Improved Regret Bounds of Bilinear Bandits using Action Space Analysis

Kyoungseok Jang [1] Kwang-Sung Jun [2] Se-Young Yun [3] Wanmo Kang [1]

Abstract

We consider the bilinear bandit problem where the learner chooses a pair of arms, each from two different action spaces of dimension d_1 and d_2, respectively. The learner then receives a reward whose expectation is a bilinear function of the two chosen arms with an unknown matrix parameter $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$ with rank r. Despite abundant applications such as drug discovery, the optimal regret rate is unknown for this problem, though it was conjectured to be $\tilde{O}(\sqrt{d_1 d_2 (d_1 + d_2) r T})$ by Jun et al. (2019) where $\tilde{O}$ ignores polylogarithmic factors in T. In this paper, we make progress towards closing the gap between the upper and lower bound on the optimal regret. First, we reject the conjecture above by proposing algorithms that achieve the regret $\tilde{O}(\sqrt{d_1 d_2 (d_1 + d_2) T})$ using the fact that the action space dimension $O(d_1+d_2)$ is significantly lower than the matrix parameter dimension $O(d_1 d_2)$. Second, we additionally devise an algorithm with better empirical performance than previous algorithms.

1. Introduction

Recently, researchers have shown much attention in the application of the bandit algorithms to the matching problem. Imagine a newly starting marriage agency company. Since they have less knowledge about how each factor of the customer (e.g., wealth, height, education) makes synergy with the opponent customer, they will want to try several matchings to learn the importance of each feature. However, they will also not want to lose their ratings by poor matchings caused by excessive exploration, so someday they should arrange couples based on their experiences to get better ratings and rewards. Balancing exploration and exploitation is the core framework of the bandit approach, and researchers start to involve in this approach to construct a better recommendation system for the matching problem. Few good examples are protein-drug pair approach (Luo et al., 2017), dating market (Das & Kamenica, 2005), duel matching system (Sui et al., 2018), and a cloth recommendation system.

However, research on this two-sided bandit problem has not been done well for even the simplest form, the bilinear model. While researchers have shown interest for a long time in pure exploration perspectives such as the matrix sensing and the matrix completion problem (Chi et al., 2019; Keshavan et al., 2009), there have been only few studies on the bilinear bandit problem.

We consider the stochastic bilinear bandit problem. Let $\mathcal{X} \subset \mathbb{R}^{d_1}$ and $\mathcal{Z} \subset \mathbb{R}^{d_2}$ be the left and right action space, respectively. For each round t, the agent chooses a pair of actions $x_t \in \mathcal{X}$ and $z_t \in \mathcal{Z}$ and then receives a reward r_t as a noisy bilinear function:

$$r_t = x_t^\top \Theta^* z_t + \eta_t$$

where $\eta_t \in \mathbf{R}$ is a σ sub-Gaussian noise. The objective is to maximize the cumulative rewards.

The lack of research on the bilinear bandit problem was partly due to the belief that the bilinear model can be sufficiently explained by the linear bandit model. The bilinear term $x_t^\top \Theta^* z_t$ in the reward with action spaces of dimension d_1 and d_2 can be re-written as $\langle vec(x_t z_t^\top), vec(\Theta^*) \rangle$ in the sense of $d_1 \times d_2$ dimensional linear bandit problem. Moreover, in the linear bandit field, several algorithms such as LinUCB (Abbasi-Yadkori et al., 2011) have proven their effectiveness. Naturally, specific studies aimed only for bilinear bandits are limited, and most of the existing studies have been mainly conducted only in the setting of broader structures (Johnson et al., 2016; Zimmert & Seldin, 2018), or with more powerful or peculiar structures (Katariya et al., 2017; Trinh et al., 2020; Kveton et al., 2017)

However, such naive linear bandit approaches for bilinear bandits cannot fully utilize the characteristics of the hidden parameter or action spaces, which leads to the limited regret analysis. Jun et al. (2019) proves that when the hidden

[1] Department of Mathematical Sciences, Korea Advanced Institute of Science and Technology, Daejeon, Korea [2] Department of Computation, University of Arizona, Arizona, USA [3] Graduate School of AI, Korea Advanced Institute of Science and Technology, Daejeon, Korea. Correspondence to: Wanmo Kang <wanmo.kang@kaist.ac.kr>.

Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

Table 1. Summary of bilinear bandit results where $d = \max(d_1, d_2)$ and $r = \mathsf{rank}(\Theta^*)$.

RESULTS	REGRET UPPER BOUND
LINUCB (2011)	$\tilde{O}(\sqrt{d_1^2 d_2^2 T})$
JUN (2019)	$\tilde{O}(\sqrt{d_1 d_2 d r T})$
LU(2021)	$\tilde{O}(\sqrt{d_1 d_2 d r T})$
ϵ-FALB (OURS)	$\tilde{O}(\sqrt{d_1 d_2 d T})$

parameter space has a low-rank structure, there exists an algorithm with a better regret than the naive linear bandit algorithm applications. After this, researchers have studied the structure of hidden parameters, cf., Lu et al. (2021); Hao et al. (2020); Kotlowski & Neu (2019). In contrast, existing researches have not shown much interest in the geometry of the action space. Most of the papers have only summarized how to apply the hidden parameter structure and ignored the fact that the action space has a much lower dimension than the hidden parameter space. This paper achieves a better regret result by focusing on the action space.

Our contributions can be summarized as follows.

- We construct a new algorithm ϵ-FALB (Finite Armed Linear Bandit) with an improved regret upper bound of $\tilde{O}(\sqrt{d^3 T})$ for the bilinear bandit problem, where $d = \max(d_1, d_2)$. The key idea is to leverage the low-rank nature of the action space rather than the hidden parameter space. This rejects the conjectured lower bound of $\Omega(\sqrt{d^3 r T})$ by Jun et al. (2019) where $r = \mathsf{rank}(\Theta^*)$. However, this algorithm requires discretization of the arm sets, which leads to impractical time and space complexity of $O(T^{d/2})$.
- Towards practical solutions, we construct a novel bilinear bandit algorithm called rO-UCB (rank-r Oracle UCB) that enjoys a tractable time complexity. We show that rO-UCB exhibit an excellent numerical performance and significantly outperforms baseline methods including ESTR (Jun et al., 2019), thanks to the lack of forced exploration that ESTR must perform. The design of rO-UCB is based on our novel adaptive design of confidence bound for low-rank matrices that can be used beyond rank-one measurements, which can be of independent interest.

We remark that both algorithms can be applied to the changing arm set environment whereas ESTR works only for the fixed arm set due to its forced exploration phase, which widens the applicability of bilinear bandits such as personalized recommendations based on contextual information.

The paper is structured as follows. Section 2 introduces related works. In Section 3, we define the problem settings and notations. Section 4 provides the main contribution of our paper. Section 5 describes the practical algorithms that overcomes the intractability of our main algorithm. We state new conjecture on the regret lower bound in Section 6, and discuss the future research directions in Section 7.

2. Related works

Bilinear bandit is a field that has received much attention recently. Mainly, the rank-1 bilinear bandit problem is relatively easy to analyze and has useful applications, so there are several instance-dependent regret analyses for the rank-1 bilinear bandit problem. However, it is not easy to generalize those studies to rank-r bilinear bandit since they depend profoundly on the properties of the rank-1 matrix. For example, Katariya et al. (2017) and Trinh et al. (2020) have dealt with Bernoulli rank-1 bandit, all entries are positive, and only canonical vectors are allowed for each side of actions. In these cases, they exploited the property that the maximum reward comes from multiplicating the maximum entry of vector u and v. This tendency is difficult to transfer to the rank-r case. Similarly, there is also a paper that analyzes the rank-r case (Kveton et al., 2017). However, the objective of the paper is finding the maximum entry of the hidden matrix which is again only about the action set with canonical vectors on both sides. Plus, they assumed strong hott topic matrix assumption on the hidden matrix.

Jun et al. (2019) have introduced the bilinear low rank bandit problem. They propose an algorithm ESTR (Explore Subspace Then Refine) that performs subspace exploration first to make a low-rank approximation of the hidden parameter, then performs the algorithm called LowOFUL, which is a subspace-regularized version of the algorithm OFUL (Abbasi-Yadkori et al., 2011) that exploits the learned information about the low-rank subspaces. ESTR shows $\tilde{O}(\sqrt{d^3 r T})$ regret upper bound, which is meaningful since it is the first algorithm better than the naive OFUL algorithm regret $O(d_1 d_2 \sqrt{T})$. As a follow-up study on this, Lu et al. (2021) studied the extension of the bilinear bandit problem. This paper uses the fact that one can also interpret bilinear term $x^\top \Theta z$ as $\langle vec(xz^\top), vec(\Theta)\rangle$, and proves that the ESTR could achieve almost the same regret bound for generalized action set. They also suggested a lower bound $O(rd\sqrt{T})$ for the extended model, but as will be described later, our paper shows a regret upper bound algorithm that is lower than the lower bound presented here, indicating that the setting here is too broad that this lower bound cannot wholly explain the properties of the bilinear bandit. Both Jun et al. (2019) and Lu et al. (2021) presented the conjecture that the upper bound suggested in the Jun et al. (2019) paper will be tight; however, we refute this argument in Section 4 by designing an algorithm with a lower regret bound.

Kotlowski & Neu (2019) has devised an algorithm that performs $O(\sqrt{rd^2 T})$ regret upper bound for a specific adversarial symmetric bilinear bandit called bandit PCA. However, this study differs from the general bilinear bandit study since their action set is smaller and specific. We will discuss in Section 5 and Section 6 about this algorithm and its extension in details.

There are numerous bandit papers that consider structural assumptions that bilinear bandits are subproblems. Low-rank tensor bandit (Hao et al., 2020) extends the hidden parameter from a matrix to a tensor. Structured bandits (Johnson et al., 2016; Yu et al., 2020) propose unified frameworks for bandits with structure including bilinear bandits. Lastly, factored bandit paper (Zimmert & Seldin, 2018) deals with the bandit problem, whose action set is a Cartesian product of atomic actions. While these studies allow more general structures, they do not exploit the rank-1 structure of the action space for the bilinear bandit case.

Finally, the linear bandit is indispensable to the bilinear bandit discussion (Abbasi-Yadkori et al., 2011; Dani et al., 2008; Lattimore & Szepesvári, 2020). As we mentioned in the introduction, the bilinear bandit can be reinterpreted in the form of the linear bandit as follows:

$$r_t = x_t^\top \Theta^* z_t + \eta_t = \langle vec(x_t z_t^\top), vec(\Theta^*)\rangle + \eta_t \qquad (1)$$

where η_t is a sub-Gaussian noise. Consequently, any linear bandit algorithms can be applied to bilinear bandit problems. However, these algorithms do not exploit the rank structure of the action nor the unknown parameter, leading to loose regret bounds. For example, applying OFUL (Abbasi-Yadkori et al., 2011) gives $O(d_1 d_2 \sqrt{T})$. To exploit the geometry of the action set of our problem, we get inspiration from finite armed linear bandits (Auer, 2002; Chu et al., 2011). There were a few linear bandit studies when the action set is a subspace or its perturbation (Lale et al., 2019; Hamidi et al., 2019), but the action set of the bilinear bandit interpreted as (1) are generally not the subspace of $\mathbb{R}^{d_1 d_2}$.

3. Problem definition

In this section we formally define the problem and notations. Let $\mathcal{X} \subset \mathbb{R}^{d_1}$ and $\mathcal{Z} \subset \mathbb{R}^{d_2}$ be the left and right action space, respectively. Without loss of generality, we assume that all these actions have l_2 norm bounded by 1.

Let $d = \max(d_1, d_2)$ for convenience, and $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$ be the hidden parameter matrix. Let $\lambda_i(\Theta)$ be the i-th largest singular value of Θ. Without loss of generality we assume that $\lambda_1(\Theta^*) \le 1$ to bound the expected reward. We define $r = \mathsf{rank}(\Theta^*)$, which is not necessarily known to the agent.

For each round t, the agent chooses a pair of actions $x_t \in \mathcal{X}$ and $z_t \in \mathcal{Z}$ then receives a reward r_t as a noisy bilinear function:

$$r_t = x_t^\top \Theta^* z_t + \eta_t$$

where $\eta_t \in \mathbf{R}$ is a σ sub-Gaussian noise conditioning on $(x_s, z_s)_{s\le t}$ and $(r_s)_{s<t}$. The goal of this bandit problem is to maximize the cumulative rewards, or equivalently, minimize the following pseudo-regret:

$$R_T = \sum_{t=1}^{T} (x_*^\top \Theta^* z_* - x_t^\top \Theta^* z_t)$$

where (x_*, z_*) is defined as $\arg\max_{x\in\mathcal{X}, z\in\mathcal{Z}} x^\top \Theta^* z$, optimal action pair in hindsight.

Notations Let $\mathcal{B}_d$ be the unit ball centered at the origin in $\mathbf{R}^d$. For a positive definite matrix $P \in \mathbf{R}^{d\times d}$, the weighted 2-norm of vector $x \in \mathbf{R}^d$ is $\|x\|_P = \sqrt{x^\top P x}$. For any sequence of d-dimensional vector $\{a_t\}$, we denote $a_{s:t} = [a_s | a_{s+1} | \cdots | a_t] \in \mathbf{R}^{d\times(t-s+1)}$ as the horizontally concatenated matrix of this subsequence of vectors. I_d represents the $d \times d$ identity matrix.

4. Main algorithm

Algorithm 1 ϵ-FALB

```
Input: $\beta$, Alg : finite armed linear bandit algorithm, $\epsilon$ : distance for the covering sets, $T$ : number of pulls
Construct $\epsilon$-covering set $\mathcal{X}_\epsilon$ and $\mathcal{Z}_\epsilon$
Initialize $\mathcal{A} = \{vec(xz^\top) : x \in \mathcal{X}_\epsilon, z \in \mathcal{Z}_\epsilon\}$.
Perform Alg with action set $\mathcal{A}$, time horizon $T$, and confidence bound constant $\beta$
```

In this section, we describe a new approach ϵ-FALB(finite armed linear bandit) that guarantees $\tilde{O}(\sqrt{d^3 T})$ regret for general action spaces, even applicable to the changing action spaces. Here, we focus on using the geometry of the action space without any knowledge of the rank r. Our framework first constructs ϵ-covering sets $\mathcal{X}_\epsilon$ and $\mathcal{Z}_\epsilon$. Then we run a finite armed linear bandit algorithm as described in Algorithm 1. Such a discretization of action spaces is folklore in the community; e.g., Beygelzimer et al. (2011, Theorem 5).

To our best knowledge, the best regret for the bilinear bandit setting was $\tilde{O}(\sqrt{d^3 r T})$ by Jun et al. (2019). The corresponding lower bound is not tight yet, but the authors claimed that the regret lower bound might be $\tilde{O}(\sqrt{d^3 r T})$ as well from a signal to noise ratio analysis. However, the reason why rank r should be in the regret term was not entirely clear.

To achieve the improved regret bound by applying the finite armed linear bandit algorithm in bilinear setting, it is key to control the number of discretized points of action spaces. The ε-covering produces a discretization of the cardinality of $\exp(\tilde{O}(d))$ and it is enough to obtain the desired regret upper bound.

