})$ during $T - T_1$ rounds. By combining with the regret for the first stage, we obtain an overall regret of

$$T_1 + \left(dr + \sqrt{T}\frac{d^3 r}{T_1}\right)\sqrt{T}.$$

Choosing T_1 to minimize this expression, we obtain a regret bound of $\widetilde{\mathcal{O}}(d^{3/2}\sqrt{rT})$.

3 Stage 1: Subspace estimation

The goal of stage 1 is to estimate the row and column subspaces for the true parameter $\mathbf{\Theta}^*$. How should we choose which arm pairs to pull, and what guarantee can we obtain on the subspace estimation error? One could choose to apply a noisy matrix recovery algorithm with affine rank minimization (Recht et al., 2010; Mohan & Fazel, 2010) to the measurements attained from the arm pulls. However, these methods require the measurements to be Gaussian or Rademacher, so their guarantees depend on satisfaction of a RIP property (Recht et al., 2010), or, for rank-one projection measurements, an RUB property (Cai et al., 2015). Such assumptions are not suitable for our setting since measurements are restricted to the arbitrarily given arm sets $\mathcal{X}$ and $\mathcal{Z}$. Uniform sampling from the arm set cannot guarantee RIP, as the arm set itself can be heavily biased in certain directions.

We design a simple reduction procedure though matrix recovery with noisy entry observations, leaving a more sophisticated treatment as future work. The d_1 arms in $\mathcal{X}$ are chosen according to the criterion (4), which is a combinatorial problem that is hard to solve exactly. Our analysis does not require its exact solution, however; it is enough that the objective value is nonzero (that is, the matrix $\mathbf{X}$ constructed from these d_1 arms is nonsingular). (Similar comments hold for the matrix $\mathbf{Z}$.) We remark that the problem (4) is shown to be NP-hard by Çivril & Magdon-Ismail (2009) and is related to finding submatrices with favorable spectral properties (Çivril & Magdon-Ismail, 2007; Tropp, 2009), but a thorough review on algorithms and their limits is beyond the scope of the paper. For our experiments, simple methods such as random selection were sufficient; we describe our implementation in the supplementary material.

If $\mathbf{K}^*$ is the matrix defined by $K^*_{ij} = \mathbf{x}^{(i)\top}\mathbf{\Theta}^*\mathbf{z}^{(j)}$, each time step of stage 1 obtains a noisy estimate of one element of $\mathbf{K}^*$. Since multiple measurements of each entry are made, in general, we compute average measurements for each entry. A matrix recovery algorithm applied to this matrix of average measurements yields the estimate $\widehat{\mathbf{K}}$ of the rank-r matrix $\mathbf{K}^*$. Since $\mathbf{K}^* = \mathbf{X}\mathbf{\Theta}^*\mathbf{Z}^\top$, we estimate $\mathbf{\Theta}^*$ by $\widehat{\mathbf{\Theta}} = \mathbf{X}^{-1}\widehat{\mathbf{K}}(\mathbf{Z}^\top)^{-1}$ and then compute the subspace estimate $\widehat{\mathbf{U}} \in \mathbb{R}^{d_1 \times r}$ and $\widehat{\mathbf{V}} \in \mathbb{R}^{d_2 \times r}$ by applying SVD to $\widehat{\mathbf{\Theta}}$.

We choose the recovery algorithm OptSpace by Keshavan et al. (2010) because of its strong (near-optimal) guarantee. Denoting the SVD of $\mathbf{K}^*$ by $\mathbf{U}\mathbf{R}\mathbf{V}^\top$, we use the matrix incoherence definition from Keshavan et al. (2010) and let (μ_0, μ_1) be the smallest values such that for all $i \in [d_1], j \in [d_2]$,

$$\sum_{k=1}^{r} U_{ik}^2 \le \mu_0 r/d_1, \quad \sum_{k=1}^{r} V_{jk}^2 \le \mu_0 r/d_2, \quad \text{and}$$

$$\left|\sum_{k=1}^{r} U_{ik}(\Sigma_k(\mathbf{K}^*)/\Sigma_{\max}(\mathbf{K}^*))V_{jk}\right| \le \mu_1\sqrt{\frac{r}{d_1 d_2}}.$$

Define the condition number $\kappa = \Sigma_{\max}(\mathbf{K}^*)/\Sigma_{\min}(\mathbf{K}^*)$. We present the guarantee of OptSpace (Keshavan et al., 2010) in a paraphrased form. (The proof of this result, and all subsequent proofs, are deferred to the supplementary material.)

Theorem 2. *There exists a constant C_0 such that for $T_1 \ge C_0\sigma^2(\mu_0^2 + \mu_1^2)\frac{\kappa^6}{\Sigma_{\min}(\mathbf{K}^*)^2} dr(r + \log d)$, we have that, with probability at least $1 - 2/d_2^3$,*

$$\|\widehat{\mathbf{K}} - \mathbf{K}^*\|_F \le C_1\kappa^2\sigma\frac{d^{3/2}\sqrt{r}}{\sqrt{T_1}} \tag{6}$$

where C_1 is an absolute constant.

The original theorem from Keshavan et al. (2010) assumes $T_1 \le d_1 d_2$ and does not allow repeated sampling. However,

we show in the proof that the same guarantee holds for $T_1 > d_1 d_2$ since repeated sampling of entries has the effect of reducing the noise parameter σ.

Our recovery of an estimate $\widehat{\mathbf{K}}$ of $\mathbf{K}^*$ implies the bound $\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_F \le \|\mathbf{X}^{-1}\|_2 \|\mathbf{Z}^{-1}\|_2 \tau$ where τ is the RHS of (6). However, our goal in stage 1 is to obtain bounds on the subspace estimation errors. That is, given the SVDs $\widehat{\boldsymbol{\Theta}} = \widehat{\mathbf{U}}\widehat{\mathbf{S}}\widehat{\mathbf{V}}^\top$ and $\boldsymbol{\Theta}^* = \mathbf{U}^*\mathbf{S}^*\mathbf{V}^{*\top}$, we wish to identify how close $\widehat{\mathbf{U}}$ ($\widehat{\mathbf{V}}$) is to $\mathbf{U}^*$ ($\mathbf{V}^*$ respectively). Such guarantees on the subspace error can be obtained via the $\sin\Theta$ theorem by Stewart & Sun (1990), which we restate in our supplementary material. Roughly, this theorem bounds the canonical angles between two subspaces by the Frobenius norm of the difference between the two matrices. Recall that s_r^* is the r-th largest singular value of $\boldsymbol{\Theta}^*$.

Theorem 3. *Suppose we invoke OptSpace to compute $\widehat{\mathbf{K}}$ as an estimate of the matrix $\mathbf{K}^*$. After stage 1 of ESTR with T_1 satisfying the condition of Theorem 2, we have, with probability at least $1 - 2/d_2^3$,*

$$\|\widehat{\mathbf{U}}_\perp^\top \mathbf{U}^*\|_F \|\widehat{\mathbf{V}}_\perp^\top \mathbf{V}^*\|_F \le \frac{\|\mathbf{X}^{-1}\|_2^2 \|\mathbf{Z}^{-1}\|_2^2}{(s_r^*)^2} \tau^2 \qquad (7)$$

where $\tau = C_1 \kappa^2 \sigma d^{3/2} \sqrt{r}/\sqrt{T_1}$.