In Section 4.1, we review the linear bandit algorithms to convey why we choose finite armed linear bandit algorithms for our algorithm. Section 4.2 describes how we exploited the action space geometry through ϵ-covering set construction.

Section 4.3 tells the necessary modification for the finite armed linear bandit, and Section 4.4 is for the main regret analysis. Section 4.5 is about the extension of our algorithm to the matrix action space case.

4.1. Reason for choosing finite armed linear bandit algorithms

Only for this subsection, let us assume a simple linear bandit model defined as follows. At every round, the agent selects action x_t from action set $\mathcal{A} \subset \mathbb{R}^d$, and receives a noisy reward $r_t = x_t^\top \theta_* + \eta_t$. Here, $\theta_* \in \mathbb{R}^d$ is a hidden parameter that the agent does not know, and η_t is σ sub-Gaussian noise. In this problem, $V_t = \sum_{s=1}^t x_t x_t^\top$ and $\bar{V}_t = V_t + \lambda I_d$ for some positive regularizing constant $\lambda > 0$, and $\hat{\theta}_t = \bar{V}_t^{-1} x_{1:t} r_{1:t}^\top$ is the Regularized Least Square estimator.

For each fixed action, the upper confidence bound of the expected reward is well known.

Theorem 4.1. *(Valko et al. (2014, Lemma 7), Chu et al. (2011, Lemma 1)) For each fixed $x \in \mathbb{R}^d$, the following inequality holds with probability $1 - \delta$:*

$$\langle x, \hat{\theta}_t - \theta_* \rangle \leq \|x\|_{\bar{V}_t^{-1}} O\left(\sqrt{\log \frac{1}{\delta}}\right) + \sqrt{\lambda}\|\theta_*\| \qquad (2)$$

The inequality above is one of the most trusted inequalities that can give a confidence bound for each point x, which is derived using Chernoff bound. The main difference between finite armed bandits and linear bandit with a broad action set depends on whether or not Eq. (2) can be applied directly to each action.

In the linear bandit with a broad action set, it is hard to expect all actions to satisfy Eq. (2) simultaneously by the union bound argument because there are too many actions. Instead, most of existing approaches utilize the fact that θ_* and $\hat{\theta}$ are close in terms of l_2 distance, and Cauchy's inequality: $\langle x, \hat{\theta} - \theta_* \rangle \leq \|x\|_{\bar{V}_t^{-1}} \|\hat{\theta} - \theta_*\|_{\bar{V}_t}$. However, Cauchy's inequality is generally not tight, which leads to the additional dimension dependency of the regret bound. We deferred the detailed discussion in the Appendix A.

On the other hand, in the finite armed linear bandit case, Eq. (2) is used to construct a high probability confidence bound. Since the number of action is finite, a simple union bound argument can decide the appropriate failure rate δ to satisfy the equation Eq. (2) for all actions as follows:

Theorem 4.2. *(Auer, 2002; Valko et al., 2014) For a fixed set A with $|A| = K$, The following inequality holds with probability $1 - \delta$: For all $x \in A$*

$$\langle x, \hat{\theta}_t - \theta_* \rangle \leq \|x\|_{\bar{V}_t^{-1}} O\left(\sqrt{\log \frac{K}{\delta}}\right) + \sqrt{\lambda}\|\theta_*\| \qquad (3)$$

Finite armed linear bandit algorithms do not suffer the additional dimension dependency that the general action set case has to take. Instead, the finite armed case regrets have additional $\sqrt{\log K}$ terms because of the union bound argument. In the next section, $\sqrt{\log K}$ will reflect the dimension of the action set.

4.2. Extension to the general action set case

For any given set S, the growth rate of ϵ-covering number, $N(S, \epsilon)$ is hinged on the dimension of S (see Hausdorff dimension). Since $\mathcal{X} \subset \mathcal{B}_{d_1}$ and $\mathcal{Z} \subset \mathcal{B}_{d_2}$ one can easily expect $K \approx O(d \log \frac{1}{\epsilon})$, and this is what we want to talk in this subsection. Formal proof of the bound for $N(\mathcal{X}, \epsilon)$ and $N(\mathcal{Z}, \epsilon)$ comes from the following lemma (adapted from Lattimore & Szepesvári (2020, Problem 20.3)).

Lemma 4.3. *For a bounded set $\mathcal{S} \subset \mathbb{R}^d$, its covering number $N(\mathcal{S}, \epsilon)$ satisfies the following inequality:*

$$N(\mathcal{S}, \epsilon) \leq \frac{vol(\mathcal{S}' + \frac{\epsilon}{2}\mathcal{B}_d)}{vol(\frac{\epsilon}{2}\mathcal{B}_d)} \qquad (4)$$

Here, $\mathcal{S}'$ is an arbitrary measurable set that contains S, and $\mathcal{S}' + \frac{\epsilon}{2}\mathcal{B}_d$ is a sumset between $\mathcal{S}'$ and $\frac{\epsilon}{2}\mathcal{B}_d$.

We deferred the detailed proof in Appendix D. Now since $\mathcal{X} \subset \mathcal{B}_{d_1}$ and $\mathcal{Z} \subset \mathcal{B}_{d_2}$, we can conclude that $N(\mathcal{X}, \epsilon) \leq (\frac{3}{\epsilon})^{d_1}$ and $N(\mathcal{Z}, \epsilon) \leq (\frac{3}{\epsilon})^{d_2}$ (see Lattimore & Szepesvári (2020, Lemma 20.1) for the covering number of the $\mathbb{S}^{d-1}$).

When we apply this lemma to the linearized action spaces (set of $xz^\top$) of the bilinear bandit problem, the cardinality of the discretized space can not be sharpened to a lower value than $O(\epsilon^{-d_1 d_2})$ whereas it is possible to get a cardinality of order $O(\epsilon^{-d_1 - d_2})$ if we apply the covering to the left and right action spaces separately.

4.3. Modification of the finite arm algorithm

The only remaining part is which algorithm we will use for the input of Algorithm 1. When it comes to the finite armed linear bandit algorithm, the SupLinRel based algorithms are usually the best known (Auer, 2002; Chu et al., 2011; Valko et al., 2013). However, these algorithms require some modifications for the bilinear bandit setting to optimize the regret. In particular, they use an assumption about the l_2-norm boundedness of the hidden parameter to compute the regret. We need some modifications to apply them to our current bilinear bandit problem. In the bilinear bandit setting, only the singular value limits the maximum reward, and rank of Θ^* is a factor that increases Frobenius norm from $\|\Theta^*\|_F^2 = \sum_{i=1}^r \lambda_i(\Theta^*)^2 \leq r\lambda_1(\Theta^*)^2$. Thus from Eq. 3 with replacing $\|\theta_*\|$ to $\|\Theta^*\|_F$, without proper regularization on λ the confidence bound width has an order of $\sqrt{r}$ no matter what $\log K$ is. When $\log K \ll r$ it is a severe loss of the regret upper bound since the regret upper bound of the UCB-type linear bandit algorithm is usually proportional to the confidence bound.

Algorithm 2 SupLinUCB(adapted from Chu et al. (2011))

```
Input: $\beta$, $S = \lceil \ln T \rceil$, $\Phi_t^s \leftarrow \emptyset$ for all $s \in [S]$
Initialize $\mathcal{A}_1 = \mathcal{A}, s = 1$.
for $t = 1$ to $T$ do
  repeat
    Calculate $\hat{r}_{t,a}^s$ and $w_{t,a}^s$ using BaseLinUCB with $\Phi_t^s$ for all $a \in \mathcal{A}_s$
    if $w_{t,a}^s \leq \frac{1}{\sqrt{T}}$ for all $a \in \mathcal{A}_s$ then
      Choose $a_t = \arg\max_{a \in A_t}(\hat{r}_{t,a}^s + w_{t,a}^s)$
      $\Phi_{t+1}^{s'} \leftarrow \Phi_t^{s'}$ for all $s' \in [S]$
    else if $w_{t,a}^s \leq 2^{-s}$ for all $a \in \mathcal{A}_s$ then
      $\mathcal{A}_{s+1} = \{a \in \mathcal{A}_s : \hat{r}_{t,a}^s + w_{t,a}^s \geq \max_{a' \in \mathcal{A}_s}(\hat{r}_{t,a'}^s + w_{t,a'}^s) - 2 \cdot 2^{-s}\}$
      $s \leftarrow s + 1$
    else
      Choose $a_t \in \mathcal{A}_s$ such that $w_{t,a_t}^s > 2^{-s}$
      $\Phi_{t+1}^s \leftarrow \Phi_t^s \cup \{t\}$
      $\Phi_{t+1}^{s'} \leftarrow \Phi_t^{s'}$ for all $s' \in [S] \backslash \{s\}$
    end if
  until $a_t$ is found
end for
```

Algorithm 3 BaseLinUCB (Chu et al., 2011)

```
Input: $\beta$, $\Phi_t^s = \{t_1, t_2, \cdots t_l\}$, $V_0 = \frac{1}{d} I_{d_1 d_2}$
$X_{t,s} = [a_{t_1}; a_{t_2}; \cdots a_{t_l}]$
$R_{t,s} = [r_{t_1}, r_{t_2}, \cdots, r_{t_l}]^\top$
$V_{t,s} = V_0 + \sum_{\tau \in \Phi_t^s} a_\tau a_\tau^\top$
$w_{t,a}^s = \beta \|a\|_{V_{t,s}^{-1}}$
$\hat{r}_{t,a}^s = V_{t,s}^{-1} X_{t,s} R_{t,s}$
Return $\hat{r}_{t,a}^s$ and $w_{t,a}^s$
```

Algorithm 2 is the modified SupLinUCB for the bilinear setting. Note that unlike Chu et al. (2011), we add $\frac{1}{d} I_d$ instead of I_d for the regularized gram matrix V_t, since we have to control the scale of $\sqrt{\lambda}\|\theta_*\|$ term in Eq. (2) by setting $\lambda = \frac{1}{d}$.

Considering that the proof in Chu et al. (2011) strongly depends on the fact that $\lambda_{min}(V_t) \geq 1$ and the boundedness of the reward, we need several modifications for the regret upper bound proof. The detailed proof is in the Appendix B. After that, the following regret upper bound holds:

Theorem 4.4. *If we run Algorithm 2 with* $\beta_t = 2\sigma\sqrt{14 \log \frac{2KT \log T}{\delta}} + 1$ *the regret is bounded by*

$$R_T \leq \tilde{O}\left(\sqrt{d_1 d_2 T \log \frac{K}{\delta}}\right)$$

with probability $1 - \delta$.

The main advantage of SupLinUCB is that the algorithm can be applied to the changing arm sets since it is basically for the contextual linear bandit problem.

Algorithm 4 Phase Elimination (Valko et al., 2014)

```
Input: $T$ : the number of pulls, $\mathcal{A}$ : finite action set, $\beta$, $\{t_j = 2^{j-1}\}$ : parameters of elimination and phase
Initialize $\mathcal{A}_1 = \mathcal{A}$.
for $j = 1$ to $J$ do
  $V_{t_j} \leftarrow \frac{1}{d} I_{d_1 d_2}$
  for $t = t_j$ to $t_{j+1} - 1$ do
    $a_t \leftarrow \arg\max_{a \in \mathcal{A}_j} \|a\|_{V_t^{-1}}$
    $V_{t+1} \leftarrow V_t + a_t a_t^\top$
  end for
  $\hat{\Theta}_j = V_t^{-1} a_{t_j:t} r_{t_j:t}^\top$
  $p \leftarrow \max_{a \in \mathcal{A}_j} a^\top \hat{\Theta} - \|a\|_{V_t^{-1}} \beta$
  $\mathcal{A}_{j+1} \leftarrow \{a \in \mathcal{A}_j : a^\top \hat{\Theta} + \|a\|_{V_t^{-1}} \beta \geq p\}$
end for
```

On the other hand, if we want to consider about Spectral Eliminator (Valko et al., 2014) and Phased elimination with G-optimal exploration (Lattimore & Szepesvári, 2020; Soare et al., 2014), they are directly applicable with some tuning on the initial matrix V_0. Instead, we cannot apply these algorithms for the changing arm sets. Algorithm 4 is a Spectral Eliminator with initial matrix $V_0 = \frac{1}{d} I_{d_1 d_2}$. Again, the regularizing constant is $\frac{1}{d}$ to control the scale of the last $\|\theta_*\|$ term in Eq. (2). Without any modification of the proof, the following regret bound holds:

Theorem 4.5. *(Valko et al., 2014) If we run Algorithm 4 with failure probability* δ*, bounding constant* $\beta = 2\sigma\sqrt{14 \log \frac{2K \log_2 T}{\delta}} + 1$*, then with probability at least* $1 - \delta$ *the following regret bound holds.*

$$R_T \leq \frac{4}{\log 2}\left(2\sigma\sqrt{14 \log \frac{2K \log_2 T}{\delta}} + 1\right) \times \sqrt{d_1 d_2 T \log(1 + (d_1 + d_2)T)}$$

In short, both algorithm shows the regret upper bound of $\tilde{O}(\sqrt{d_1 d_2 T \log K})$ with probability at least $1 - \delta$.