4 Stage 2: Almost-low-dimensional linear bandits

The goal of stage 2 is to exploit the subspaces $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{V}}$ estimated in stage 1 to perform efficient bandit learning. At first, it is tempting to project all the left and right arms to r-dimensional subspaces using $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{V}}$, respectively, which seems to be a bilinear bandit problem with an r by r unknown matrix. One can then reduce it to an r^2-dimensional linear bandit problem and solve it by standard algorithms such as OFUL (Abbasi-Yadkori et al., 2011). Indeed, if $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{V}}$ *exactly* span the row and column spaces of $\boldsymbol{\Theta}^*$, this strategy yields a regret bound of $\widetilde{\mathcal{O}}(r^2\sqrt{T})$. In reality, these matrices (subspaces) are not exact, so there is model mismatch, making it difficult to apply standard regret analysis. The upper confidence bound (UCB) used in popular algorithms becomes invalid, and there is no known correction that leads to a regret bound lower than $\widetilde{\mathcal{O}}(d_1 d_2 \sqrt{T})$, to the best of our knowledge.

In this section, we show how stage 2 of our approach avoids the mismatch issue by returning to the full $d_1 d_2$-dimensional space, allowing the subspace estimates to be inexact, but penalizing those components that are complementary to $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{V}}$. This effectively constrains the hypothesis space to be much smaller than the full $d_1 d_2$-dimensional space. We show how the bilinear bandit problem with good subspace estimates can be turned into the *almost low-dimensional linear bandit problem*, and how much penalization / regularization is needed to achieve a low overall regret bound.

Finally, we state our main theorem showing the overall regret bound of ESTR.

Reduction to linear bandits. Recall that $\boldsymbol{\Theta}^* = \mathbf{U}^*\mathbf{S}^*\mathbf{V}^{*\top}$ is the SVD of $\boldsymbol{\Theta}^*$ (where $\mathbf{S}^*$ is $r \times r$ diagonal) and that $\widehat{\mathbf{U}}_\perp$ and $\widehat{\mathbf{V}}_\perp$ are the complementary subspace of $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{V}}$ respectively. Let $\mathbf{M} = [\widehat{\mathbf{U}}\ \widehat{\mathbf{U}}_\perp]^\top \boldsymbol{\Theta}^* [\widehat{\mathbf{V}}\ \widehat{\mathbf{V}}_\perp]$ be a rotated version of $\boldsymbol{\Theta}^*$. Then we have

$$\boldsymbol{\Theta}^* = [\widehat{\mathbf{U}}\ \widehat{\mathbf{U}}_\perp]\mathbf{M}[\widehat{\mathbf{V}}\ \widehat{\mathbf{V}}_\perp]^\top \quad \text{and}$$
$$\mathbf{x}^\top \boldsymbol{\Theta}^* \mathbf{z} = ([\widehat{\mathbf{U}}\ \widehat{\mathbf{U}}_\perp]^\top \mathbf{x})^\top \mathbf{M} ([\widehat{\mathbf{V}}\ \widehat{\mathbf{V}}_\perp]^\top \mathbf{z}) .$$

Thus, the bilinear bandit problem with the unknown $\boldsymbol{\Theta}^*$ with arm sets $\mathcal{X}$ and $\mathcal{Z}$ is equivalent to the one with the unknown $\mathbf{M}$ with arm sets $\mathcal{X}' = \{\mathbf{x}' = [\widehat{\mathbf{U}}\ \widehat{\mathbf{U}}_\perp]^\top \mathbf{x} \mid \mathbf{x} \in \mathcal{X}\}$ and $\mathcal{Z}'$ (defined similarly). As mentioned earlier, this problem can be cast as a $d_1 d_2$-dimensional linear bandit problem by considering the unknown vector $\boldsymbol{\theta}^* = \mathrm{vec}(\mathbf{M})$. The difference is, however, that we have learnt something about the subspace in stage 1. We define $\boldsymbol{\theta}^*$ to be a rearranged version of $\mathrm{vec}(\mathbf{M})$ so that the last $(d_1 - r)\cdot(d_2 - r)$ dimensions of $\boldsymbol{\theta}^*$ are M_{ij} for $i \in \{r+1, \dots, d_1\}$ and $j \in \{r+1, \dots, d_2\}$; that is, letting $k := d_1 d_2 - (d_1 - r)\cdot(d_2 - r)$,

$$\boldsymbol{\theta}^*_{1:k} = [\mathrm{vec}(\mathbf{M}_{1:r,1:r});\ \mathrm{vec}(\mathbf{M}_{r+1:d_1,1:r});\ \mathrm{vec}(\mathbf{M}_{1:r,r+1:d_2})],$$
$$\boldsymbol{\theta}^*_{k+1:p} = \mathrm{vec}(\mathbf{M}_{r+1:d_1,r+1:d_2}) .$$

Then we have

$$\|\boldsymbol{\theta}^*_{k+1:p}\|_2^2 = \sum_{i>r \wedge j>r} M_{ij}^2 = \|\widehat{\mathbf{U}}_\perp^\top (\mathbf{U}^*\mathbf{S}^*\mathbf{V}^{*\top}) \widehat{\mathbf{V}}_\perp\|_F^2 \qquad (9)$$
$$\le \|\widehat{\mathbf{U}}_\perp^\top \mathbf{U}^*\|_F^2 \|\mathbf{S}^*\|_2^2 \|\widehat{\mathbf{V}}_\perp^\top \mathbf{V}^*\|_F^2 ,$$

which implies $\|\boldsymbol{\theta}^*_{k+1:p}\|_2 = \mathcal{O}(d^3 r/T_1)$ by Theorem 3. Our knowledge on the subspace results in the knowledge of the norm of certain coordinates! Can we exploit this knowledge to enjoy a better regret bound than $\widetilde{\mathcal{O}}(d_1 d_2 \sqrt{T})$? We answer this question in the affirmative below.

Almost-low-dimensional OFUL (LowOFUL). We now focus on an abstraction of the conversion described in the previous paragraph, which we call the *almost-low-dimensional linear bandit problem*. In the standard linear bandit problem in p dimensions, the player chooses an arm $\mathbf{a}_t$ at time t from an arm set $\mathcal{A} \subseteq \mathbb{R}^p$ and observes a noisy reward $y_t = \langle \mathbf{a}_t, \boldsymbol{\theta}^* \rangle + \eta_t$, where the noise η_t has the same properties as in (1). We assume that $\|\mathbf{a}\|_2 \le 1$ for all $\mathbf{a} \in \mathcal{A}$, and $\|\boldsymbol{\theta}^*\|_2 \le B$ for some known constant $B > 0$. In almost-low-dimensional linear bandits, we have additional knowledge that $\|\boldsymbol{\theta}^*_{k+1:p}\|_2 \le B_\perp$ for some index k and some constant $B_\perp$ (ideally $\ll B$). This means that all-but-k dimensions of $\boldsymbol{\theta}^*$ are close to zero.

To exploit the extra knowledge on the unknown, we propose *almost-low-dimensional OFUL* (LowOFUL) that extends the standard linear bandit algorithm OFUL (Abbasi-Yadkori et al., 2011). To describe OFUL, define the design matrix $\mathbf{A} \in \mathbb{R}^{t \times p}$ with rows $\mathbf{a}_s^\top$, $s = 1, 2, \dots, t$ and the vector of rewards $\mathbf{y} = [y_1, \dots, y_t]^\top$. The key estimator is based on regression with the standard squared ℓ_2-norm regularizer, as

Algorithm 1 LowOFUL

Input: T, k, the arm set $\mathcal{A} \subseteq \mathbb{R}^p$, failure rate δ, and positive constants B, $B_\perp$, λ, $\lambda_\perp$.
Set $\mathbf{\Lambda} = \mathrm{diag}(\lambda, \dots, \lambda, \lambda_\perp, \dots, \lambda_\perp)$ where λ occupies the first k diagonal entries.
for $t = 1, 2, \dots, T$ **do**
  Compute $\mathbf{a}_t = \arg\max_{\mathbf{a}\in\mathcal{A}} \max_{\boldsymbol{\theta}\in c_{t-1}} \langle \boldsymbol{\theta}, \mathbf{a}\rangle$.
  Pull arm $\mathbf{a}_t$.
  Receive reward y_t.
  Set c_t as (12).
end for

follows:

$$\widehat{\boldsymbol{\theta}}_t = \arg\min_{\boldsymbol{\theta}} \frac{1}{2}\|\mathbf{A}\boldsymbol{\theta} - \mathbf{y}\|_2^2 + \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2 = (\lambda\mathbf{I} + \mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top\mathbf{y}\,. \tag{10}$$

OFUL then defines a confidence ellipsoid around $\widehat{\boldsymbol{\theta}}_t$ based on which one can compute an upper confidence bound on the mean reward of any arm. In our variant, we allow a different regularization for each coordinate, replacing the regularizer $\frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2$ by $\frac{1}{2}\|\boldsymbol{\theta}\|_{\mathbf{\Lambda}}^2 = \frac{1}{2}\boldsymbol{\theta}^\top\mathbf{\Lambda}\boldsymbol{\theta}$ for some positive diagonal matrix $\mathbf{\Lambda}$. Specifically, we define $\mathbf{\Lambda} = \mathrm{diag}(\lambda, \dots, \lambda, \lambda_\perp, \dots, \lambda_\perp)$, where λ occupies the first k diagonal entries and $\lambda_\perp$ the last $p-k$ positions. With this modification, the estimator becomes

$$\widehat{\boldsymbol{\theta}}_t = \arg\min_{\boldsymbol{\theta}} \frac{1}{2}\|\mathbf{A}\boldsymbol{\theta} - \mathbf{y}\|_2^2 + \frac{1}{2}\|\boldsymbol{\theta}\|_{\mathbf{\Lambda}}^2 = (\mathbf{\Lambda} + \mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top\mathbf{y}\,. \tag{11}$$

Define $\mathbf{V}_t = \mathbf{\Lambda} + \sum_{s=1}^t \mathbf{a}_t\mathbf{a}_t^\top = \mathbf{\Lambda} + \mathbf{A}^\top\mathbf{A}$ and let δ be the failure rate we are willing to endure. The confidence ellipsoid for $\boldsymbol{\theta}^*$ is

$$c_t = \left\{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_t\|_{\mathbf{V}_t} \le \sqrt{\beta_t}\right\} \quad \text{where}$$
$$\sqrt{\beta_t} = \sigma\sqrt{\log\frac{|\mathbf{V}_t|}{|\mathbf{\Lambda}|\delta^2}} + \sqrt{\lambda}B + \sqrt{\lambda_\perp}B_\perp. \tag{12}$$

This ellipsoid enjoys the following guarantee, which is a direct consequence of Valko et al. (2014, Lemma 3) that is based on the self-normalized martingale inequality of Abbasi-Yadkori et al. (2011, Theorem 1).

Lemma 1. *With probability at least $1-\delta$, we have $\boldsymbol{\theta}^* \in c_t$ for all $t \ge 1$.*

We summarize LowOFUL in Algorithm 1, where $\max_{\boldsymbol{\theta}\in c_{t-1}}\langle\boldsymbol{\theta}, \mathbf{a}\rangle$ can be simplified to $\langle\widehat{\boldsymbol{\theta}}_{t-1}, \mathbf{a}\rangle + \sqrt{\beta_{t-1}}\|\mathbf{a}\|_{\mathbf{V}_{t-1}^{-1}}$.

We now state the regret bound of LowOFUL in Theorem 4, which is based on the standard linear bandit regret analysis dating back to Auer (2002).

Theorem 4. *The regret of LowOFUL is, with probability at least $1-\delta$,*

$$\mathcal{O}\left(\sqrt{\log\frac{|\mathbf{V}_T|}{|\mathbf{\Lambda}|}}\left(\sigma\sqrt{\log\frac{|\mathbf{V}_T|}{|\mathbf{\Lambda}|\delta^2}} + \sqrt{\lambda}B + \sqrt{\lambda_\perp}B_\perp\right)\cdot\sqrt{T}\right). \tag{13}$$

In the standard linear bandit setting where $\lambda_\perp = \lambda$ and $B_\perp = B$, we recover the regret bound $\widetilde{\mathcal{O}}(p\sqrt{T})$ of OFUL, since $\log\frac{|\mathbf{V}_T|}{|\mathbf{\Lambda}|} = \mathcal{O}(p\sqrt{T})$ (Abbasi-Yadkori et al., 2011, Lemma 10).

To alleviate the dependence on p in the regret bound, we propose a carefully chosen value of $\lambda_\perp$ in the following corollary.

Corollary 1. *Then, the regret of LowOFUL with $\lambda_\perp = \frac{T}{k\log(1+\frac{T}{\lambda})}$ is, with probability at least $1-\delta$,*

$$\widetilde{\mathcal{O}}\left((\sigma k + \sqrt{k\lambda}B + \sqrt{T}B_\perp)\sqrt{T}\right).$$

The bound improves the dependence on dimensionality from p to k, but introduces an extra factor of $\sqrt{T}$ to $B_\perp$, resulting in linear regret. While this choice is not interesting in general, we show that it is useful for our case: Since $\|\boldsymbol{\theta}^*_{k+1:p}\|_2 = \mathcal{O}(1/T_1)$, we can set $B_\perp = \mathcal{O}(1/T_1)$ to be a valid upper bound of $\|\boldsymbol{\theta}^*_{k+1:p}\|_2$. By setting $T_1 = \Theta(\sqrt{T})$, the regret bound in Corollary 1 scales with $\sqrt{T}$ rather than T.

Concretely, using (9), we set

$$B = S_F \quad \text{and} \quad B_\perp = S_2\cdot\gamma(T_1) \quad \text{where}$$
$$\gamma(T_1) = \frac{\|\mathbf{X}^{-1}\|_2^2\|\mathbf{Z}^{-1}\|_2^2}{(S_r)^2}\cdot C_1^2\left(\frac{S_2}{S_r}\right)^4\sigma^2\cdot\frac{d^3r}{T_1}\,. \tag{14}$$

B and $B_\perp$ are valid upper bounds of $\|\boldsymbol{\theta}^*\|_2$ and $\|\boldsymbol{\theta}^*_{k+1:p}\|_2$, respectively, with high probability. Note we must use S_2, S_r, and S_2/S_r instead of s_1^*, s_r^*, and κ, respectively, since the latter variables are unknown to the learner.

Overall regret. Theorem 5 shows the overall regret bound of ESTR.

Theorem 5. *Suppose we run ESTR (Algorithm 1) with $T_1 \ge C_0\sigma^2(\mu_0^2+\mu_1^2)\frac{\kappa^6}{\Sigma_{\min}(\mathbf{K}^*)^2}dr(r+\log d)$. We invoke LowOFUL in stage 2 with $p = d_1d_2$, $k = r\cdot(d_1+d_2-r)$, $\boldsymbol{\theta}^*$ defined as (8), the rotated arm sets $\mathcal{X}'$ and $\mathcal{Z}'$, $\lambda_\perp = \frac{T_2}{k\log(1+T_2/\lambda)}$, and B and $B_\perp$ as in (14). The regret of ESTR is, with probability at least $1-\delta-2/d_2^3$,*

$$\widetilde{\mathcal{O}}\left(s_1^*T_1 + T\cdot\frac{\|\mathbf{X}^{-1}\|_2^2\|\mathbf{Z}^{-1}\|_2^2(S_2^5/S_r^6)\sigma^2d^3r}{T_1}\right).$$

One can see that there exists an optimal choice of T_1, which we state in the following corollary.