4.4. Regret analysis

Theorem 4.6. *Algorithm 1 with input* $\epsilon = \frac{1}{\sqrt{T}}$, Alg *as Algorithm 2 or Algorithm 4, and* β *for suitable constant for* Alg *in Theorem 4.4 and Theorem 4.5 satisfies the following regret upper bound with probability* $1 - \delta$*:*

$$R_T \leq \tilde{O}(\sqrt{d_1 d_2 (d_1 + d_2) T \log \frac{1}{\delta}}) \tag{5}$$

Proof. Let $K = |\mathcal{A}|$, $x_\epsilon = \arg\min_{x \in \mathcal{X}_\epsilon} \|x_* - x\|$, and $z_\epsilon = \arg\min_{z \in \mathcal{Z}_\epsilon} \|z_* - z\|$. We can separate the regret of the Algorithm 1 to the following three terms:

$$
\begin{aligned}
R_T &= \sum_{t=1}^{T} x_*^\top \Theta^* z_* - \sum_{t=1}^{T} x_t^\top \Theta^* z_t \\
&= \sum_{t=1}^{T} x_*^\top \Theta^* z_* - \sum_{t=1}^{T} x_\epsilon^\top \Theta^* z_\epsilon \\
&+ \sum_{t=1}^{T} x_\epsilon^\top \Theta^* z_\epsilon - \sum_{t=1}^{T} \max_{x,z\in\mathcal{E}} x^\top \Theta^* z \\
&+ \sum_{t=1}^{T} \max_{x,z\in\mathcal{E}} x^\top \Theta^* z - \sum_{t=1}^{T} x_t^\top \Theta^* z_t \\
&= R_1 + R_2 + R_3
\end{aligned}
$$

Here, $R_1 = \sum_{t=1}^T x_*^\top \Theta^* z_* - \sum_{t=1}^T x_\epsilon^\top \Theta^* z_\epsilon$ represents the reward difference between the optimal action and its closest ϵ-covering set element x_ϵ, z_ϵ. $R_2 = \sum_{t=1}^T x_\epsilon^\top \Theta^* z_\epsilon - \sum_{t=1}^T \max_{x\in\mathcal{X}_\epsilon, z\in\mathcal{Z}_\epsilon} x^\top \Theta^* z$ is the difference between the action closest to the optimal action and the optimal action among ϵ-covering set elements. $R_3 = \sum_{t=1}^T \max_{x,z\in\mathcal{E}} x^\top \Theta^* z - \sum_{t=1}^T x_t^\top \Theta^* z_t$ is the regret of the finite armed linear bandit algorithm. Now those three regret terms are calculated as follows:

- By definition, $R_2 \leq 0$
- R_3 can be bounded by $O(\sqrt{d_1 d_2 T \log \frac{K}{\delta}})$ by Theorem 4.4 or Theorem 4.5.
- Lastly, since $\|x_* z_*^\top - x_\epsilon z_\epsilon^\top\|_F \leq \|(x_* - x_\epsilon) z_*^\top\|_F + \|x_\epsilon (z_*^\top - z_\epsilon^\top)\|_F \leq 2\epsilon$ by the ϵ-cover construction, R_1 is bounded as follows:

$$
\begin{aligned}
&\sum_{t=1}^{T} x_*^\top \Theta^* z_* - \sum_{t=1}^{T} x_\epsilon^\top \Theta^* z_\epsilon \\
&\quad = \sum_{t=1}^{T} \langle vec(\Theta^*), vec(x_* z_*^\top - x_\epsilon z_\epsilon^\top) \rangle \\
&\quad \leq \sum_{t=1}^{T} \|\Theta^*\|_F \cdot \|x_* z_*^\top - x_\epsilon z_\epsilon^\top\|_F \leq 2\epsilon T \|\Theta^*\|_F
\end{aligned}
$$

Overall, the regret bound is

$$
\begin{aligned}
R_T &\leq R_1 + R_2 + R_3 \\
&\leq 2\epsilon T \sqrt{r} \lambda_1(\Theta^*) + 0 + \tilde{O}(\sqrt{d_1 d_2 T \ln(\frac{K}{\delta})})
\end{aligned}
$$

Substituting $\epsilon = \frac{1}{\sqrt{T}}$ and using the fact $K = N(\mathcal{A}, \epsilon) = O((\frac{1}{\epsilon})^{d_1+d_2})$ from Section 4.2 concludes the theorem. □

Remark 1 Note that from the proof the final regret bound is $\tilde{O}(\sqrt{d_1 d_2 T \ln \frac{K}{\delta}})$, and the regret of Eq. 5 is from $\log K = \tilde{O}(d)$. The bound can be even lower when the scale of $N(\mathcal{X}, \epsilon)$ (or $N(\mathcal{Z}, \epsilon)$) is much smaller than d_1 (or d_2, respectively), thanks to the modifications and initialization of V_0 discussed in 4.3. One of the cases is when $\mathcal{X}$ and $\mathcal{Z}$ are finite action spaces.

Remark 2 One might wonder which ϵ shows the best empirical performance of ϵ-FALB in practice. We can get the same order of regret upper bound when $\epsilon \in [\frac{1}{\sqrt{T}}, \frac{d}{\sqrt{T}}]$, and this range is also the best choice for empirical perspectives. Appendix E.1 includes the experiment about the ϵ-value selection.

4.5. Extension to the action set of matrices

In the previous section, we used the fact that the action space of the bilinear bandit has much smaller dimensions than $d_1 \times d_2$ – from the perspective of (1), the action space is a set of some rank-1 matrices. Then, one natural question is whether we can extend the previous result to the action space consists of matrices with rank $\leq \rho$ for some constant ρ. Specifically, for the linear bandit problem

$$y_t = \langle vec(A_t), vec(\Theta^*) \rangle + \eta_t$$

with the action space $\mathcal{A} \subset \{\Theta \in \mathbb{R}^{d_1 \times d_2} : \lambda_{max}(\Theta) \leq 1, \mathsf{rank}(\Theta) \leq \rho\}$, we can expect to achieve a better regret-bound compared to the naive $d_1 d_2$ dimensional linear bandit, and we show that it holds partially as we will see in the following corollary. We can prove the corollary in a similar way to the proof of Theorem 4.6.

Corollary 4.7. *Let $O_{d',\rho} = \{M \in \mathbb{R}^{d' \times \rho} : MM^\top = I_{d'}\}$, $D_\rho = \{Diag(\theta) : \theta \in [-1,1]^\rho\}$. Suppose there are three sets $\mathcal{X} \subset O_{d_1,\rho}, \mathcal{Z} \subset O_{d_2,\rho}, \mathcal{D} \subset D_\rho$ such that the action set can be represented as the product of three sets, namely $\mathcal{A} = \{vec(U\Sigma V^\top) : U \in \mathcal{X}, V \in \mathcal{Z}, \Sigma \in D_\rho\}$. If we run Algorithm 1 with action set $\mathcal{A}_\epsilon$(ϵ-covering set of $\mathcal{A}$) and other hyperparameters described as Theorem 4.6, then the regret is bounded as below with probability $1 - \delta$*

$$R_T \leq \tilde{O}(\sqrt{d_1 d_2 \rho (d_1 + d_2) T \log(\frac{T}{\delta}})$$

We left the details in the Appendix D. Note that all rank ρ matrix can be decomposed as $\Theta = U\Sigma V^\top$ by singular value decomposition, the Corollary 4.7 covers wide range of rank-ρ action sets.

5. Practical algorithms

Although the Algorithm 1 shows better regret bound than the previous studies, it is not tractable to apply Algorithm 1 in practice since the cardinality of $\mathcal{X}_\epsilon$ and $\mathcal{Z}_\epsilon$ grows in the order of $O((\frac{1}{\epsilon})^d) = O(T^{d/2})$ in general, which is spatially intractable. This spatial drawback leads a serious computational time disadvantage - see Appendix E.2 for details.

In addition, finite armed linear bandit algorithms are well known to be inefficient in practice compare to the linear bandit algorithms with general action space (Valko et al., 2014; Chu et al., 2011).

Instead, we devise two practical algorithms that one shows superior empirical performance, and the other shows provable computational complexity.

Table 2. Summary of our additional algorithms. Here Forced exp. is about whether the algorithm requires first forced exploration phase.

RESULTS	REGRET BOUND	FORCED EXP.	ACTION SPACE
ϵ-FALB	$\tilde{O}(\sqrt{d^3T})$	NO	CHANGABLE
RO-UCB	$\tilde{O}(\sqrt{d^3rT})$	NO	CHANGABLE
B-PCA (2019)[1]	$\tilde{O}(\sqrt{d^3T})$[2]	NO	$\mathbb{S}^{d-1}$
ESTR (2019)	$\tilde{O}(\sqrt{d^3rT})$	YES	FIXED

5.1. Considering hidden parameter structure

We have verified that Algorithm 1 can guarantee regret bound $\tilde{O}(\sqrt{d^3T})$ even for the worst-case by considering the geometry of the action set, although we do not know whether it is optimal or not. From the result, the rank of the hidden parameter might not affect much on the worst-case regret of the bilinear bandit.

However, it is undeniable that knowing the rank of the problem might help better approximation, evidenced by historical low-rank studies (Chi et al., 2019).

Suppose that there exists an oracle that solves the following optimization problem, and the answer is $\hat{\Theta}_t$

$$\text{(Opt)} \begin{cases} \min_\Theta & \sum_{s=1}^{t}(x_s^\top \Theta z_s - r_s)^2 \\ \text{subject to} & \mathrm{rank}(\Theta) \le r, \\ & \|\Theta\|_F \le C \end{cases}$$

In practice, the existing low-rank estimation algorithms usually depend on the gradient descent-based methods. They need several conditions about action x_s and z_s to guarantee to find the solution of (Opt), such as the restricted isometry condition (Chi et al., 2019; Bhojanapalli et al., 2016) as gradient descent methods usually require convexity conditions on the landscape. Those conditions are usually hard to achieve in the action history of the bandits. However, assuming that the oracle for (Opt) exists, we can create a concentration inequality like follows:

Theorem 5.1. *For all $t \in \{1, \cdots, T\}$, $\hat{\Theta}_t$ defined as above satisfies the following inequality with probability at least $1-\delta$:*

$$\|vec(\hat{\Theta} - \Theta^*)\|_{W_t} \le O\left(\sqrt{rd \log \frac{CT}{\delta}}\right)$$

where $W_t = I_{d_1 d_2} + \sum_{s=1}^{t-1} vec(x_s z_s^\top) vec(x_s z_s^\top)^\top$.

With this oracle, we can construct an algorithm, adapted from linUCB, that has a regret of order $\tilde{O}(\sqrt{rd^3T})$. See Appendix C for its proof.

Algorithm 5 rO-UCB (rank r Oracle UCB)

```
Input: $\beta$, $W_0 = I_{d_1 d_2}$, $C = \sqrt{r}$
for $t = 1$ to $T$ do
    $W_t = W_0 + \sum_{s=1}^{t-1} vec(x_s z_s^\top) vec(x_s z_s^\top)^\top$
    $\hat{\Theta}_t = \mathsf{Oracle}(x_{1:t-1}, z_{1:t-1}, r_{1:t-1}, r, C)$
    $\mathsf{UCB}_t(x, z) = x^\top \hat{\Theta}_t z + \beta \|vec(xz^\top)\|_{W_t^{-1}}$
    Choose $(x_t, z_t) = \arg\max_{(x,z)\in\mathcal{X}\times\mathcal{Z}} \mathsf{UCB}_t(x, z)$ and receive reward $r_t$
end for
```

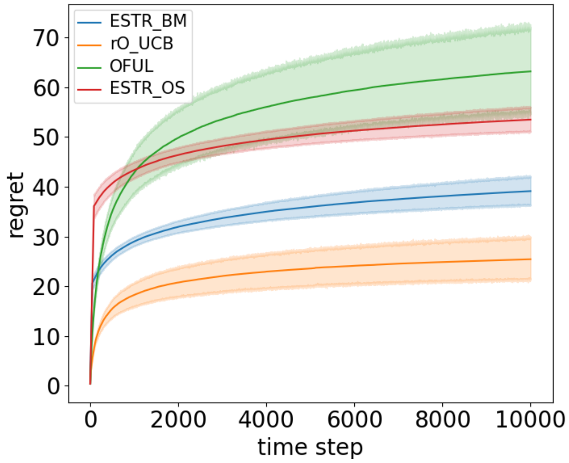

Figure 1. Simulation result for $d = 8$ and $r = 1$, and $\sigma = 0.01$. We plot the average regret of the methods and the .95 confidence intervals. Our method outperforms all the known general bilinear bandit algorithms.

We present an experiment to compare the performance of the existing bilinear algorithms and our rO-UCB algorithm.

In the experiment, we consider the four methods: ESTR-OS, which is the proposed method of Jun et al. (2019); ESTR-BM, the best heuristic method in Jun et al. (2019); OFUL, naive OFUL extension as discussed in (1); and rO-UCB, our proposed algorithm. We use the grid search method to adjust the forced exploration time of ESTR and confidence bound width of all algorithms. Instead of the true oracle, we used one of the low rank approximation of Burer & Monteiro

[1] Though the algorithm was designed by Kotlowski & Neu (2019), we adapted this algorithm to the stochastic environment and calculated the regret upper bound result.

[2] This bound is about the expected regret upper bound. It is another challenging problem to calculate the high probability regret bound for the bandit PCA algorithm.

(2003) instead. The graph is about the best regret result for each algorithm. Our rO-UCB outperforms every other algorithms, and one can see several additional experiments in the Appendix E.3 that in another environment with larger σ, our rO-UCB outperforms other algorithms with stability while ESTR algorithms fail because of the unsuccessful forced exploration. We leave the hyper-parameters and additional experiments in the Appendix E.3.

5.2. Stochastic bandit PCA analysis

Kotlowski & Neu (2019) is one of main inspirations of our research. The bandit PCA of Kotlowski & Neu (2019) is a specialized version of the adversarial bilinear problem, which repeats the following steps for each round:

- Agent selects action $x_t \in \mathbb{S}^{d-1}$ through history.
- The environment choose a d by d symmetric matrix L_t with a spectral norm of less than 1.
- Agent receives loss(or reward) $l_t = x_t^\top L_t x_t$.

Indeed, this is a partial problem of the bilinear problem, where the right and left actions are the same. Thus, we analyze the regret of the stochastic bandit PCA problem to check the regret lower bound of the bilinear bandit problem. The problem changes as follows:

- The environment decides the d by d symmetric matrix L with a spectral norm less than 1 at the start of the game. That is, $L_t = L$ for all t.
- Agent selects action $x_t \in \mathbb{S}^{d-1}$ through history.
- Agent receives loss(or reward) $l_t = x_t^\top L x_t + \eta_t$.

As a result, we have the following theorem.

Theorem 5.2. *The expected cumulative regret of FTRL with Sparse sampling algorithm (Kotlowski & Neu, 2019) on stochastic bandit PCA problem is bounded as follows:*

$$\mathbb{E}[R_T] \leq \tilde{O}\left(\sqrt{d^3 T}\right)$$

We defer the details to the Appendix F. The main advantage of this stochastic bandit PCA is that it requires only $\tilde{O}(dT)$ computational complexity (Kotlowski & Neu, 2019).

6. Discussion on the lower bound

One of the shortcomings in our study is the gap between the known regret lower bound ($\Omega(d\sqrt{T})$, Jun et al. (2019)) and the regret upper bound of our algorithm. Motivations mentioned in Section 4 also lead us to suspect that $\Omega(\sqrt{d^3 T})$ might be the minimax lower bound for the bilinear bandit problem, while a parallel work of Lattimore & Hao (2021) has proposed the existence of the algorithm with a better regret upper bound. In this section, we will briefly discuss about those evidences.

Signal to Noise Ratio Jun et al. (2019) provide the signal to noise ratio(SNR) as the evidence of the $\sqrt{d^3}$ term in the upper bound. Please refer to Section 6 of Jun et al. (2019) for the details.

Stochastic Bandit PCA As mentioned in the additional algorithm section, while studying Bandit PCA, stochastic bandit PCA was able to obtain only the regret of order $\tilde{O}(\sqrt{d^3 T})$, unlike adversarial bandit PCA regret $\tilde{O}(\sqrt{rd^2 T})$. The reason for this difference was intriguing because the noise factor completely obscures the parameter's properties, similar to the relationship between the adversarial linear bandit and the stochastic linear bandit.

In Appendix F, we bound the regret of the online mirror descent algorithm by the following inequality.

$$R_T \leq \frac{d \log T}{\eta} + \eta \times \sum_t B_t$$

The main difference between stochastic and adversarial bandit PCA problem comes from the calculation of B_t:

- Adversarial : $B_t \leq \cdots \leq d\|L_t\|_F^2 \leq dr$
- Stochastic : $B_t \leq \cdots \leq d\|L\|_F^2 + d^2\sigma^2 \leq dr + d^2\sigma^2$

Here, this new term $d^2\sigma^2$ is created by the sum of noises and has a larger dimensional dependency than the term created by the original loss matrix. Therefore, no matter what property does the hidden matrix L possesses, all of which are obscured by the noise term.

A similar phenomenon happens in the linear bandit problem. Apparently, contradictory result between the upper bound for adversarial bandits on the unit ball and the lower bound for stochastic bandits for the unit ball is one of the famous phenomenons in the linear bandit field (Bubeck et al., 2012; Lattimore & Szepesvári, 2020). From the close relationship between the linear bandit and the bilinear bandit, and from the SNR ratio analysis, we can expect that our Algorithm 1 might be asymptotically optimal.