Corollary 2. *Suppose the assumptions in Theorem 5 hold. If $T_1 = \Theta\left(\|\mathbf{X}^{-1}\|_2\|\mathbf{Z}^{-1}\|_2\frac{S_2^2}{S_r^3}\sigma d^{3/2}\sqrt{rT}\right)$, then the regret of*

ESTR is, with probability at least $1 - \delta - 2/d_2^3$,

$$\widetilde{\mathcal{O}}\left(\frac{S_2^3}{S_r^3}\|\mathbf{X}^{-1}\|_2\|\mathbf{Z}^{-1}\|_2\sigma d^{3/2}\sqrt{rT}\right).$$

Note that, for our problem, the incoherence constants μ_0 and μ_1 do not play an important role with large enough T.

Remark One might notice that we can also regularize the submatrices $\mathbf{M}_{r+1:d_1,1:r}$ and $\mathbf{M}_{1:r,r+1:d_2}$ since they are coming partly from the complementary subspace of $\widehat{\mathbf{U}}$ and partly from the complement of $\widehat{\mathbf{V}}$ (but not both). In practice, such a regularization can be done to reduce the regret slightly, but it does not affect the order of the regret. We do not have sufficient decrease in the magnitude to provide interesting bounds. One can show that, while $\|\mathbf{M}_{r+1:d_1,r+1:d_2}\|_F^2 = \mathcal{O}(1/T_1)$, the quantities $\|\mathbf{M}_{1:r,r+1:d_2}\|_F^2$ and $\|\mathbf{M}_{r+1:d_1,1:r}\|_F^2$ are $\mathcal{O}(1/\sqrt{T_1})$.

5 Lower bound

A simple lower bound is $\Omega(d\sqrt{T})$, since when the arm set $\mathbf{Z}$ is a singleton the problem reduces to a d_1-dimensional linear bandit problem. We have attempted to extend existing lower-bound proof techniques in Rusmevichientong & Tsitsiklis (2010), Dani et al. (2008), and Lattimore & Szepesvári (2018), but the bilinear nature of the problem introduces cross terms between the left and right arm, which are difficult to deal with in general. However, we conjecture that the lower bound is $\Omega(d^{3/2}\sqrt{rT})$. We provide an informal argument below that the dependence on d must be $d^{3/2}$ based on the observation that the rank-one bilinear reward model's signal-to-noise ratio (SNR) is significantly worse than that of the linear reward model.

Consider a rank-one $\mathbf{\Theta}^*$ that can be decomposed as $\mathbf{u}\mathbf{v}^\top$ for some $\mathbf{u}, \mathbf{v} \in \{\pm 1/\sqrt{d}\}^d$. Suppose the left and right arm sets are $\mathcal{X} = \mathcal{Z} = \{\pm 1/\sqrt{d}\}^d$. Let us choose $\mathbf{x}_t$ and $\mathbf{z}_t$ uniformly at random (which is the sort of pure exploration that must be performed initially). Then a simple calculation shows that the expected squared signal strength with such a random choice is $\mathbb{E}|\mathbf{x}_t^\top\mathbf{\Theta}^*\mathbf{z}_t|^2 = \frac{1}{d^2}$. In contrast, the expected squared signal strength for a linear reward model is $\mathbb{E}|\mathbf{x}_t^\top\mathbf{u}|^2 = \frac{1}{d}$. The effect of this is analogous to increasing the sub-Gaussian scale parameter of the noise η_t by a factor of $\sqrt{d}$. We thus conjecture that the $\sqrt{d}$ difference in the SNR introduces the dependence $d^{3/2}$ in the regret rather than d.

6 Experiments

We present a preliminary experimental result and discuss practical concerns.

Bandits in practice requires tuning the exploration rate to perform well, which is usually done by adjusting the confidence bound width (Chapelle & Li, 2011; Li et al., 2010; Zhang et al., 2016), which amounts to replacing β_t with $c\beta_t$ for some $c > 0$ for OFUL or its variants (including LowOFUL). An efficient parameter tuning in bandits is an open problem and is beyond our scope. For the sake of comparison, we tune c by grid search and report the result with the smallest average regret. For ESTR, the value of T_1 used in the proof involves some unknown constants; to account for this, we tune T_1 by grid search. We consider the following methods:

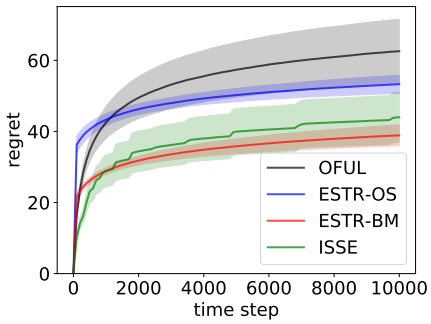

Figure 2. Simulation results for $d = 8$ and $r = 1$. Our method ESTR-OS, its variant ESTR-BM, and an implicit exploration variant of ESTR called ISSE all outperform the baseline linear bandit method OFUL.

- OFUL: The OFUL reduction described in (3), which ignores the low-rank structure.
- ESTR-OS: Our proposed method; we simplify $B_\perp$ in (14) to $S_2\sigma^2 d^3 r/T_1$.
- ESTR-BM: We replace OptSpace with the Burer-Monteiro formulation and perform the alternating minimization (Burer & Monteiro, 2003).
- ISSE (Implicit SubSpace Exploration): LowOFUL with a heuristic subspace estimation that avoids an explicit exploration stage. We split the time intervals with knots at $t \in \{10^{0.5}, 10^1, 10^{1.5}, \ldots\}$. At the beginning time t' of each interval, we perform the matrix recovery with the Burer-Monteiro formulation using all the past data, estimate the subspaces, and use them to initialize LowOFUL with $B_\perp = S_2\sigma^2 d^3 r/t'$ and all the past data.

Note that OFUL and ISSE only require tuning c whereas ESTR methods require tuning both c and T_1.

We run our simulation with $d_1 = d_2 = 8$, $r = 1$, $\sigma = 0.01$. We set $\lambda = 1$ for both OFUL and LowOFUL. We draw 16 arms from the unit sphere for each arm set $\mathcal{X}$ and $\mathcal{Z}$ and simulate the bandit game for $T = 10^4$ iterations, which we repeat 60 times for each method. Figure 2 plots the average regret of the methods and the .95 confidence intervals. All the methods outperform OFUL, and the regret differences from OFUL are statistically significant. We observe that ESTR-BM performs better than ESTR-OS. We believe this is due to our limit on the number of iterations of OptSpace set to

1000, which we imposed due to its slow convergence in our experiments.[7] The Burer-Monteiro formulation, however, converged within 200 iterations. Finally, ISSE performs close to ESTR-BM, but with a larger variance. Although ISSE does not have a theoretical guarantee, it does not require tuning T_1 and performs better than OFUL.

7 Related work

There exist a few studies on pulling a pair of arms as a unit action, as we do. Kveton et al. (2017) consider the K-armed bandit with N_1 left arms and N_2 right arms. The expected rewards can be represented as a matrix $\bar{\mathbf{R}} \in \mathbb{R}^{N_1 \times N_2}$ where the authors assume $\bar{\mathbf{R}}$ has rank $r \ll \min\{N_1, N_2\}$. The main difference from our setting is that they do not assume that the arm features are available, so our work is related to Kveton et al. (2017) in the same way as the linear bandits are related to K-armed bandits. The problem considered in Katariya et al. (2017b) is essentially a rank-one version of Kveton et al. (2017), which is motivated by a click-feedback model called position-based model with N_1 items and N_2 positions. This work is further extended to have a tighter KL-based bound by Katariya et al. (2017a). All these studies successfully exploit the low-rank structure to enjoy regret bounds that scale with $r(N_1 + N_2)$ rather than $N_1 N_2$. Zimmert & Seldin (2018) propose a more generic problem called factored bandits whose action set is a product of atomic L action sets rather than two. While they achieve generality by not require to know the explicit reward model, factored bandits do not leverage the known arm features nor the low-rank structure, resulting in large regret in our problem.

There are other works that exploit the low-rank structure of the reward matrix, although the action is just a single arm pull. Sen et al. (2017) consider the contextual bandit setting where there are N_1 discrete contexts and N_2 arms, but do not take into account the observed features of contexts or arms. Under the so-called separability assumption, the authors make use of Hottopix algorithm to exploit the low-rank structure. Gopalan et al. (2016) consider a similar setting, but employ the robust tensor power method for recovery. Kawale et al. (2015) study essentially the same problem, but make assumptions on the prior that generates the unknown matrix and perform online matrix factorization with particle filtering to leverage the low-rank structure. These studies also exploit the low-rank structure successfully and enjoy regret bounds that scale much better than $N_1 N_2$.

There has been a plethora of contextual bandit studies that exploit structures other than the low-rank-ness, where the context is usually the user identity or features. For example, Gentile et al. (2014) and its followup studies (Li et al., 2016; Gentile et al., 2017) leverage the clustering structure of the contexts. In Cesa-Bianchi et al. (2013) and Vaswani et al. (2017), a graph structure of the users is leveraged to enjoy regret bound that is lower than running bandits on each context (i.e., user) independently. Deshmukh et al. (2017) introduce a multitask learning view and exploit arm similarity information via kernels, but their regret guarantee is valid only when the similarity is known ahead of time. In this vein, if we think of the right arm set $\mathcal{Z}$ as tasks, we effectively assume different parameters for each task but with a low-rank structure. That is, the parameters can be written as a linear combination of a few hidden factors, which are estimated on the fly rather than being known in advance. Johnson et al. (2016) consider low-rank structured bandits but in a different setup. Their reward model has expected reward of the form $\mathrm{tr}(\mathbf{X}_t^\top \mathbf{\Theta}^*)$ with the arm $\mathbf{X}_t \in \mathbb{R}^{d \times p}$ and the unknown $\mathbf{\Theta}^* \in \mathbb{R}^{d \times p}$. While $\mathbf{X}_t$ corresponds to $\mathbf{x}_t \mathbf{z}_t^\top$ in our setting, they consider a continuous arm set only, so their algorithm cannot be applied to our problem.

Our subroutine LowOFUL is quite similar to SpectralUCB of (Valko et al., 2014), which is designed specifically for graph-structured arms in which expected rewards of the two arms are close to each other (i.e., "smooth") when there is an edge between them. Although technical ingredients for Corollary 1 stem from Valko et al. (2014), LowOFUL is for an inherently different setup in which we *design* the regularization matrix $\mathbf{\Lambda}$ to maximally exploit the subspace knowledge and minimize the regret, rather than receiving $\mathbf{\Lambda}$ from the environment as a part of the problem definition. Gilton & Willett (2017) study a similar regularizer in the context of sparse linear bandits under the assumption that a superset of the sparse locations is known ahead of time. Yue et al. (2012b) consider a setup similar to LowOFUL. They assume an estimate of the subspace is available, but their regret bound still depends on the total dimension p.

8 Conclusion

In this paper, we introduced the bilinear low-rank bandit problem and proposed the first algorithm with a nontrivial regret guarantee. Our study opens up several future research directions. First, there is currently no nontrivial lower bound, and showing whether the regret of $\widetilde{\mathcal{O}}(d^{3/2}\sqrt{rT})$ is tight or not remains open. Second, while our algorithm improves the regret bound over the trivial linear bandit reduction, the algorithm requires to tune an extra parameter T_1. It would be more natural to continuously update the subspace estimate and the amount of regularization, just like ISSE. However, proving a theoretical guarantee would be challenging since most matrix recovery algorithms require some sort of uniform sampling with a "nice" set of measurements. We speculate that one can employ randomized arm selection and use importance-weighted data to perform effective and provable matrix recoveries on-the-fly.

[7] We used the authors' C implementation that is wrapped in python (`https://github.com/strin/pyOptSpace`).

Acknowledgements

This work was supported by NSF 1447449 - IIS, NIH 1 U54 AI117924-01, NSF CCF-0353079, NSF 1740707, and AFOSR FA9550-18-1-0166.

References

Abbasi-Yadkori, Y., Antos, A., and Szepesvári, C. Forced-exploration based algorithms for playing in stochastic linear bandits. In *COLT Workshop on On-line Learning with Limited Feedback*. Citeseer, 2009.

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

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

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning*, 47(2–3):235–256, 2002.

Bhargava, A., Ganti, R., and Nowak, R. Active Positive Semidefinite Matrix Completion: Algorithms, Theory and Applications. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 54, pp. 1349–1357, 2017.

Bubeck, S. and Cesa-Bianchi, N. Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. *Foundations and Trends in Machine Learning*, 5: 1–122, 2012.

Burer, S. and Monteiro, R. D. C. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming, Series B*, 2003.

Cai, T. T., Zhang, A., and Others. ROP: Matrix recovery via rank-one projections. *The Annals of Statistics*, 43(1): 102–138, 2015.

Çivril, A. and Magdon-Ismail, M. Finding maximum Volume sub-matrices of a matrix. *RPI Comp Sci Dept TR*, pp. 7–8, 2007.

Çivril, A. and Magdon-Ismail, M. On selecting a maximum volume sub-matrix of a matrix and related problems. *Theoretical Computer Science*, 410(47):4801, 2009.

Cesa-Bianchi, N., Gentile, C., and Zappella, G. A Gang of Bandits. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 737–745. 2013.

Chapelle, O. and Li, L. An Empirical Evaluation of Thompson Sampling. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2249–2257, 2011.

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.

Deshmukh, A. A., Dogan, U., and Scott, C. Multi-Task Learning for Contextual Bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4848–4856. 2017.

Filippi, S., Cappe, O., Garivier, A., and Szepesvári, C. Parametric Bandits: The Generalized Linear Case. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 586–594. 2010.

Gentile, C., Li, S., and Zappella, G. Online Clustering of Bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 32, pp. 757–765, 2014.

Gentile, C., Li, S., Kar, P., Karatzoglou, A., Zappella, G., and Etrue, E. On Context-Dependent Clustering of Bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 70, pp. 1253–1262, 2017.

Gilton, D. and Willett, R. Sparse linear contextual bandits via relevance vector machines. In *International Conference on Sampling Theory and Applications (SampTA)*, pp. 518–522, 2017.

Gopalan, A., Maillard, O.-A., and Zaki, M. Low-rank Bandits with Latent Mixtures. *arXiv:1609.01508 [cs.LG]*, 2016.

Johnson, N., Sivakumar, V., and Banerjee, A. Structured Stochastic Linear Bandits. *arXiv:1606.05693 [stat.ML]*, 2016.

Jun, K.-S., Bhargava, A., Nowak, R., and Willett, R. Scalable Generalized Linear Bandits: Online Computation and Hashing. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 99–109. 2017.

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, 2017a.

Katariya, S., Kveton, B., Szepesvári, C., Vernade, C., and Wen, Z. Stochastic Rank-1 Bandits. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 392–401, 2017b.

Kawale, J., Bui, H. H., Kveton, B., Tran-Thanh, L., and Chawla, S. Efficient Thompson Sampling for Online Matrix-Factorization Recommendation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1297–1305. 2015.

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

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

Lai, T. L. and Robbins, H. Asymptotically Efficient Adaptive Allocation Rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

Lattimore, T. and Szepesvári, C. Bandit Algorithms. 2018. URL http://downloads.tor-lattimore.com/book.pdf.

Li, L., Chu, W., Langford, J., and Schapire, R. E. A Contextual-Bandit Approach to Personalized News Article Recommendation. *Proceedings of the International Conference on World Wide Web (WWW)*, pp. 661–670, 2010.