Bandit Phase Retrieval On the other hand, Lattimore & Hao (2021) suggests the possibility of $\tilde{O}(d\sqrt{T})$ bilinear bandit algorithm by analyzing the bandit phase retrieval problem, which is a sub-problem to our bilinear bandit problem. The work of Lattimore & Hao (2021) is a tight result of the known regret lower bound ($\Omega(d\sqrt{T})$, (Jun et al., 2019)), and similar strategies may lead to the bilinear bandit algorithm with the regret upper bound $\tilde{O}(d\sqrt{T})$. Note that for the case where the left and right arm sets are both the unit balls and the parameter Θ^* is symmetric, one can apply their algorithm to solve the bilinear problem with regret $\tilde{O}(d\sqrt{T})$. Whether or not the same is true for the more generic bilinear problems and whether or not $\mathsf{rank}(\Theta^*)$ affects the regret upper bound are important open problems for bilinear bandits.

Proving or refuting these lower bound conjectures will be a meaningful research subject in the future. Although Jun et al. (2019) verified a lower bound of $\Omega(d\sqrt{T})$ through the singleton action set case, it was hard to be generalized to the action spaces with multiple actions since the lower bound calculation of the bilinear bandit requires computing the cross-terms of the paired action. Interested readers can check our lower bound analysis in the Appendix G, which is about the lower bound of the nontrivial action spaces.

7. Conclusion

In this paper, we have proposed new algorithms that enjoy either improved regret bound or much better numerical performance over prior art. Specifically, by focusing on the action set dimension, ϵ-FALB achieves an improved regret bound that disproves a conjectured optimal regret rate from Jun et al. (2019). Furthermore, our algorithm rO-UCB achieves significantly better numerical results over existing algorithms by leveraging our novel concentration inequality, which allows us to avoid forced exploration.

Our new results tell us that we are yet far from understanding the optimal regret rate for bandits with matrix parameters, which opens up numerous future directions. First, studying the optimal regret of bilinear bandits with the landmark arm sets like the unit ball or finite set remains to be a challenging open problem. Second, it seems that UCB-type algorithms with the adaptive design confidence inequalities are not amenable to exploiting the action set's true dimension, as far as known proof techniques are concerned. While fixed design confidence bounds lead to tighter theoretical bounds for finite arm sets such as SupLinRel-type algorithms, the community has seen that algorithms based on the adaptive design confidence bounds such as OFUL are simple yet enjoy better empirical performance. It would be interesting to develop novel algorithmic frameworks that can exploit the true dimension of the action set, which can lead to practical algorithms with tighter regret guarantees.

8. Acknowledgements

The first and fourth authors' work was supported by the National Research Foundation of Korea (NRF) grants funded by the Korea government (MOE, MSIT) (NRF-2017R1A2B4011546). The first, third, and fourth authors thank support by the Stochastic Analysis and Application Research Center (SAARC) (NRF-2019R1A5A1028324). The fourth author was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2019-0-00075, Artificial Intelligence Graduate School Program(KAIST)). The authors thank Chicheng Zhang for sharing a seed idea for this paper.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pp. 2312–2320, 2011.

Abbasi-Yadkori, Y., Pal, D., and Szepesvari, C. Online-to-confidence-set conversions and application to sparse stochastic bandits. In *Artificial Intelligence and Statistics*, pp. 1–9, 2012.

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

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

Beygelzimer, A., Langford, J., Li, L., Reyzin, L., and Schapire, R. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 19–26, 2011.

Bhojanapalli, S., Neyshabur, B., and Srebro, N. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 3873–3881, 2016.

Brochu, E., Cora, V. M., and De Freitas, N. A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*, 2010.

Bubeck, S., Cesa-Bianchi, N., and Kakade, S. M. Towards minimax policies for online linear optimization with bandit feedback. In *Conference on Learning Theory*, pp. 41–1. JMLR Workshop and Conference Proceedings, 2012.

Burer, S. and Monteiro, R. D. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.

Candes, E. J. and Plan, Y. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011.

Chapelle, O. and Li, L. An empirical evaluation of thompson sampling. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural*

Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain, pp. 2249–2257, 2011.

Chi, Y., Lu, Y. M., and Chen, Y. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, 2019.

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

Dani, V., Hayes, T. P., and Kakade, S. M. Stochastic linear optimization under bandit feedback. In *21st Annual Conference on Learning Theory - COLT 2008, Helsinki, Finland, July 9-12, 2008*, pp. 355–366. Omnipress, 2008.

Das, S. and Kamenica, E. Two-sided bandits and the dating market. In *IJCAI*, volume 5, pp. 19. Citeseer, 2005.

Hamidi, N., Bayati, M., and Gupta, K. Personalizing many decisions with high-dimensional covariates. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 11469–11480, 2019.

Hao, B., Zhou, J., Wen, Z., and Sun, W. W. Low-rank tensor bandits. *arXiv preprint arXiv:2007.15788*, 2020.

Johnson, N., Sivakumar, V., and Banerjee, A. Structured stochastic linear bandits. *arXiv preprint arXiv:1606.05693*, 2016.

Jun, K., Willett, R., Wright, S., and Nowak, R. D. Bilinear bandits with low-rank structure. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3163–3172. PMLR, 2019.

Katariya, S., Kveton, B., Szepesvári, C., Vernade, C., and Wen, Z. Stochastic rank-1 bandits. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, volume 54 of *Proceedings of Machine Learning Research*, pp. 392–401. PMLR, 2017.

Keshavan, R. H., Montanari, A., and Oh, S. Matrix completion from noisy entries. In *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. Proceedings of a meeting held 7-10 December 2009, Vancouver, British Columbia, Canada*, pp. 952–960. Curran Associates, Inc., 2009.

Kotlowski, W. and Neu, G. Bandit principal component analysis. In *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, volume 99 of *Proceedings of Machine Learning Research*, pp. 1994–2024. PMLR, 2019.

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

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

Lattimore, T. and Hao, B. Bandit phase retrieval, 2021.

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

Li, L., Chu, W., Langford, J., and Schapire, R. E. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010*, pp. 661–670. ACM, 2010. doi: 10.1145/1772690.1772758.

Li, Y., Wang, Y., and Zhou, Y. Nearly minimax-optimal regret for linearly parameterized bandits. In *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, volume 99 of *Proceedings of Machine Learning Research*, pp. 2173–2174. PMLR, 2019.

Lu, Y., Meisami, A., and Tewari, A. Low-rank generalized linear bandit problems. In *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, 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):1–13, 2017.

Rigollet, P. 18.s997: High dimensional statistics, 2015.

Soare, M., Lazaric, A., and Munos, R. Best-arm identification in linear bandits. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pp. 828–836, 2014.

Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. W. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, pp. 1015–1022. Omnipress, 2010.

Sui, Y., Zoghi, M., Hofmann, K., and Yue, Y. Advancements in dueling bandits. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pp. 5502–5510. ijcai.org, 2018. doi: 10.24963/ijcai.2018/776.

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.

Valko, M., Korda, N., Munos, R., Flaounas, I. N., and Cristianini, N. Finite-time analysis of kernelised contextual bandits. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence, UAI 2013, Bellevue, WA, USA, August 11-15, 2013*. AUAI Press, 2013.

Valko, M., Munos, R., Kveton, B., and Kocák, T. Spectral bandits for smooth graph functions. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, volume 32 of *JMLR Workshop and Conference Proceedings*, pp. 46–54. JMLR.org, 2014.

Yu, T., Kveton, B., Wen, Z., Zhang, R., and Mengshoel, O. J. Influence diagram bandits: Variational thompson sampling for structured bandit problems. *arXiv preprint arXiv:2007.04915*, 2020.

Zhang, L., Yang, T., Jin, R., Xiao, Y., and Zhou, Z. Online stochastic linear optimization under one-bit feedback. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pp. 392–401. JMLR.org, 2016.

Zimmert, J. and Seldin, Y. Factored bandits. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 2840–2849, 2018.

Appendix

A. Review of the linear bandit theory

In this section, we demonstrate why general action set approach usually has additional dimensional dependency –through an example– by analyzing the framework of LinUCB (Abbasi-Yadkori et al., 2011), a popular algorithm in this area.

The ESTR-lowOFUL algorithm of (Jun et al., 2019) is based on spectralUCB (Valko et al., 2014) and, by extension, on LinUCB (Abbasi-Yadkori et al., 2011). However, these algorithms have no restrictions on the action set other than its bound. Since the number of actions may be infinite, it is difficult to prove that all actions satisfy Eq. 2 at the same time. Instead, they focus on the fact that θ_* and $\hat{\theta}$ are close in l_2 sense as the following inequality:

Theorem A.1. *(Abbasi-Yadkori et al., 2011) At round t, the following inequality holds with probability at least $1-\delta$:*

$$\|\hat{\theta}_t - \theta_*\|_{V_t^{-1}} \leq \beta_t \,. \tag{6}$$

UCB-based linear bandit algorithms construct a confidence region $\mathcal{C}_t$ that contains θ_* with a probability of at least $1-\delta$ using Eq. 6, and calculate upper confidence bound of an action a as follows:

$$UCB_t(a) = \max_{\theta \in \mathcal{C}_t} \langle a, \theta \rangle \,.$$

This quantity can be interpreted as the maximum inner product value possible in the confidence region. Define $\bar{\theta} = \arg\max_{\theta \in \mathcal{C}_t} \langle x_t, \theta \rangle$. From the following inequalities

$$\langle \theta_*, x_* \rangle \leq UCB_t(x_*) \leq UCB_t(x_t) = \langle \tilde{\theta}, x_t \rangle \,, \tag{7}$$

The regret occurring at t can be bounded by $\bar{r}_t = \langle x_* - x_t, \theta_* \rangle \leq \langle x_t, \tilde{\theta} - \theta_* \rangle$. Bounding the second inner product is the core task of UCB-based algorithm regret analysis. However, since there is almost no restriction on the action set except the norm bound, we cannot assure that Eq. 2 holds for all actions. Therefore, we rely on Cauchy's inequality and obtain a loose bound as follows:

$$\begin{aligned}
\bar{r}_t &\leq \langle x_t, \tilde{\theta} - \theta_* \rangle \\
&\leq \langle x_t, \tilde{\theta} - \hat{\theta} \rangle + \langle x_t, \hat{\theta} - \theta_* \rangle \\
&\leq \|x_t\|_{V_t^{-1}} (\|\tilde{\theta} - \hat{\theta}\|_{V_t} + \|\hat{\theta} - \theta_*\|_{V_t}) \\
&\leq 2\|x_t\|_{V_t^{-1}} \beta_T \,.
\end{aligned}$$

Finally the regret is summarized as $R_T = \sum_{t=1}^T \bar{r}_t \leq 2\beta_T \sum_{t=1}^T \|x_t\|_{V_t^{-1}}$. In this calculation, note that β_t of Eq. 6 is usually much larger than the bound of Eq. 2 (e.g., in LinUCB it is about $\sqrt{d}$ order larger). This additional loss comes mainly from the step using Cauchy's inequality, which bounds only the inner product by l_2 sense along all possible directions. On the other hand, Eq. 2 bounds the inner product for a specific direction, which leads to a tighter bound.

For the summation $\sum_{t=1}^T \|x_t\|_{V_t^{-1}}$, we note that the geometry of the action space is not exploited well except for $span(\mathcal{A})$. This summation is usually bounded by the following lemma,

Lemma A.2. *(Elliptic Potential Lemma, Abbasi-Yadkori et al. (2011))*

$$\begin{aligned}
\min(1, \|x_t\|_{V_t^{-1}}^2) &\leq 2\log\frac{|V_t|}{|V_0|} \\
&\leq 2d\log\frac{Tr(V_0)+T}{d} - 2\log|V_0| \,.
\end{aligned}$$

The first inequality in the elliptic potential lemma is known to be relatively tight (Li et al., 2019), but it is not easy to control directly the determinant of the gram matrix V_t. Therefore, the second inequality in Lemma A.2 uses the determinant-trace inequality to bound the sum by a tractable value. This bound may not help since this determinant-trace inequality ignores most of the geometry of the action set. For example, the bound in this lemma cannot distinguish the action set of d canonical vectors and the infinite action set of whole sphere, S^{d-1}, since both sets span $\mathbb{R}^d$. Jun et al. (2019) considered the low-rank parameter structure by modifying V_0, but no known method exploits the exact geometry of the action set by manipulating V_0.

In short, previous studies analyze the linear bandit algorithm for the general action set along the following framework.

- It is hard to construct an event that Eq. 2 holds for all actions since the number of actions could be infinite. Instead, consider a bound on the distance between $\hat{\theta}$ and θ_* in l_2 sense.
- Based on the confidence ellipsoid generated through Eq. 6, calculate the regret upper bound by separating the action norm and the $\hat{\theta}$ estimation error through Cauchy's inequality. This process may produce a large regret upper bound due to the inefficiency of the Cauchy's inequality.

B. SupLinUCB modification

For this section, please recall the notations in Section 4.1. In this section, we will analyze about the case when the action set $\mathcal{A}$ is fixed for the notational convenience. However, this algorithm can also be applied to the changing arm set case as the original SupLinUCB algorithm is for the contextual bandit with linear payoff function. The proof does not change much when $\mathcal{A}$ changes over time. Plus, from the assumptions in Section 3 we can naturally assume that the absolute value of the mean reward $|\langle a, \theta_* \rangle|$ is bounded by 1 for all action $a \in \mathcal{A}$, and $\|\theta_*\| \le C \le \sqrt{d}$ for some constant C.

We analyze SupLinUCB algorithm with $V_0 = \frac{1}{\lambda} I_d$, and choose an appropriate λ later.

The main differences from Chu et al. (2011) are:

- The original paper (Chu et al., 2011) only deals with bounded reward models. We extend this case to the sub-Gaussian noise model.
- From the reason briefly discussed in Section 4.3, we change the initial matrix from $V_0 = I$ to $V_0 = \frac{1}{\lambda} I$ for some $\lambda > 1$. Due to this change, one has to modify several proof steps. For example, Lemma 2 of Chu et al. (2011) is no longer applicable.
- The main proof of SupLinUCB is based on the framework of SupLinRel (Auer et al., 2002), but this framework overlooked the scale of $\ln K$, which leads to an excessive estimation of the order of $\ln K$. We decrease the order of $\ln K$ term since $\ln K$ is crucial in our approach.

Original SupLinUCB paper uses Azuma Hoeffding's inequality to bound the inner product $\langle x, \hat{\theta} - \theta_* \rangle$. Here, we use another inequality to bound general sub-Gaussian random variables.