Li, S., Karatzoglou, A., and Gentile, C. Collaborative Filtering Bandits. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 539–548, 2016.

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.

Ma, H., Yang, H., Lyu, M. R., and King, I. SoRec: Social Recommendation Using Probabilistic Matrix Factorization. *Proceeding of the ACM Conference on Information and Knowledge Mining (CIKM)*, 2008.

Mohan, K. and Fazel, M. New restricted isometry results for noisy low-rank recovery. In *IEEE International Symposium on Information Theory - Proceedings*, 2010.

Recht, B., Fazel, M., and Parrilo, P. A. Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization. *SIAM Rev.*, 52(3):471–501, 2010.

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

Sen, R., Shanmugam, K., Kocaoglu, M., Dimakis, A., and Shakkottai, S. Contextual Bandits with Latent Confounders: An NMF Approach. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 54, pp. 518–527, 2017.

Stewart, G. W. and Sun, J.-g. *Matrix Perturbation Theory*. Academic Press, 1990.

Tropp, J. A. Column subset selection, matrix factorization, and eigenvalue optimization. In *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 978–986. Society for Industrial and Applied Mathematics, 2009.

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.

Vaswani, S., Schmidt, M., and Lakshmanan, L. V. S. Horde of Bandits using Gaussian Markov Random Fields. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 690–699, 2017.

Yue, Y., Broder, J., Kleinberg, R., and Joachims, T. The K-armed dueling bandits problem. In *Journal of Computer and System Sciences*, 2012a.

Yue, Y., Hong, S. A. S., and Guestrin, C. Hierarchical exploration for accelerating contextual bandits. *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 1895–1902, 2012b.

Zhang, L., Yang, T., Jin, R., Xiao, Y., and Zhou, Z.-h. Online Stochastic Linear Optimization under One-bit Feedback. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 48, pp. 392–401, 2016.

Zimmert, J. and Seldin, Y. Factored Bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2840–2849, 2018.

Supplementary Material

A Proof of Theorem 2

Theorem 2 (Restated) *There exists a constant C_0 such that for $T_1 \geq C_0\sigma^2(\mu_0^2+\mu_1^2)\frac{\kappa^6}{\Sigma_{\min}(\mathbf{K}^*)^2}dr(r+\log d)$, we have that, with probability at least $1-2/d_2^3$,*

$$\|\widehat{\mathbf{K}}-\mathbf{K}^*\|_F \leq C_1\kappa^2\sigma\frac{d^{3/2}\sqrt{r}}{\sqrt{T_1}} \tag{15}$$

where C_1 is an absolute constant.

Proof. There are a number of assumptions required for the guarantee of OptSpace to hold. Given a noise matrix $\mathbf{Z}$, let $\widetilde{\mathbf{K}} = \mathbf{K}^* + \mathbf{Z}$ be the noisy observation of matrix $\mathbf{K}^*$. Among various noise models in Keshavan et al. (2010, Theorem 1.3), the independent ρ-sub-Gaussian model fits our problem setting well. Let $E \in [m]\times[n]$ be the indicator of observed entries and let $\mathbf{Z}^E$ be a censored version of $\mathbf{Z}$ in which the unobserved entries are zeroed out. Recall that we assume $d_2 = \Theta(d_1)$, that $d = \max\{d_1, d_2\}$, and that κ is the condition number of $\mathbf{K}^*$.

We first state the guarantee and then describe the required technical assumptions. Keshavan et al. (2010, Theorem 1.2) states that the following is true for some constant $C' > 0$:

$$\|\widehat{\mathbf{K}}-\mathbf{K}^*\|_F \leq C'\kappa^2\frac{d^2\sqrt{r}}{|E|}\|\mathbf{Z}^E\|_2 \,.$$

Here, by Keshavan et al. (2010, Theorem 1.3), $\|\mathbf{Z}^E\|_2$ is no larger than $C''\rho\sqrt{|E|/d}$, for some constant $C'' > 0$, under Assumption **(A3)** below, where ρ is the sub-Gaussian scale parameter for the noise $\mathbf{Z}$. (ρ can be different from σ, as we explain below). The original version of the statement has a preprocessed version $\|\tilde{\mathbf{Z}}^E\|_2$ rather than $\|\mathbf{Z}^E\|_2$, but they are the same under our noise model, according to Keshavan et al. (2010, Section 1.5). Together, in our notation, we have

$$\|\widehat{\mathbf{K}}-\mathbf{K}^*\|_F \leq C'C''\kappa^2\rho\frac{d^{3/2}\sqrt{r}}{\sqrt{|E|}}.$$

In the case of $T_1 < d_1d_2$, the guarantee above holds true with $\rho = \sigma$ and $|E| = T_1$. If $T_1 \geq d_1d_2$, the guarantee holds true with $\rho = \sigma\cdot\left(\left\lfloor\frac{T_1}{d_1d_2}\right\rfloor\right)^{-1/2}$ and $|E| = d_1d_2$. In both cases, we arrive at (6).

We now state the conditions. Let $\alpha = d_1/d_2$. Define $\Sigma_{\min}$ to be the smallest nonzero singular values of $\mathbf{M}$.

- **(A1)**: $\mathbf{M}$ is (μ_0,μ_1)-incoherent. Note $\mu_0 \in [1, \max\{d_1,d_2\}/r]$.
- **(A2)**: (Sufficient observation) For some $C' > 0$, we have
$$|E| \geq C'd_2\sqrt{\alpha}\kappa^2\max\left\{\mu_0 r\sqrt{\alpha}\log d_2,\ \mu_0^2r^2\alpha\kappa^4,\ \mu_1^2r^2\alpha\kappa^4\right\}\,,$$
which we loosen and simplify to (using $\mu_0 \geq 1$)
$$|E| \geq C'\kappa^6(\mu_0^2+\mu_1^2)dr(r+\log d).$$
- **(A3)**: $|E| \geq n\log n$.
- **(A4)**: We combine the bound on $\|\mathbf{Z}^E\|_2$ and the condition in Keshavan et al. (2010, Theorem 1.2) that says "provided that the RHS is smaller than $\sigma_{\min}$", which results in requiring
$$\frac{|E|}{\rho^2} \geq C''\frac{\kappa^4}{\Sigma_{\min}^2}dr,$$
for some $C'' > 0$, Using the same logic as before, either $T_1 < d_1d_2$ or $T_1 \geq d_1d_2$, we can rewrite the same statement in terms of T_1, as follows:
$$T_1 \geq C''\sigma^2\frac{\kappa^4}{\Sigma_{\min}^2}dr\,.$$

All these conditions can be merged to

$$T_1 \geq C_0\sigma^2(\mu_0^2+\mu_1^2)\frac{\kappa^6}{\Sigma_{\min}^2}dr(r+\log d)$$

for some constant $C_0 > 0$. □

B Proof of Theorem 3

We first restate the $\sin\Theta$ theorem due to Wedin (Stewart & Sun, 1990) in a slightly simplified form. Let the SVDs of matrices $\mathbf{A}$ and $\tilde{\mathbf{A}}$ be defined as follows:

$$(\mathbf{U}_1\ \mathbf{U}_2\ \mathbf{U}_3)^\top \mathbf{A} (\mathbf{V}_1\ \mathbf{V}_2) = \begin{pmatrix} \mathbf{\Sigma}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_2 \\ \mathbf{0} & \mathbf{0} \end{pmatrix},$$

$$(\tilde{\mathbf{U}}_1\ \tilde{\mathbf{U}}_2\ \tilde{\mathbf{U}}_3)^\top \tilde{\mathbf{A}} (\tilde{\mathbf{V}}_1\ \tilde{\mathbf{V}}_2) = \begin{pmatrix} \tilde{\mathbf{\Sigma}}_1 & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{\Sigma}}_2 \\ \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

Let $\mathbf{R} = \mathbf{A}\tilde{\mathbf{V}}_1 - \tilde{\mathbf{U}}_1\tilde{\mathbf{\Sigma}}_1$ and $\mathbf{S} = \mathbf{A}^\top\tilde{\mathbf{U}}_1 - \tilde{\mathbf{V}}_1\tilde{\mathbf{\Sigma}}_1$, and define $\mathbf{U}_{1\perp} = [\mathbf{U}_2\ \mathbf{U}_3]$ and $\mathbf{V}_{1\perp} = [\mathbf{V}_2\ \mathbf{V}_3]$. Wedin's $\sin\Theta$ theorem, roughly speaking, bounds the sin canonical angles between two matrices by the Frobenius norm of their difference.

Theorem 6 (Wedin)**.** *Suppose that there is a number $\delta > 0$ such that*

$$\min_{i,j} \left|\sigma_i\left(\tilde{\mathbf{\Sigma}}_1\right) - \sigma_j\left(\mathbf{\Sigma}_2\right)\right| \geq \delta \quad \text{and} \quad \min_i \sigma_i\left(\tilde{\mathbf{\Sigma}}_1\right) \geq \delta\ .$$

Then,

$$\sqrt{\|\mathbf{U}_{1\perp}^\top\tilde{\mathbf{U}}_1\|_F^2 + \|\mathbf{V}_{1\perp}^\top\tilde{\mathbf{V}}_1\|_F^2} \leq \frac{\sqrt{\|\mathbf{R}\|_F^2 + \|\mathbf{S}\|_F^2}}{\delta}$$

We now prove Theorem 3.

Theorem 3 (Restated) *Suppose we invoke OptSpace to compute $\widehat{\mathbf{K}}$ as an estimate of the matrix $\mathbf{K}^*$. After stage 1 of ESTR with T_1 satisfying the condition of Theorem 2, we have, with probability at least $1 - 2/d_2^3$,*

$$\|\widehat{\mathbf{U}}_\perp^\top\mathbf{U}^*\|_F\|\widehat{\mathbf{V}}_\perp^\top\mathbf{V}^*\|_F \leq \frac{\|\mathbf{X}^{-1}\|_2^2\|\mathbf{Z}^{-1}\|_2^2}{(s_r^*)^2}\tau^2 \tag{16}$$

where $\tau = C_1\kappa^2\sigma d^{3/2}\sqrt{r}/\sqrt{T_1}$.

Proof. In our case, the variables defined for Wedin's theorem are as follows:

$$\mathbf{A} = \widehat{\mathbf{\Theta}} \qquad \tilde{\mathbf{A}} = \mathbf{\Theta}^*$$
$$\mathbf{U}_1 = \widehat{\mathbf{U}} \qquad \tilde{\mathbf{U}}_1 = \mathbf{U}^*$$
$$\mathbf{\Sigma}_1 = \widehat{\mathbf{S}} \qquad \tilde{\mathbf{\Sigma}}_1 = \mathbf{S}^*$$
$$\mathbf{V}_1 = \widehat{\mathbf{V}} \qquad \tilde{\mathbf{V}}_1 = \mathbf{V}^*\ .$$

Let $\mathbf{E} = \widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*$. Note that

$$\mathbf{R} = \widehat{\mathbf{\Theta}}\mathbf{V}^* - \mathbf{U}^*\mathbf{S}^* = (\mathbf{\Theta}^* + \mathbf{E})\mathbf{V}^* - \mathbf{U}^*\mathbf{S}^* = \mathbf{E}\widehat{\mathbf{V}}$$
$$\mathbf{S} = -\mathbf{E}^\top\mathbf{U}^* \quad \text{(similarly)}\ .$$

Then, $\|\mathbf{R}\|_F = \|\mathbf{E}\widehat{\mathbf{V}}\|_F \leq \|\mathbf{E}\|_F$, using the fact that

$$\|\mathbf{E}\|_F = \|\mathbf{E}[\widehat{\mathbf{V}}\ \widehat{\mathbf{V}}_\perp]\|_F = \sqrt{\|\mathbf{E}\widehat{\mathbf{V}}\|_F^2 + \|\mathbf{E}\widehat{\mathbf{V}}_\perp\|_F^2}.$$

Similarly, $\|\mathbf{S}\|_F \leq \|\mathbf{E}\|_F$. We now apply the $\sin\Theta$ theorem to obtain

$$\sqrt{2\|\widehat{\mathbf{U}}_\perp^\top\mathbf{U}^*\|_F\|\widehat{\mathbf{V}}_\perp^\top\mathbf{V}^*\|_F} \leq \sqrt{\|\widehat{\mathbf{U}}_\perp^\top\mathbf{U}^*\|_F^2 + \|\widehat{\mathbf{V}}_\perp^\top\mathbf{V}^*\|_F^2} \leq \frac{\sqrt{\|\mathbf{R}\|_F^2 + \|\mathbf{S}\|_F^2}}{\delta} \leq \frac{\sqrt{2\|\mathbf{E}\|_F^2}}{s_r^*}$$

where the first inequality follows from the Young's inequality. To summarize, we have

$$\|\widehat{\mathbf{U}}_\perp^\top\mathbf{U}^*\|_F\|\widehat{\mathbf{V}}_\perp^\top\mathbf{V}^*\|_F \leq \frac{\|\mathbf{\Theta}^* - \widehat{\mathbf{\Theta}}\|_F^2}{s_r^{*2}}\ . \tag{17}$$

Theorem 2 and the inequality $\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_F \leq \|\mathbf{X}^{-1}\|_2\|\mathbf{Z}^{-1}\|_2\|\widehat{\mathbf{K}} - \mathbf{K}^*\|_F$ conclude the proof. □

C Proof of Lemma 1

This lemma is a direct consequence of Valko et al. (2014, Lemma 3). We just need to characterize the constant C therein that upper bounds $\|\boldsymbol{\theta}^*\|_{\boldsymbol{\Lambda}}$. The observation that

$$\|\boldsymbol{\theta}^*\|_{\boldsymbol{\Lambda}} \leq \sqrt{\lambda\|\boldsymbol{\theta}_{1:k}\|_2^2 + \lambda_\perp\|\boldsymbol{\theta}_{k+1:p}\|_2^2} \leq \sqrt{\lambda}S + \sqrt{\lambda_\perp}S_\perp$$

completes the proof.

D Proof of Theorem 4

Let r_t be the instantaneous pseudo-regret at time t: $r_t = \langle\boldsymbol{\theta}^*, \mathbf{a}^*\rangle - \langle\boldsymbol{\theta}^*, \mathbf{a}_t\rangle$. The assumptions $\|\mathbf{a}_t\|_2 \leq 1$ and $\|\boldsymbol{\theta}^*\|_2 \leq B$ imply that $r_t \leq 2B$. Using the fact that the OFUL ellipsoidal confidence set contains $\boldsymbol{\theta}^*$ w.p. $\geq 1-\delta$, one can show that $r_t \leq 2\sqrt{\beta_t}\|\mathbf{x}_t\|_{\mathbf{V}_{t-1}^{-1}}$ as shown in Abbasi-Yadkori et al. (2011, Theorem 3). Then, using the monotonicity of β_t,