Lemma B.1. *Valko et al. (2014, Lemma 7) : For any fixed $x \in \mathbf{R}^d$ and any $\delta > 0$, we have that if $\beta_0 = 2\sigma\sqrt{14 \log \frac{2}{\delta}} + \frac{\|\theta_*\|}{\sqrt{\lambda}}$, then at time t, with probability greater than $1 - \delta$:*

$$|\langle x, \theta_* - \hat{\theta} \rangle| \le \beta_0 \|x\|_{V_t^{-1}}$$

From Lemma B.1, we get a confidence bound width by adapting the proof of Lemma 15 in Auer et al. (2002):

Lemma B.2. *(Chu et al., 2011) Let $\beta = 2\sigma\sqrt{14 \log \frac{2TK \ln T}{\delta}} + \frac{C}{\sqrt{\lambda}}$. Then with probability at least $1 - \delta$, for all time $t \in [T]$ and $s \in [S]$, the followings hold:*

- *$|\langle a, \theta_* - \hat{\theta}_t \rangle| \le w^s_{t,a}$ for all $a \in \mathcal{A}_s$;*
- *$a^* \in \hat{A}_s$, where a^* is the best action in hindsight.*
- *$|\langle a^* - a, \theta_* \rangle| \le 8 \times 2^{-s}$ for all $a \in \mathcal{A}_s$*

Note here that if $\sqrt{\log TK} \ll \frac{C}{\sqrt{\lambda}}$, the bound β is dominated by the factor $\frac{C}{\sqrt{\lambda}}$. The regret upper bound is usually proportional to the confidence bound width, and it might be too massive to be useful when $\sqrt{\log TK} \ll \frac{C}{\sqrt{\lambda}}$. In the bilinear bandit problem, it might be crucial since higher rank of the hidden parameter matrix makes the upper bound of the Frobenius norm looser. To avoid this, we define $\lambda = C^2$.

Denote the event mentioned in Lemma B.2 as E'. We discuss the regret bound under this event. Let Φ_0 be the set of time steps for which an action is chosen by the condition $w_{t,a}^s \le \frac{1}{\sqrt{T}}$ for all $a \in \mathcal{A}_s$. Based on Lemma B.2,

$$
\begin{aligned}
R_T &= \sum_{t=1}^{T} (a^* - a_t)^\top \theta_* \\
&= \sum_{t\in\Phi_0} (a^* - a_t)^\top \theta_* + \sum_{s=1}^{S} \sum_{t\in\Phi_s} (a^* - a_t)^\top \theta_* \\
&\le \sum_{t\in\Phi_0} a_t^\top (\bar{\theta}_t - \theta_*) + \sum_{s=1}^{S} \sum_{t\in\Phi_s} (a^* - a_t)^\top \theta_* \\
&\le \frac{2}{\sqrt{T}} |\Phi_0| + \sum_{s=1}^{S} \sum_{t\in\Phi_s} \min\{(a^* - a_t)^\top \theta_*, 2\} \\
&\le \frac{2}{\sqrt{T}} |\Phi_0| + \sum_{s=1}^{S} \sum_{t\in\Phi_s} \min\{2^{3-s}, 2\} \\
&\le \frac{2}{\sqrt{T}} |\Phi_0| + 8 \sum_{s=1}^{S} \sum_{t\in\Phi_s} \min\{w_{t,a_t}^s, 1\}
\end{aligned}
$$

The first inequality is from the same UCB trick in Eq. 7, and the second inequality is from the fact that $w_{t,a}^0 \le \frac{1}{\sqrt{T}}$ holds at all time steps in Φ_0, and the boundedness of $\langle a, \theta_* \rangle$ for all $a \in \mathcal{A}$. The third inequality comes from the Lemma B.2, and the final inequality is from the definition of a_t for $t \in \Phi_s$.

SupLinUCB strongly uses the property that the least eigenvalue of V_t is greater than or equal to 1 (since $V_0 = I_d$), which is not applicable to our case $V_0 = \frac{1}{\lambda} I_d$. Hence, we use Eq. A.2 instead of Lemma 2 and 3 of Chu et al. (2011). Now applying Eq. A.2 for each Φ_s leads the following result:

$$
\begin{aligned}
\sum_{t\in\Phi_s} \min(w_{t,a_t}^s, 1) &\le \max(\beta, 1) \sum_{t\in\Phi_s} \min(1, \|a_t\|_{V_{s,t}^{-1}}) \\
&\le \max(\beta, 1) \sqrt{|\Phi_s| \sum_{t\in\Phi_s} \min(1, \|a_t\|_{V_{s,t}^{-1}}^2)} \\
&\le \max(\beta, 1) \sqrt{|\Phi_s| (2d \log \frac{\frac{d}{\lambda} + |\Phi_s|}{d} + d \log \lambda)} \\
&\le \max(\beta, 1) \sqrt{|\Phi_s| (2d \log \frac{\frac{d}{\sqrt{\lambda}} + \sqrt{\lambda}|\Phi_s|}{d})} \\
&\le \tilde{O}(\sqrt{d|\Phi_s| \ln K}) \,.
\end{aligned}
$$

Finally the regret is bounded as follows:

$$
\begin{aligned}
R_T &\le \frac{2}{\sqrt{T}} |\Phi_0| + 8 \sum_{s=1}^{S} \sum_{t\in\Phi_s} \min(w_{t,a_t}^s, 1) \\
&\le 2\sqrt{T} + \sum_{s\in[S]} \tilde{O}(\sqrt{d|\Phi_s| \ln K}) \\
&\le 2\sqrt{T} + \tilde{O}(\sqrt{dST \ln K}) = \tilde{O}(\sqrt{dT \ln K}) \,.
\end{aligned}
$$

For the last inequality, we use Cauchy's inequality: $(\sum_{s\in[S]} \sqrt{|\Phi_s|})^2 \le S(\sum_{s\in[S]} |\Phi_s|) \le ST$.

C. Proof of Theorem 5.1

In this section, we derive the confidence bound width for rO-UCB algorithm. Here, we define C as the Frobenius norm bound for both $\hat{\Theta}_t$ and Θ^*. Since we know that $\lambda_{max}(\Theta^*) \leq 1$, one possible option for C might be $C = \sqrt{r}$.

Theorem C.1. *Suppose that $\hat{\Theta}_t$ is the solution of the Opt in Section 5. Then, with probability at least $1-\delta$ the following holds: for all $t \in \{1, \cdots, T\}$, $\hat{\Theta}_t$ satisfies*

$$\|vec(\hat{\Theta} - \Theta^*)\|_{W_t} \leq O\left(\sqrt{rd \log \frac{CT}{\delta}}\right).$$

when $\delta \geq \frac{1}{2\exp(T)}$

Proof. Using Cramer-Chernoff inequality for σ sub-Gaussian (Lattimore & Szepesvári (2020), Theorem 5.3) with the probability at least $1-\delta$, the following holds:

$$2\sum_{s=1}^{t} \eta_s((x_s^\top \Theta z_s) - (x_s^\top \Theta_* z_s)) \leq 2\sqrt{2\sigma^2 \sum_{s=1}^{t}((x_s^\top \Theta z_s) - (x_s^\top \Theta_* z_s))^2 \log \frac{1}{\delta}}. \tag{8}$$

Since we assumed $\max(\|\hat{\Theta}_t\|, \|\Theta_*\|) \leq C$, $\hat{\Theta}_t \in \Xi = \{\Theta \in \mathbf{R}^{d_1 \times d_2} : \mathsf{rank}(\Theta) \leq r, \|\Theta\| \leq C\}$. Since Ξ is a bounded set, its ϵ-covering set Ξ_ϵ has a finite cardinality. This means that for all $\Theta \in \Xi_\epsilon$, the following inequality holds with probability at least $1-\delta$:

$$2\sum_{s=1}^{t} \eta_s((x_s^\top \Theta z_s) - (x_s^\top \Theta_* z_s)) \leq 2\sqrt{2\sigma^2 \sum_{s=1}^{t}((x_s^\top \Theta z_s) - (x_s^\top \Theta_* z_s))^2 \log \frac{|\Xi_\epsilon|}{\delta}}. \tag{9}$$

However, we have to focus on the bound of $\hat{\Theta}_t$ not a fixed point Θ. For the bound of $\hat{\Theta}_t$, we first compute the bound of $|(x_s^\top \hat{\Theta}_t z_s) - (x_s^\top \Theta_\epsilon z_s)|$ where $\Theta_\epsilon \in \Xi_\epsilon$ is a point that satisfies $\|\hat{\Theta}_t - \Theta_\epsilon\|_F \leq \epsilon$. Since $\|x\|, \|z\| \leq 1$ for all $x \in \mathcal{X}, z \in \mathcal{Z}$, the following inequality holds for all round $s \in [T]$.

$$|(x_s^\top \hat{\Theta}_t z_s) - (x_s^\top \Theta_\epsilon z_s)| = |(x_s^\top (\hat{\Theta}_t - \Theta_\epsilon) z_s)| \leq \epsilon \tag{10}$$

The last inequality comes from the fact that the maximum singular value of a matrix is smaller than its Frobenius norm.

Let $Obj(\Theta) = \sum_{s=1}^{t}(x_s^\top \Theta z_s - r_s)^2 = \sum_{s=1}^{t}(x_s^\top \Theta z_s - x_s^\top \Theta_* z_s - \eta_s)^2$. By the minimality of $\hat{\Theta}_t$, the following inequality holds for $\Theta^* \in \Xi$:

$$Obj(\hat{\Theta}_t) - Obj(\Theta_*) = \sum_{s=1}^{t}((x_s^\top \hat{\Theta}_t z_s) - (x_s^\top \Theta_* z_s))^2 - 2\sum_{s=1}^{t} \eta_s((x_s^\top \hat{\Theta}_t z_s) - (x_s^\top \Theta_* z_s)) \leq 0. \tag{11}$$

In addition, by (10) and the fact $|((x_s^\top \Theta_\epsilon z_s) + (x_s^\top \hat{\Theta}_t z_s) - 2(x_s^\top \Theta_* z_s))| \leq |((x_s^\top \Theta_\epsilon z_s)| + |(x_s^\top \hat{\Theta}_t z_s)| + 2|(x_s^\top \Theta_* z_s))| \leq 4C + o(\epsilon)$, we obtain the following result

$$\begin{aligned}
&\sum_{s=1}^{t}\left[((x_s^\top \Theta_\epsilon z_s) - (x_s^\top \Theta_* z_s))^2 - ((x_s^\top \hat{\Theta}_t z_s) - (x_s^\top \Theta_* z_s))^2\right] \\
&= \sum_{s=1}^{t}((x_s^\top \Theta_\epsilon z_s) - (x_s^\top \hat{\Theta}_t z_s))((x_s^\top \Theta_\epsilon z_s) + (x_s^\top \hat{\Theta}_t z_s) - 2(x_s^\top \Theta_* z_s)) \\
&\leq \sum_{s=1}^{t} \epsilon \times 4C + O(\epsilon^2) = 4Ct\epsilon + O(\epsilon^2 t).
\end{aligned} \tag{12}$$

This implies, ignoring constant and ϵ^2 terms,

$$\begin{aligned}
\sum_{s=1}^{t}((x_s^\top \Theta_\epsilon z_s) - (x_s^\top \Theta_* z_s))^2 &\leq \sum_{s=1}^{t}((x_s^\top \hat{\Theta}_t z_s) - (x_s^\top \Theta_* z_s))^2 + 4Ct\epsilon \\
&\leq 2\sum_{s=1}^{t} \eta_s((x_s^\top \hat{\Theta}_t z_s) - (x_s^\top \Theta_* z_s)) + 4Ct\epsilon
\end{aligned}$$

$$\begin{aligned}
&= 2\sum_{s=1}^{t} \eta_s[(x_s^\top \hat{\Theta}_t z_s) - (x_s^\top \Theta_\epsilon z_s)] + 2\sum_{s=1}^{t} \eta_s[(x_s^\top \Theta_\epsilon z_s) - (x_s^\top \Theta_* z_s)] + 4Ct\epsilon \\
&\le 2\sum_{s=1}^{t} |\eta_s|\epsilon + 2\sqrt{2\sigma^2 \sum_{s=1}^{t}((x_s^\top \Theta_\epsilon z_s) - (x_s^\top \Theta_* z_s))^2 \log\frac{|\Xi_\epsilon| t}{\delta}} + 4Ct\epsilon\,. \qquad (13)
\end{aligned}$$

Here,

- The first inequality holds by (12);
- The second inequality is by (11);
- The third inequality can be achieved by applying triangular inequality on the first term, and applying (9) on the second term (holds with probability $1-\delta$).

As will be discussed later in section F.2, the sub-Gaussian maxima inequality shows that $\max_{s=1}^{T} |\eta_s| \le O(\sqrt{\sigma^2 \log\frac{T}{\delta'}})$ with probability $1-\delta'$, which means $\frac{\sum_{s=1}^{t}|\eta_s|}{T^2} \le O(\frac{\sqrt{\log(T/\delta')}}{T})$ for all t. By letting $\epsilon = \frac{1}{T^2}$ and $\delta' = \delta$, the following inequality holds with probability $1-2\delta$:

$$\begin{aligned}
\sqrt{\sum_{s=1}^{t}((x_s^\top \Theta_\epsilon z_s) - (x_s^\top \Theta_* z_s))^2} &\le 2\sqrt{2\sigma^2 \log\frac{|\Xi_\epsilon| t}{\delta}} + O(\sqrt{\frac{\sqrt{\sigma^2\log(T/\delta)} + C}{T}}) \\
&\le \sqrt{8\sigma^2 r(d_1+d_2+1)\log(9CT^2/\delta)} + O(\sqrt{\frac{\sqrt{\sigma^2\log(T/\delta)} + C}{T}})
\end{aligned}$$

Here we used a Proposition 9 in Abbasi-Yadkori et al. (2012) on (13) for the first inequality – if $z^2 \le a + bz$ and $a, b > 0$, then $z \le b + \sqrt{a}$ by a simple quadratic equation computation. The second inequality is from Lemma 3.1 of Candes & Plan (2011), $\log|\Xi_\epsilon| \le r(d_1+d_2+1)\log\frac{9C}{\epsilon}$. Note that the order of the latter term is ignorable compare to the first term if $T \ge \frac{1}{\sigma^2}$.

Finally, combining Eq. 13 and Eq. 10 makes the following result:

$$\begin{aligned}
\sum_{s=1}^{t}((x_s^\top \hat{\Theta}_t z_s) - (x_s^\top \Theta_* z_s))^2 &\le 2\sum_{s=1}^{t}((x_s^\top \hat{\Theta}_t z_s) - (x_s^\top \Theta_\epsilon z_s))^2 + 2\sum_{s=1}^{t}((x_s^\top \Theta_\epsilon z_s) - (x_s^\top \Theta_* z_s))^2 \\
&\le 2\epsilon^2 t + 16\sigma^2 r(d_1+d_2+1)\log\frac{9CT^2}{\delta} \\
&\approx 16\sigma^2 r(d_1+d_2+1)\log\frac{9CT^2}{\delta}\,. \qquad (14)
\end{aligned}$$

Now we change Eq. 14 to be appropriate for the linear bandit form,

$$\begin{aligned}
&\sum_{s=1}^{t}(\langle vec(x_s z_s^\top), vec(\hat{\Theta} - \Theta_*)\rangle)^2 \le 16\sigma^2(r(d_1+d_2+1)\log(\frac{9CT^2}{\delta})) \\
\Rightarrow\ & vec(\hat{\Theta} - \Theta_*)^\top(\sum_{s=1}^{t} vec(x_s z_s^\top) vec(x_s z_s^\top)^\top) vec(\hat{\Theta} - \Theta_*) \le 16\sigma^2 r(d_1+d_2+1)\log(\frac{9CT^2}{\delta}) \\
\Rightarrow\ & \|vec(\hat{\Theta} - \Theta_*)\|_{W_t}^2 \le 16\sigma^2 r(d_1+d_2+1)\log(\frac{9CT^2}{\delta}) + 4C^2 \\
\Rightarrow\ & \|vec(\hat{\Theta} - \Theta_*)\|_{W_t} \le \sqrt{16\sigma^2 r(d_1+d_2+1)\log(\frac{9CT^2}{\delta}) + 4C^2}\,.
\end{aligned}$$

with probability at least $1-2\delta$ for each t. Let this event E_t. Now $\mathbb{P}(\cup_{t=1}^{T} E_t^c) \le 2T\delta$ by the union bound argument. Substituting δ to $\frac{\delta}{2T}$ leads the desired result. □

C.1. Regret upper bound of rO-UCB

About the regret upper bound of rO-UCB, one can follow the usual LinUCB proof introduced in Appendix A by replacing Theorem A.1 to Theorem C.1.