$$\begin{aligned}
r_t &\leq \min\{2B, 2\sqrt{\beta_t}\|\mathbf{x}_t\|_{\mathbf{V}_{t-1}^{-1}}\} \leq \min\{2B, 2\sqrt{\beta_T}\|\mathbf{x}_t\|_{\mathbf{V}_{t-1}^{-1}}\} \\
&= 2\sqrt{\beta_T}\min\{B/\sqrt{\beta_T}, \|\mathbf{x}_t\|_{\mathbf{V}_{t-1}^{-1}}\} \\
\implies \sum_{t=1}^T r_t &\overset{(a)}{\leq} 2\sqrt{\beta_T}\sqrt{T\sum_{t=1}^T \min\{B^2/\beta_T, \|\mathbf{x}_t\|_{\mathbf{V}_{t-1}^{-1}}^2\}} \\
&\overset{(b)}{\leq} 2\sqrt{\beta_T}\sqrt{T\max\{2, B^2/\beta_T\}\sum_{t=1}^T \log\left(1 + \|\mathbf{x}_t\|_{\mathbf{V}_{t-1}^{-1}}^2\right)} \\
&\overset{(c)}{\leq} 2\sqrt{\max\{2, 1/\lambda\}}\sqrt{\beta_T}\sqrt{\log\frac{|\mathbf{V}_T|}{|\boldsymbol{\Lambda}|}}\sqrt{T}
\end{aligned}$$

where (a) is due to Cauchy-Schwarz, (b) is due to $\min\{a, x\} \leq \max\{2, a\}\log(1+x), \forall a, x > 0$ (see Jun et al. (2017, Lemma 3)), and (c) is by $\sum_{t=1}^T \log(1 + \|\mathbf{x}_t\|_{\mathbf{V}_{t-1}^{-1}}^2) = \log\frac{|\mathbf{V}_T|}{|\boldsymbol{\Lambda}|}$ (see Abbasi-Yadkori et al. (2011, Lemma 11)) and $\beta_T \geq (\sqrt{\lambda}B + \sqrt{\lambda_\perp}B_\perp)^2 \geq \lambda B^2$.

E Proof of Corollary 1

The lemma from Valko et al. (2014) characterizes how large $\log\frac{|\mathbf{V}_T|}{|\boldsymbol{\Lambda}|}$ can be.

Lemma 2. *(Valko et al., 2014, Lemma 5) For any T, let $\boldsymbol{\Lambda} = diag([\lambda_1, \ldots, \lambda_p])$.*

$$\log\frac{|\mathbf{V}_T|}{|\boldsymbol{\Lambda}|} \leq \max\sum_{i=1}^p \log\left(1 + \frac{t_i}{\lambda_i}\right)$$

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

We specify a desirable value of $\lambda_\perp$ in the following lemma.

Lemma 3. *If $\lambda_\perp = \frac{T}{k\log\left(1+\frac{T}{\lambda}\right)}$, then*

$$\log\frac{|\mathbf{V}_T|}{|\boldsymbol{\Lambda}|} \leq 2k\log\left(1 + \frac{T}{\lambda}\right)$$

Proof. Inheriting the setup of Lemma 2,

$$\begin{aligned}
\log\frac{|\mathbf{V}_T|}{|\boldsymbol{\Lambda}|} &\leq \max\sum_{i=1}^p \log(1 + \frac{t_i}{\lambda_i}) \\
&\leq k\log(1 + \frac{T}{\lambda}) + \sum_{i=k+1}^p \log(1 + \frac{t_i}{\lambda_\perp})
\end{aligned}$$

Then,

$$\sum_{i=k+1}^p \log(1 + \frac{t_i}{\lambda_\perp}) \leq \sum_{i=k+1}^p \frac{t_i}{\lambda_\perp} \leq \frac{T}{\lambda_\perp} = k\log(1 + T/\lambda)\,.$$

□

By Lemma 3, the regret bound is, ignoring constants,

$$\begin{aligned}\left(\sigma k \log(1+T/\lambda) + \sqrt{k\log(1+T/\lambda)}\cdot(\sqrt{\lambda}S + \sqrt{\lambda_\perp}S_\perp)\right)\cdot\sqrt{T} &= \widetilde{\mathcal{O}}\left(\left(\sigma k + \sqrt{k}\cdot(\sqrt{\lambda}S + \sqrt{\lambda_\perp}S_\perp)\right)\sqrt{T}\right)\\ &= \widetilde{\mathcal{O}}((\sigma k + \sqrt{k\lambda}S + \sqrt{k}\cdot\sqrt{\frac{T}{k}}S_\perp)\sqrt{T})\\ &= \widetilde{\mathcal{O}}((\sigma k + \sqrt{k\lambda}S + \sqrt{T}S_\perp)\sqrt{T})\end{aligned}$$

F Proof of Theorem 5 and Corollary 2

Let us define $r_t = \max_{\mathbf{x}\in\mathcal{X},\mathbf{z}\in\mathcal{Z}} \mathbf{x}^\top\mathbf{\Theta}^*\mathbf{z} - \mathbf{x}_t^\top\mathbf{\Theta}^*\mathbf{z}_t$, the instantaneous regret at time t. Using $\max_{\|x\|_2,\|z\|_2\le 1}|\mathbf{x}^\top\mathbf{\Theta}^*\mathbf{z}| \le \|\mathbf{\Theta}^*\|_2$, we bound the cumulative regret incurred up to the stage 1 as $\sum_{t=1}^{T_1} r_t \le 2S_2T_1$. In the second stage, by the choices of S and S_F of (14),

$$\begin{aligned}\sum_{t=T_1+1}^{T_2} r_t &= \widetilde{\mathcal{O}}((\sigma k + \sqrt{k\lambda}S + \sqrt{T_2}S_\perp)\sqrt{T_2})\\ &= \widetilde{\mathcal{O}}\left(\left(\sigma k + \sqrt{k\lambda}S_F + \sqrt{T_2}\cdot\|\mathbf{X}^{-1}\|_2^2\|\mathbf{Z}^{-1}\|_2^2 C_1^2\frac{S_2^5}{S_r^6}\sigma^2 d^3 r\cdot\frac{1}{T_1}\right)\sqrt{T_2}\right)\end{aligned}$$

Then, the overall regret is, using $T_2 \le T$

$$\sum_{t=1}^{T} r_t = \widetilde{\mathcal{O}}\left(s_1^* T_1 + T\cdot\|\mathbf{X}^{-1}\|_2^2\|\mathbf{Z}^{-1}\|_2^2\frac{S_2^5}{S_r^6}\sigma^2 d^3 r\cdot\frac{1}{T_1}\right)$$

With the choice of $T_1 = \Theta\left(\sqrt{T\|\mathbf{X}^{-1}\|_2^2\|\mathbf{Z}^{-1}\|_2^2\frac{S_2^4}{S_r^6}\sigma^2 d^3 r}\right)$, the regret is

$$\widetilde{\mathcal{O}}\left(\frac{S_2^3}{S_r^3}\|\mathbf{X}^{-1}\|_2\|\mathbf{Z}^{-1}\|_2\sigma d^{3/2}\sqrt{rT}\right)$$

G Heuristics for selecting arms in stage 1

We describe heuristics for solving (4) when the arm set is finite.

Let $\mathbf{X}\in\mathbb{R}^{N_1\times d_1}$ be a matrix that takes each arm $\mathbf{x}\in\mathcal{X}$ as its row. One way to develop algorithms for solving (4) is to relax the cardinality constraint to a continuous one:

$$\begin{aligned}\min_{\boldsymbol{\lambda}}\quad & -t\\ \text{s.t.}\quad & \mathbf{X}^\top\mathrm{diag}(\boldsymbol{\lambda})\mathbf{X} \succeq t\mathbf{I}\\ & \lambda_i \ge 0 \quad \forall i\in[N_1]\\ & \textstyle\sum_{i=1}^{N_1}\lambda_i = 1\end{aligned}$$

We then choose the top d_1 arms with the largest λ_i.

We found that choosing the best among the solution above and 20 candidate subsets drawn uniformly at random (total 21 candidate subsets) returns reasonable solutions for our purpose.