Let $\beta_T = \sqrt{16\sigma^2 r(d_1+d_2+1)\log(\frac{18\sqrt{r}T^3}{\delta}) + 4r}$. By Theorem C.1, with probability $1-\delta$, $\|vec(\hat{\Theta} - \Theta^*)\|_{W_t} \le \beta_T$. for all $t \in \{1, 2, \cdots, T\}$. By the steps introduced in Appendix A, $R_T \le 2\beta_T \sum_{t=1}^{T} \|vec(x_t z_t^T)\|_{W_t^{-1}}$. Now the regret of rO-UCB algorithm can be summarized as follows:

$$\begin{aligned} R_T &\le \beta_T \sum_{t=1}^{T} \|vec(x_t z_t^T)\|_{W_t^{-1}} \\ &\le \beta_T \left[\sqrt{T \sum_{t=1}^{T} \min(1, \|vec(x_t z_t^T)\|^2_{W_t^{-1}})} \right] \\ &\le \beta_T \sqrt{T(2d_1 d_2 \log \frac{Tr(W_0) + T}{d_1 d_2} - 2\log|W_0|)} \\ &\le \tilde{O}(\sqrt{r d_1 d_2 d T}) \end{aligned}$$

Here the second inequality comes from Cauchy's inequality and $\|vec(x_t z_t^T)\|_{W_t^{-1}} \le \|vec(x_t z_t^T)\|_{W_0^{-1}} = \|vec(x_t z_t^T)\| \le 1$. The third inequality is from the elliptic potential lemma (Lemma A.2).

D. Proof of Lemma 4.3 and Corollary 4.7

Lemma D.1. *(Lemma 4.3) For a bounded set $\mathcal{S} \subset \mathbb{R}^d$, its covering number $N(\mathcal{S}, \epsilon)$ satisfies the following inequality:*

$$N(\mathcal{S}, \epsilon) \le \frac{vol(\mathcal{S}' + \frac{\epsilon}{2}\mathcal{B}_d)}{vol(\frac{\epsilon}{2}\mathcal{B}_d)} \tag{15}$$

Here, $\mathcal{S}'$ is an arbitrary measurable set that contains $\mathcal{S}$, and $\mathcal{S}' + \frac{\epsilon}{2}\mathcal{B}_d$ is a sumset between $\mathcal{S}'$ and $\frac{\epsilon}{2}\mathcal{B}_d$.

Proof. From Lattimore & Szepesvári (2020), the packing number $M(\mathcal{S}, \epsilon)$ is always greater than or equal to $N(\mathcal{S}, \epsilon)$. Let P be a maximum cardinality ϵ-packing set of $\mathcal{S}$. Then for any $p, q \in P$, $(\frac{\epsilon}{2}\mathcal{B}_d + \{p\}) \cap (\frac{\epsilon}{2}\mathcal{B}_d + \{q\}) = \emptyset$. By definition $(\frac{\epsilon}{2}\mathcal{B}_d + \{a\}) \subset \mathcal{S}' + \frac{\epsilon}{2}\mathcal{B}_d$. Therefore, $\cup_{p\in P}(\frac{\epsilon}{2}\mathcal{B}_d + \{p\}) \subset \mathcal{S}' + \frac{\epsilon}{2}\mathcal{B}_d$. By the monotonicity and additivity of the volume measure, $\mathrm{vol}(\cup_{p\in P}(\frac{\epsilon}{2}\mathcal{B}_d + \{p\})) = M(\mathcal{S}, \epsilon)\mathrm{vol}(\frac{\epsilon}{2}\mathcal{B}_d) \le \mathrm{vol}(\mathcal{S}' + \frac{\epsilon}{2}B)$, and the theorem holds. □

Corollary D.2. *(Corollary 4.7) Let $O_{d',\rho} = \{M \in \mathbf{R}^{d' \times \rho} : MM^\top = I_{d'}\}$, $D_\rho = \{Diag(\theta) : \theta \in [-1, 1]^\rho\}$. Suppose that we have three sets $\mathcal{X} \subset O_{d_1,\rho}, \mathcal{Z} \subset O_{d_2,\rho}, \mathcal{D} \subset D^\rho$, and the action set $\mathcal{A}$ is defined as $\mathcal{A} = \{vec(U\Sigma V^\top) : U \in \mathcal{X}, V \in \mathcal{Z}, \Sigma \in D^\rho\}$. Now if we perform Algorithm 2 and Algorithm 4 with the action set $\mathcal{A}_\epsilon$, the ϵ-covering set of $\mathcal{A}$, we get the following regret upper bound:*

$$R_T \le O(\sqrt{d_1 d_2 \rho(d_1 + d_2)\log(T)})$$

First, we will prove the following lemma about the cardinality of the $\mathcal{A}_\epsilon$:

Lemma D.3. $|\mathcal{A}_\epsilon| \le (\frac{9\rho}{\epsilon})^{\rho(d_1+d_2+1)}$

Proof. Let's decompose $\mathcal{A}$ into $\mathcal{X}$, $\mathcal{Z}$, and $\mathcal{D}$. Consider the cardinality of the $\epsilon/3$-covering set of each set.

- First, since $\mathcal{X} \subset O_{d_1,\rho} \subset B_{d_1\rho}(\sqrt{\rho})$, $N(\mathcal{X}, \epsilon/3) \le (\frac{9\rho}{\epsilon})^{d_1\rho}$ by the Lemma 4.3;
- Similarly, $N(\mathcal{Z}, \epsilon/3) \le (\frac{9\rho}{\epsilon})^{d_1\rho}$.

For $\Sigma_1, \Sigma_2 \in \mathcal{D}$, $\|\Sigma_1 - \Sigma_2\|_F = \|diag(\Sigma_1) - diag(\Sigma_2)\|$. Thus $N(\mathcal{D}, \epsilon/3) \le N([-1, 1]^\rho, \epsilon/3) \le N(B_\rho(\sqrt{\rho}), \epsilon/3) \le (\frac{9\rho}{\epsilon})^\rho$. This implies, if we let $\mathcal{A}' = \{vec(U\Sigma V^\top) : U \in \mathcal{X}_{\epsilon/3}, V \in \mathcal{Z}_{\epsilon/3}, \Sigma \in D^\rho_{\epsilon/3}\}$, then this $\mathcal{A}'$ is the ϵ-covering set for $\mathcal{A}$ by following the logic of Lemma 3.1 of Candes & Plan (2011) (note that we are not sure about minimality). Therefore, $N(\mathcal{A}, \epsilon) \le (\frac{9\rho}{\epsilon})^{\rho(d_1+d_2+1)}$ holds, and the proof is finished. □

After obtaining this cardinality bound, all the steps follow the framework of Theorem 4.6. For the notational convenience, let $\langle\langle W_1, W_2 \rangle\rangle := \langle vec(W_1), vec(W_2) \rangle$ for matrices $W_1, W_2 \in \mathbb{R}^{d_1 \times d_2}$. Define A_* as the optimal action in hindsight and $A_\epsilon = \arg\min_{A \in a_\epsilon} \|A_* - A\|$. Then,

$$
\begin{aligned}
R_T &= \sum_{t=1}^{T} \langle\langle A_*, \Theta \rangle\rangle - \sum_{t=1}^{T} \langle\langle A_t, \Theta \rangle\rangle \\
&= \sum_{t=1}^{T} \langle\langle A_*, \Theta \rangle\rangle - \sum_{t=1}^{T} \langle\langle A_\epsilon, \Theta \rangle\rangle \\
&+ \sum_{t=1}^{T} \langle\langle A_\epsilon, \Theta \rangle\rangle - \sum_{t=1}^{T} \max_{A \in \mathcal{A}_\epsilon} \langle\langle A, \Theta \rangle\rangle \\
&+ \sum_{t=1}^{T} \max_{A \in \mathcal{A}_\epsilon} \langle\langle A, \Theta \rangle\rangle - \sum_{t=1}^{T} \langle\langle A_t, \Theta \rangle\rangle \\
&= R_1 + R_2 + R_3 \,.
\end{aligned}
$$

Here,

- By definition, $R_2 \leq 0$;
- R_3 can be bounded by $O(\sqrt{d_1 d_2 T \log \frac{K}{\delta}})$ by Theorem 4.4 or Theorem 4.5;
- Lastly, R_1 is bounded by the following ϵ-covering argument:

$$
\begin{aligned}
&\sum_{t=1}^{T} \langle\langle A_*, \Theta \rangle\rangle - \sum_{t=1}^{T} \langle\langle A_\epsilon, \Theta \rangle\rangle \\
&\quad = \sum_{t=1}^{T} \langle\langle A_* - A_\epsilon, \Theta \rangle\rangle \\
&\quad \leq \sum_{t=1}^{T} \|\Theta\|_F \cdot \|A_* - A_\epsilon\|_F \\
&\quad \leq \sum_{t=1}^{T} \sqrt{d} \lambda_{max} \epsilon \,.
\end{aligned}
$$

Sum up R_1, R_2 and R_3 leads

$$
\begin{aligned}
R_T &\leq R_1 + R_2 + R_3 \\
&\leq \epsilon T \sqrt{d} \lambda_{max} + 0 + \tilde{O}(\sqrt{d_1 d_2 T \ln(K)}) \,.
\end{aligned}
$$

Substituting $\epsilon = \frac{\sqrt{d_1 d_2 \rho}}{\sqrt{T}}$ with the fact $\ln K \leq \rho(d_1 + d_2 + 1) \ln \frac{9\sqrt{T\rho}}{\sqrt{d_1 d_2}}$ leads desired $\tilde{O}(\sqrt{d_1 d_2 d \rho T})$ regret upper bound.

E. Additional experimental results

E.1. Experiment about the best ϵ

We have performed an additional experiment for various ϵ as shown in Figure 2. We set the experiment setting as follows:

- $\mathcal{X} = \mathcal{Z} = \mathbb{S}^1$, unit circles, $d = 2$
- Noise: Gaussian distribution with standard deviation $\sigma = 0.01$
- Number of total rounds: $T = 2500$
- Number of repeated experiments: 60
- $\Theta = \begin{bmatrix} 1 & 0 \\ 0 & 0.3 \end{bmatrix}$, best arm $x^* = z^* = [1, 0]$

- Size of ϵ: $8\epsilon_0, 4\epsilon_0, 2\epsilon_0, \epsilon_0$, where $\epsilon_0 = \frac{\pi}{80} < \frac{1}{\sqrt{T}}$.
- Covering set $\mathcal{X}_\epsilon = \mathcal{Z}_\epsilon = \{(\cos(\theta), \sin(\theta)) : \theta = (i + 0.5)\epsilon, i = 0, 1, \cdots, \frac{\pi}{\epsilon} - 1\}$
- Finite armed linear bandit algorithm for ϵ-FALB: SupLinUCB

We chose the covering set as the farthest one from the true best arm to show the effect of the coarse ϵ-covering set. We have found that a coarse ϵ-covering set construction could severely harm the regret when ϵ is over $d/\sqrt{T}$, and proved this observation through more precise theoretical calculations.

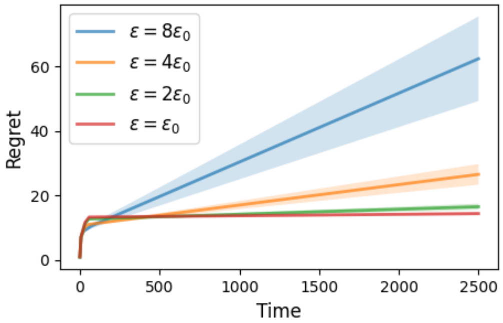

Figure 2. Simulation result for the ϵ-FALB algorithm. $d = 2$, $\mathcal{X} = \mathcal{Z} = \mathbb{S}^1$, $\Theta^* = \mathbf{diag}(1, 0.3)$, $T = 2500$, and $\sigma = 0.01$. Here $\epsilon_0 = 1/\sqrt{T}$ is the parameter of the covering set from the theoretical analysis in our paper.

E.2. Experiment about the inefficiency of ϵ-FALB

Before describing the experiment result, we have to explain how we calculate argmax in the continuous action sets. One of the main obstacle in the UCB-like setting is $\arg\max_{x,z \in \mathcal{X}\times\mathcal{Z}} UCB_t(x, z)$, which is computationally difficult for the continuous action sets. Note that it can be hard even for the standard linear bandits (cf. Section 19.3.1 in Lattimore & Szepesvári (2020), Section 3.4 in Dani et al. (2008)). One can use the global search heuristics (Srinivas et al., 2010; Brochu et al., 2010) which are known to be effective in practice. In our case, we use alternating maximization. Note that UCB function is a summation of a hyperplane and a cone, which is quite exploitable geometry. Few more calculation suggests that when one side of the action is fixed, the function $UCB_t(\cdot, z)$ is again the summation of a hyperplane and a cone, which implies when the action sets $\mathcal{X}$ and $\mathcal{Z}$ are sufficiently good, one can perform alternating maximization with few computations. Below are the experimental conditions.

- Left and right action sets: $\mathcal{B}_2(0)$
- Rank $r = 1$
- Noise: gaussian distribution with standard deviation $\sigma = 0.01$
- Number of total rounds: $T = 2500$
- Number of repeated experiments: 60
- Optimization method for alternating maximization: COBYLA
- Finite armed linear bandit algorithm for ϵ-FALB: SupLinUCB

Table 3. Experimental result of the inefficiency of ϵ-FALB.

RESULTS	LINUCB WITH COBYLA	ϵ-FALB ($\epsilon = 1/25$)	ϵ-FALB($\epsilon = 1/50$)
COMP. TIME (MIN)	3	3	35

This experiment shows how the spatial complexity growth of the ϵ-FALB affects the computational time – even when $d = 2$, the computation time is seriously longer than the LinUCB with COBYLA with $\epsilon = 1/50$, and the case of ϵ-FALB with $\epsilon = 1/25$ shows that this time disadvantage is mainly from the spatial complexity. Considering that the computational cost of the ϵ-FALB increases at a much faster rate (exponentially) as the dimension increases, you can see that the ϵ-FALB is computationally inefficient in virtually all cases.

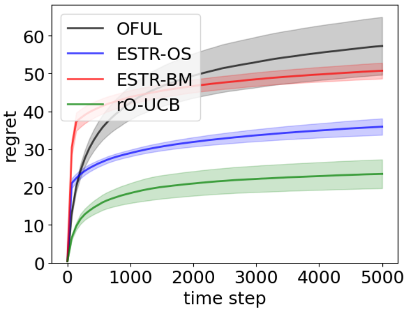

Figure 3. $T = 5000$, $R = 0.01$

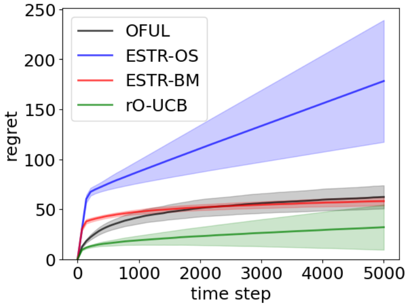

Figure 4. $T = 5000$, $R = 0.1$

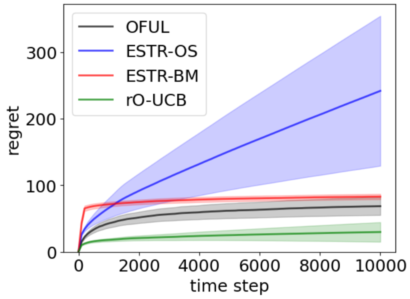

Figure 5. $T = 10000$, $R = 0.1$

Figure 6. Simulation result for various standard deviations (R) and total times (T). Our method outperforms all known general bilinear bandit algorithms, while the forced exploration algorithms ESTR-OS and ESTR-BM show excessive explorations or linear regret failures.

E.3. Additional experiments about noise and time

While performing our simulation experiments, we follow the experimental conditions of Jun et al. (2019), to compare our algorithm with their best condition. Since bandit problems in practice require fine-tuning the hyperparameters such as confidence bound width β_t (Chapelle & Li, 2011; Li et al., 2010; Zhang et al., 2016), we adjusted the confidence bound width β_t with $c\beta_t$ for OFUL, LowOFUL of Jun et al. (2019), and rO-UCB. For the fair comparison, we calibrated c by grid search and report the result with the smallest average regret. In addition, since ESTR-OS and ESTR-BM requires exploration time adjustment, we also tune exploration time T_1 to $C_{T_1}T_1$, and find the best C_{T_1} by grid search as in the experiment of Jun et al. (2019). Finally, our rO-UCB uses Burer & Monteiro (2003) instead of the oracle, which requires an initial point as the input of the algorithm. We set the initial point close to the true parameter Θ to make our optimization work as the true oracle.

- Left action dimension $d_1 = 8$, right action dimension $d_2 = 8$
- Rank $r = 1$
- Number of right and left action $|\mathcal{X}| = |\mathcal{Z}| = 16$
- Confidence bound width calibration constant: $c = 10^{-2}, 10^{-1.5}, 10^{-1}, 10^{-0.5}, 10^{0}$
- Exploration time multiplication constant: $C_{T_1} = 10^{-1}, 10^{-0.75}, 10^{-0.5}, 10^{-0.25}, 10^{0}$
- Noise: gaussian distribution with standard deviation $R = 0.01, 0.1$
- Number of total rounds: $T = 5000, 10000$
- Number of repeated experiments: 60

As one can see from the experimental results, forced exploration based algorithms (ESTR-OS, ESTR-BM) frequently fails or requires too much initial exploration when the condition about T or R changes. This means ESTR algorithms are not rigorous, and requires excessive amount of fine-tunning of exploration time and confidence bound width. On the other hand, our oracle based algorithm stably shows the best performance among all the other algorithms.

F. Proof of Theorem 5.2

F.1. Preliminary

Here we consider the stochastic bandit PCA model,

$$l_t = w_t^\top L w_t + \epsilon_t$$

where $w_t \in S^{d-1}$ is an action vector in round t. Basically, this section follows the steps of the proof framework of Algorithm 1 in Kotlowski & Neu (2019) with sparse sampling.

- W_t : Positive semidefinite density matrix which decides the distribution of the action w_t. $Tr(W_t) = 1$.
    - $\tilde{W}_{t+1}$: Intermediate $d \times d$ matrix after the update step of the mirror descent on W_t. Applying the projection step on $\tilde{W}_{t+1}$ is W_{t+1}. For more details, see Section 4 of Kotlowski & Neu (2019).
- $L : \mathbb{R}^{d\times d}$ hidden matrix. $L_t = L + \epsilon_t I$

- $\tilde{L}_t$: Estimator of the loss matrix L, and later we will show that it is an unbiased estimator. It is defined as follows:

$$\tilde{L}_t = \begin{cases} \frac{l_t}{\lambda_I^2} u_I u_I^\top & \text{if } i = j \\ \frac{s l_t}{2\lambda_I \lambda_J}(u_I u_J^\top + u_J u_I^\top) & \text{otherwise} \end{cases} \tag{16}$$

  - s: Uniform random variable with probability $1/2$ for each ± 1
  - I, J: Index random variable for sparse sampling in $[d] = \{1, \cdots, d\}$.
  - u_i: Eigenvectors of $W_t = \sum \lambda_i u_i u_i^\top$
  - l_{ij} : loss when i, j indices are chosen. It is defined as follows

$$l_{ij} = \begin{cases} u_i^\top L_t u_i & \text{if } i = j \\ \frac{1}{2}(u_i + s u_j)^\top L_t (u_i + s u_j) & \text{otherwise} \end{cases} \tag{17}$$

- η : Mirror descent step size
- γ : Forced uniform exploration probability. Details are in Kotlowski & Neu (2019) Algorithm 3, Sparse sampling section.
- $V = \frac{\gamma}{d} I + (1-\gamma) W_t = \sum \mu_i u_i u_i^\top$ ($\mu_i = \frac{\gamma}{d} + (1-\gamma)\lambda_i$)
- $B_t = W_t^{1/2} \tilde{L}_t W_t^{1/2}$, $\{b_{ti}\}$ are the eigenvalues of B_t.

We assume that noise ϵ_t is bounded, and the reward is always bounded by a constant C. This bounded noise assumption is just for convenience, and as will be discussed in the latter section, it is easily relaxed to the sub-gaussian condition.

The following lemma assures that $\tilde{L}_t$ is the unbiased estimator of L.

Lemma F.1. $\mathbb{E}_t[\tilde{L}_t] = L$

Proof.

$$\tilde{L}_t = \begin{cases} \frac{l_t}{\lambda_I^2} u_I u_I^\top = \frac{w_t^\top L w_t}{\lambda_I^2} u_I u_I^\top + \frac{\epsilon_t}{\lambda_I^2} u_I u_I^\top & \text{if } i = j \\ \frac{s l_t}{2\lambda_I \lambda_J}(u_I u_J^\top + u_J u_I^\top) = \frac{s w_t^\top L w_t}{2\lambda_I \lambda_J}(u_I u_J^\top + u_J u_I^\top) + \frac{s \epsilon_t}{2\lambda_I \lambda_J}(u_I u_J^\top + u_J u_I^\top) & \text{otherwise} \end{cases} \tag{18}$$

In any case, the latter term including ϵ_t has zero mean since ϵ_t is mean 0 independent noise under $\mathbb{E}_t$. The former term which includes $w_t^\top L w_t$ can be processed in the same way as in the Appendix A.2 of Kotlowski & Neu (2019), which implies $\mathbb{E}_t[\tilde{L}_t] = L$. □

After this, following the traditional step of the mirror descent framework, regret is bounded by the following form.

Lemma F.2.

$$R_T \leq \frac{d \log T}{\eta} + \gamma T + (1-\gamma) \sum_t \mathbb{E}[\langle\langle W_t - \tilde{W}_{t+1}, \tilde{L}_t \rangle\rangle] \,.$$

The above Lemma F.2 will be proven by the following Lemma F.3.

Lemma F.3. *For all positive semi-definite matrix U, the following holds.*

$$\sum_{t=1}^{\top} \langle\langle W_t - U, \tilde{L}_t \rangle\rangle \leq \frac{D_R(U \| W_1)}{\eta} + \sum_{t=1}^{\top} \langle\langle W_t - \tilde{W}_{t+1}, \tilde{L}_t \rangle\rangle \,.$$

The proof of this lemma follows the exactly same steps of Lemma 12 as in Kotlowski & Neu (2019), so we omit it.

Proof. (Proof of Lemma F.2) First we have to modify the left side of Lemma F.3 considering the form of the regret. Using Lemma F.1 and conditional independence with W_t the following relationship holds:

$$(1-\gamma)\mathbb{E}_t[\langle\langle W_t, \tilde{L}_t \rangle\rangle] = (1-\gamma)\langle\langle W_t, L \rangle\rangle = \mathbb{E}_t[\langle\langle w_t w_t^\top, L \rangle\rangle] - \frac{\gamma}{d}\langle\langle I, L \rangle\rangle$$

Here the last equality comes from the fact $\mathbb{E}_t[w_t w_t^\top] = \frac{\gamma}{d} I + (1-\gamma) W_t$. Now subtract $\langle\langle U, L \rangle\rangle$ on both sides where U is a positive semi-definite matrix of trace 1

$$\mathbb{E}_t[\langle\langle w_t w_t^\top - U, L\rangle\rangle] = (1-\gamma)\mathbb{E}_t[\langle\langle W_t - U, \tilde{L}_t\rangle\rangle] + \gamma\langle\langle \frac{I}{d} - U, L\rangle\rangle$$

Since the spectral norm of L is bounded by 1,

$$\langle\langle \frac{I}{d} - U, L\rangle\rangle \leq \|\frac{I}{d} - U\|_1 \|L\|_\infty \leq tr(\frac{I}{d} + U) \leq 2$$

Here the norm on the matrix is the Schatten norm. Now substitute U to $w_* w_*^\top$ where w_* is the optimal action, and applying Lemma F.3 on the above inequality leads the following regret bound:

$$\begin{aligned} R_T = \sum_{t=1}^{\top} \mathbb{E}[\langle\langle U - w_t w_t^\top, L\rangle\rangle] &= \sum_{t=1}^{\top} \mathbb{E}[\mathbb{E}_t[\langle\langle U - w_t w_t^\top, L_t\rangle\rangle]] \\ &\leq (1-\gamma)\mathbb{E}[\mathbb{E}_t[\langle\langle U - W_t, \tilde{L}_t\rangle\rangle]] + 2\gamma T \\ &\leq (1-\gamma)\frac{D_R(U\|W_1)}{\eta} + (1-\gamma)\sum_{t=1}^{\top} \mathbb{E}[\langle\langle W_t - \tilde{W}_{t+1}, \tilde{L}_t\rangle\rangle] + 2\gamma T\,. \end{aligned}$$

Here, one minor challenge is that $D_R(U\|W_1)$ is infinite since $\mathsf{rank}(U) = 1$. One can easily check how to deal with this issue referring the proof of Lemma 9 in Kotlowski & Neu (2019). □

The main difference between Kotlowski & Neu (2019) and our stochastic badit PCA comes from the calculation of the latter term of Lemma F.2, $\mathbb{E}[\langle\langle W_t - \tilde{W}_{t+1}, \tilde{L}_t\rangle\rangle]$. Readers are encouraged to continue comparing the proof of the Lemma 11 of Kotlowski & Neu (2019) and that in this paper.

Lemma F.4. *Suppose that the hyperparameter η and γ satisfy $\eta \leq \min(\frac{1}{2d}, \frac{1}{2C}, \frac{1}{2Cd})$ and $\gamma = Cd\eta$. Then the sparse sampling method guarantees*

$$\mathbb{E}[\langle\langle W_t - \tilde{W}_{t+1}, \tilde{L}_t\rangle\rangle] \leq 8\eta(dr + d^2\sigma^2)\,.$$

Proof. Let's recall the definition of the several variables before the proof.

- $W_t = \sum \lambda_i u_i u_i^\top$ is the eigenvalue decomposition of W_t
- $V = \frac{\gamma}{d}I + (1-\gamma)W_t = \sum \mu_i u_i u_i^\top$ ($\mu_i = \frac{\gamma}{d} + (1-\gamma)\lambda_i$)
- $B_t = W_t^{1/2}\tilde{L}_t W_t^{1/2}$, $\{b_{ti}\}_{i=1}^d$ are the eigenvalues of B_t.
- l_{ij} : loss when i, j indices are chosen. It is defined as follows

$$\tilde{l_{ij}} = \begin{cases} u_i^\top L_t u_i & \text{if } i = j \\ \frac{1}{2}(u_i + su_j)^\top L_t (u_i + su_j) & \text{otherwise} \end{cases}$$

In addition, note that from (3) in Kotlowski & Neu (2019), the following equalities hold:

$$\begin{aligned} \langle\langle W_t - \tilde{W}_{t+1}, \tilde{L}_t\rangle\rangle &= \eta Tr(B_t(I + \eta B_t)^{-1}B_t) \\ &= \sum_{i=1}^{d} \frac{\eta b_{ti}^2}{1 + \eta b_{ti}}\,. \end{aligned}$$

First, assume $I = J$. This event occurs with probability μ_i^2. From the definition of the sparse sampling algorithm, $\tilde{L}_t = \frac{l_{ii}}{\mu_i^2}u_i u_i^\top$ and $B_t = \frac{l_{ii}\lambda_i}{\mu_i^2}u_i u_i^\top$. In other word the only nonzero eigenvalue of B_t is $b_{t1} = \frac{l_{ii}\lambda_i}{\mu_i^2}$. From the bound $\mu_i^2 = ((1-\gamma)\lambda_i + \gamma/d)^2 \geq 4\frac{(1-\gamma)\gamma\lambda_i}{d} \geq 2\frac{\gamma\lambda_i}{d}$, we can bound $|b_{t1}| \leq \frac{|l_{ii}|d}{2\gamma} \leq \frac{1}{2\eta}$. Applying this bound, we obtain

$$\langle\langle W_t - \tilde{W}_{t+1}, \tilde{L}_t\rangle\rangle = \frac{\eta b_{t1}^2}{1 + \eta b_{t1}} \leq 2\eta b_{t1}^2 = 2\eta\frac{l_{ii}^2\lambda_i^2}{\mu_i^4}\,.$$

Similarly for $I \neq J$, (which occurs with probability $2\mu_i\mu_j$)

- $B_t = \frac{sl_{ij}\sqrt{\lambda_i\lambda_j}}{2\mu_i\mu_j}(u_i u_j^\top + u_j u_i^\top)$ has two nonzero eigenvalues;
- $b_{t\pm} = \pm\frac{sl_{ij}\sqrt{\lambda_i\lambda_j}}{2\mu_i\mu_j}$;

- Similarly from the previous calculation $|b_{t\pm}| \le \frac{1}{2\eta}$.

Finally we again get the bound of $\langle\langle W_t - \tilde{W}_{t+1}, \tilde{L}_t \rangle\rangle$ as follows:

$$\langle\langle W_t - \tilde{W}_{t+1}, \tilde{L}_t \rangle\rangle \le 2\eta b_{t+}^2 + 2\eta b_{t-}^2 \le 2\eta \frac{l_{ij}^2 \lambda_i \lambda_j}{\mu_i^2 \mu_j^2} .$$

Now we can calculate the total expectation of $\langle\langle W_t - \tilde{W}_{t+1}, \tilde{L}_t \rangle\rangle$.

$$\mathbb{E}[\langle\langle W_t - \tilde{W}_{t+1}, \tilde{L}_t \rangle\rangle] \le 2\eta \sum_{i,j} \frac{E_s[l_{ij}^2]\lambda_i \lambda_j}{\mu_i \mu_j} \le 8\eta \sum_{i,j} E_s[l_{ij}^2] = (8\eta \sum_{i,j} E_s[(w_t^\top L w_t)^2]) + 8\eta d^2 \sigma^2 .$$

Here $E_s$is the expectation from the random sign variable s, and the second inequality comes from $\mu_i > (1-\gamma)\lambda_i \ge \frac{1}{2}\lambda_i$.

Note that since the problem is now about stochastic bandit PCA problem, additional last noise term of $l_{ij} = w_t^\top L w_t + \epsilon_t$ leads an additional $8\eta d^2\sigma^2$ term at the bound of $\mathbb{E}[\langle\langle W_t - \tilde{W}_{t+1}, \tilde{L}_t \rangle\rangle]$. The steps of the Lemma 11 in Kotlowski & Neu (2019) leads $\sum_{i,j} E_s[(w_t^\top L w_t)]^2 \le d\|L\|_F^2 \le dr$, and applying this bound concludes the lemma. □

Combining Lemma F.4, Lemma F.3 and Lemma F.2 the regret is bounded as follows:

$$R_T \le \frac{d \log T}{\eta} + \gamma T + 8\eta(1-\gamma)T(dr + d^2\sigma^2) .$$

Substituting $\eta = \frac{1}{\sqrt{Td}}$ and $\gamma = Cd\eta$, we can get regret upper bound of order $O(\sqrt{d^3 T \log T})$.

F.2. In the case of the sub-Gaussian noise

For the convenience of the previous proofs, we assumed that ϵ_t is a bounded noise with variance σ^2. However, we can easily extend the result to the case of the σ sub-Gaussian noise. The bounded noise condition is used only when bounding $1 + \eta b_{ti}$. If $\eta \le \frac{1}{Cd}$ is satisfied, then the bounded noise condition was not necessary for the rest of the process (note that we set $\eta = \frac{1}{\sqrt{Td}}$ to get the regret upper bound of $\tilde{O}(\sqrt{d^3 T})$). Therefore, what we need to check is whether $\frac{1}{\sqrt{Td}} \le \frac{1}{Cd}$ holds even in the σ sub-Gaussian noise condition. For the sub-Gaussian maxima, the following inequality holds (Rigollet, 2015):

$$P(\max_t \epsilon_t > \sqrt{2\sigma^2(\log T + \log \frac{1}{\delta})}) \le \delta$$

Therefore, if we substitute $C \approx \sqrt{2\sigma^2(\log T + \log\frac{1}{\delta})}$, then most of our arguments hold with high probability $1-\delta$. In most cases $\log T \lll \sqrt{T}$, so the condition $\eta = \frac{1}{\sqrt{dT}} < \frac{1}{Cd}$ can be regarded as a reasonable assumption.

G. Discussion about the bilinear bandit lower bound

In this section, we will discuss the lower bounds of the bilinear bandit model. As mentioned in Jun et al. (2019), one of the simple case for the lower bound is when the arm set $\mathcal{X}$ or $\mathcal{Z}$ is a singleton. Then by using known linear bandit regret lower bound (Dani et al., 2008; Lattimore & Szepesvári, 2020), one can easily achieve the regret lower bound $\Omega(\max(d_1, d_2)\sqrt{T})$.

However, most of the algorithm assumes multiple entries for each side of the action set. Especially they require each $\mathcal{X}$ and $\mathcal{Z}$ spans the whole dimension d_1 and d_2, respectively; since if not, one can reduce the problem by projecting the action space to the lower dimension. For example, Jun et al. (2019) selects d_1(and d_2) independent actions at the start of the algorithm, and $\mathcal{X}$ and $\mathcal{Z}$ of the rank-1 bandits (Katariya et al., 2017; Trinh et al., 2020) can be seen as sets of the canonical vectors. In other words, the dimension of the action sets discussed in most of the bilinear algorithms is not only the nominal dimension, but the dimension of the $span(\mathcal{X})$ and $span(\mathcal{Z})$. From this point of view, lower bound using singleton example only represents $d_1 \times 1$ and $1 \times d_2$ cases.

When it comes to the non-singleton action sets, one has to deal with the bilinear nature of the problem, which introduces cross terms between the left and right arms. It makes the problem challenging to deal with in general (Jun et al., 2019). We are curious that the d order of regret lower bound also holds to all $d_1 \times d_2$ bilinear bandits, and the theorem below verifies that is true.

Theorem G.1. *Consider the arm sets* $\mathcal{X} := \{\pm 1/\sqrt{d_1}\}^{d_1}$,$\mathcal{Z} := \{\pm 1/\sqrt{d_2}\}^{d_2}$ *There exists* $\Theta \in \mathbb{R}^{d_1 \times d_2}$ *that satisfies the following regret lower bound:*

$$\mathbb{E}[R(\Theta)] \geq \Omega(\max(d_1, d_2)\sqrt{T}) \,.$$

Proof. WLOG assume $d_1 \geq d_2$. Let the hidden parameter hypothesis space $\Sigma = \{\theta\xi^\top : \theta \in \{\pm\sqrt{\epsilon/d_1}\}^{d_1}, \xi \in \{\pm\sqrt{\epsilon/d_2}\}^{d_2}\}$. The best expected reward $\max_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{z}\in\mathcal{Z}}(\mathbf{x}^\top\theta)(\xi^\top\mathbf{z})$ is achieved at $sign(\mathbf{x}) = sign(\theta)$ and $sign(\mathbf{z}) = sign(\xi)$ with the maximum of ϵ. Fix the bandit algorithm. Denote by $\mathbf{x}_t$ and $\mathbf{z}_t$ the arms pulled at time t. Then,

$$R_n(\mathcal{X}, \mathcal{Z}, \theta\xi^\top) = \sum_{t=1}^{T}[\epsilon - \mathbb{E}_{\theta,\xi}(\mathbf{x}_t^\top\theta) \cdot (\xi^\top\mathbf{z}_t)]$$

Here, $\mathbf{x}_t^\top\theta = \sum_i x_{ti}\theta_i = \sum_i(\frac{\sqrt{\epsilon}}{d_1}\mathbf{1}\{sign(x_{ti}) = sign(\theta_i)\} - \frac{\sqrt{\epsilon}}{d_1}\mathbf{1}\{sign(x_{ti}) \neq sign(\theta_i)\}) = \frac{\sqrt{\epsilon}}{d_1}\sum_i(1 - 2 \cdot \mathbf{1}\{sign(x_{ti}) \neq sign(\theta_i)\})$. Define $\#_t^T\{A_t\} = \sum_{t=1}^T \mathbf{1}\{A_t\}$, where A_t is an event. Using $\mathbf{1}\{sign(x_{ti}) \neq sign(\theta_i)\} = \mathbf{1}\{x_{ti}\theta_i < 0\}$,
$R_n(\mathcal{X}, \mathcal{Z}, \theta\theta^\top)$

$$\begin{aligned}
&= \sum_{t=1}^{T}\left[\epsilon - \frac{\epsilon}{d_1 d_2}\mathbb{E}_{\theta,\xi}\sum_{i=1}^{d_1}\sum_{j=1}^{d_2}(1 - 2\cdot\mathbf{1}\{x_{ti}\theta_i < 0\})\cdot(1 - 2\cdot\mathbf{1}\{z_{tj}\xi_j < 0\})\right] \\
&= \left[\sum_{t=1}^{T}\epsilon - \frac{\epsilon}{d_1 d_2}\mathbb{E}_{\theta,\xi}\sum_{i=1}^{d_1}\sum_{j=1}^{d_2}(1 - 2\cdot\mathbf{1}\{x_{ti}\theta_i z_{tj}\xi_j < 0\})\right] \\
&\overset{(a)}{=} 2\frac{\epsilon}{d_1 d_2}\sum_{t=1}^{T}\mathbb{E}_{\theta,\xi}\sum_{i=1}^{d_1}\sum_{j=1}^{d_2}\mathbf{1}\{x_{ti}\theta_i z_{tj}\xi_j < 0\} \\
&\geq 2\frac{\epsilon}{d_1 d_2}\sum_{i=1}^{d_1}\sum_{j=1}^{d_2}\mathbb{E}_{\theta,\xi}\;\left[\mathbf{1}\left\{\#_{t=1}^T\left\{x_{ti}\theta_i z_{tj}\xi_j < 0\right\} \geq \frac{T}{2}\right\}\cdot\frac{T}{2} + \mathbf{1}\left\{\#_t^T\left\{x_{ti}\theta_i z_{tj}\xi_j < 0\right\} < \frac{T}{2}\right\}\cdot\#_t^T\left\{x_{ti}\theta_i z_{tj}\xi_j < 0\right\}\right] \\
&\geq 2\frac{\epsilon}{d_1 d_2}\sum_i\sum_j\mathbb{P}_{\theta,\xi}\left(\#_{t=1}^T\left\{x_{ti}\theta_i z_{tj}\xi_j < 0\right\} \geq \frac{T}{2}\right)\cdot\frac{T}{2}
\end{aligned}$$

where (a) is by $d_1 d_2 = \sum_i\sum_j 1$.

We hope to enumerate all possible $\theta\xi^\top \in \Sigma$ and average out the regret:

$$\frac{1}{|\Sigma|}\sum_\theta R_n(\mathcal{X}, \mathcal{Z}, \theta\xi^\top) \geq C$$

for some C and then claim that there exists a $\theta\xi^\top$ such that the regret is greater than C. Define $p_{\theta\xi,ij} = \mathbb{P}_{\theta,\xi}\left(\#_{t=1}^T\left\{x_{ti}\theta_i z_{tj}\xi_j < 0\right\} \geq \frac{T}{2}\right)$. Let $\sum_{\theta_{-j}} := \sum_{\theta_{1:j-1}}\sum_{\theta_{j+1:d}}$, summation over all other coordinates except j-th coordinate. Define $\theta^{'(j)}$ to be the "flipped-j-th-coordinate' version of θ: $\theta_j^{'(j)} = -\theta_j$ and $\theta_k^{'(j)} = \theta_k, \forall k \neq j$. Then,

$$\begin{aligned}
\frac{1}{|\Sigma|}\sum_{\theta\in\Sigma}R_n(\mathcal{X}, \mathcal{Z}, \theta\theta^\top) &\geq \frac{T\epsilon}{d_1 d_2|\Sigma|}\sum_{\theta,\xi}\sum_i\sum_j p_{\theta\xi,ij} \\
&= \frac{T\epsilon}{d_1 d_2|\Sigma|}\sum_i\sum_j\sum_{\theta_{-i}}\sum_{\xi_{-j}}(p_{\theta\xi,ij} + p_{\theta'^{(i)}\xi,ij} + p_{\theta\xi'^{(j)},ij} + p_{\theta'^{(i)}\xi'^{(j)},ij})
\end{aligned}$$

Now, we realize that it all boils down the lower-bounding

$$p_{\theta\xi,ij} + p_{\theta'^{(i)}\xi,ij} + p_{\theta\xi'^{(j)},ij} + p_{\theta'^{(i)}\xi'^{(j)},ij} \;,$$

Without loss of generality, suppose that $d_1 > d_2$. Then, we can apply the Bretagnolle's inequality (Lattimore & Szepesvári (2020, Theorem 14.2)) for each pair $p_{\theta\xi,ij} + p_{\theta'^{(i)}\xi,ij}$ and $p_{\theta\xi'^{(j)},ij} + p_{\theta'^{(i)}\xi'^{(j)},ij}$

$$p_{\theta'^{(i)}\xi,ij} = \mathbb{P}_{\theta'^{(i)},\xi}\left(\#_{t=1}^T\left\{x_{ti}\theta_i^{'(i)}z_{tj}\xi_j < 0\right\} \geq \frac{T}{2}\right)$$

$$
\begin{aligned}
&= \mathbb{P}_{\theta'^{(i)}\xi}\left(\#_{t=1}^{T}\left\{x_{ti}\theta_i z_{tj}\xi_j > 0\right\} \geq \frac{T}{2}\right)\\
&= \mathbb{P}_{\theta'^{(i)}\xi}\left(\#_{t=1}^{T}\left\{x_{ti}\theta_i z_{tj}\xi_j < 0\right\} <\frac{T}{2}\right)
\end{aligned}
$$

Due to the Bretagnolle's inequality and the noise model $N(0,1)$,

$$
\begin{aligned}
p_{\theta\xi,ij} + p_{\theta'^{(i)}\xi,ij} &\geq \mathbb{P}_{\theta,\xi}\left(\#_{t=1}^{T}\left\{x_{ti}\theta_i z_{tj}\xi_j < 0\right\} \geq \frac{T}{2}\right) + \mathbb{P}_{\theta'^{(i)}\xi}\left(\#_{t=1}^{T}\left\{x_{ti}\theta_i z_{tj}\xi_j < 0\right\} < \frac{T}{2}\right)\\
&\geq \frac{1}{2}\exp\left(-\mathbb{E}_{\theta,\xi}\sum_{t=1}^{T}\frac{(\mathbf{x}_t^\top(\theta\xi^\top - \theta'^{(i)}\xi^\top)\mathbf{z}_t)^2}{2}\right)\\
&\geq \frac{1}{2}\exp\left(-2\mathbb{E}_{\theta,\xi}\sum_{t=1}^{T}(x_{ti}\theta_i\xi^\top\mathbf{z}_t)^2\right)\\
&\overset{(a)}{\geq} \frac{1}{2}\exp\left(-2T\frac{\epsilon^2}{d_1^2}\right)
\end{aligned}
$$

where in (a) we consider the worse case of $\mathbf{x}_t$ and $\mathbf{z}_t$. Now we see that

$$
p_{\theta\xi,ij} + p_{\theta'^{(i)}\xi,ij} \geq \frac{1}{2}\exp\left(-2T\frac{\epsilon^2}{d_1^2}\right)
$$

Similarly, one can calculate

$$
p_{\theta\xi^{(j)},ij} + p_{\theta'^{(i)}\xi^{(j)},ij} \geq \frac{1}{2}\exp\left(-2T\frac{\epsilon^2}{d_1^2}\right)
$$

Then,

$$
\begin{aligned}
\frac{1}{|\Sigma|}\sum_{\theta,\xi}R_n(\mathcal{X},\mathcal{Z},\theta\theta^\top) &\geq \frac{n\epsilon/d_1d_2}{|\Sigma|}\sum_i\sum_j\sum_{\theta_{-i}}\sum_{\xi_{-j}}(p_{\theta\xi,ij} + p_{\theta'^{(i)}\xi,ij} + p_{\theta\xi'^{(j)},ij} + p_{\theta'^{(i)}\xi'^{(j)},ij})\\
&\geq \frac{n\epsilon/d_1d_2}{|\Sigma|}\sum_i\sum_j 2^{d_1-1}2^{d_2-1}\frac{1}{2}\exp(-2T\frac{\epsilon^2}{d_1^2})\\
&\geq \frac{n\epsilon/d_1d_2}{2^{d_1+d_2}}d_1d_2 2^{d_1-1}2^{d_2-1}\exp(-2T\frac{\epsilon^2}{d_1^2})\\
&= \frac{n\epsilon}{4}\exp(-2T\frac{\epsilon^2}{d_1^2})\\
&\overset{(a)}{=} \frac{d_1\sqrt{n}}{4}\exp(-2)
\end{aligned}
$$

where (a) by choosing $\epsilon = \frac{d_1}{\sqrt{n}}$. This concludes the proof. □