Personalizing Many Decisions with High-Dimensional Covariates

Nima Hamidi[*] **Mohsen Bayati**[†] **Kapil Gupta**[‡]

Abstract

We consider the k-armed stochastic contextual bandit problem with d dimensional features, when both k and d can be large. To the best of our knowledge, all existing algorithms for this problem have a regret bound that scale as polynomials of degree at least two in k and d. The main contribution of this paper is to introduce and theoretically analyse a new algorithm (REAL-Bandit) with a regret that scales by $r^2(k+d)$ when r is the rank of the $k \times d$ matrix of unknown parameters. REAL-Bandit relies on ideas from low-rank matrix estimation literature and a new row-enhancement subroutine that yields sharper bounds for estimating each row of the parameter matrix that may be of independent interest. We also show via simulations that REAL-Bandit algorithm outperforms existing algorithms that do not leverage the low-rank structure of the problem.

1 Introduction

Running online experiments has recently become a popular approach in data-centric enterprises. However, running an experiment involves an opportunity cost or *regret* (e.g., exposing some users to a potentially inferior experience by chance). To reduce this opportunity cost, a growing number of companies leverage multi-armed bandit (MAB) experiments [38, 39, 19] that were initially motivated by the cost of experimentation in clinical trials [41, 27]. Another common feature of online experiments is personalization; users have heterogenous preferences that means the optimal decisions depend on users or product characteristics (also referred to by *context*). MAB approach for personalizing decisions is therefore called contextual MAB (or contextual bandit) [29]. For example, [30] uses contextual bandits to propose a personalized news article recommender system.

There is a large body of literature on algorithms with theoretical guarantees for contextual bandits with linear reward functions. An admittedly incomplete list is [5, 13, 2, 12, 14, 4, 34, 37, 42, 25, 6], and we defer to [7] for additional references. While these papers study the problem under a variety of different assumptions, they can be divided into two groups: (A) when context vectors are arbitrary and can be potentially selected by an adversary, and (B) when context vectors are i.i.d. samples from a fixed (but unknown) probability distribution. Our focus in this paper is the latter group (first studied by [14]). As the number of decisions T (also called time horizon) grows, the regret bounds for algorithms in group (A) grow with $\sqrt{T}$. But the algorithms in group (B) take advantage of the i.i.d. assumption and have a significantly lower (logarithmic) dependence in T.

Two other important parameters are the number of arms k, and the dimension of context vectors d. For example, when d grows, the regret bound of [14] grows as d^3 which can dominate the dependence on T. [6] tackles this difficulty by imposing sparseness assumption and can replace d^3 with s^2 (up to logarithmic factors) where s is sparsity of the reward functions. On the other hand, a careful inspection of the bounds in [14, 6] reveals that their regret bounds could grow by k^3 in the worst case that can be very large in applications such as assortment optimization [21].

[*]Department of Statistics, Stanford University, hamidi@stanford.edu
[†]Graduate School of Business, Stanford University, bayati@stanford.edu
[‡]Airbnb, kapil.gupta@airbnb.com

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

On the other hand, the lower bounds found in [13] and [14] imply that if there is no structural assumption, the lowest possible regret would grow at least by kd which can be very large when k and d are large. The aim of this paper is to reduce this dependence under a low-rank assumption on the $k \times d$ matrix of parameters of the reward functions. Specifically, each of the k reward functions is represented by a d-vector of coefficients (one for each covariate) that is a row of the parameter matrix. This can also be interpreted as imposing a similarity between the reward function of the k different arms, like in multi-task learning literature [10]. We propose a new algorithm (called REAL-Bandit) and prove that its regret grows by $r^2(k + d)$ where r is rank of the parameter matrix.

Contributions. Our main technical contributions in design and analysis of REAL-Bandit are as follows. *(1) Stronger row-wise guarantees:* To prove regret guarantees for REAL-Bandit, we need bounds for estimating every single row of the matrix. However, existing matrix estimation results provide a bound on the estimation error of the whole matrix, which would be a crude upper bound for the estimation error of each single row. Therefore, REAL-Bandit includes a subroutine (called Row Enhancement) that refines the estimates in order to establish stronger row-wise guarantees that may be of independent interest (see §3 for details). Very recently, [11] provided bounds for the matrix completion problem that are also sharp at the row or entry level, however, their results are only for the matrix completion case and do not apply to our setting. *(2) Implementation:* Our theoretical analysis does not require that each matrix estimation phase in REAL-Bandit is solved to completion. In other words, REAL-Bandit does not need finding a global minima of the relevant estimator's optimization (penalized maximum likelihood) problem. REAL-Bandit only needs a solution with cost below a certain threshold which can be used to significantly speed up implementation of REAL-Bandit. *(3) Estimator independence:* Over the last decade, several types of estimators have been introduced for recovering low-rank matrices from noisy observations, with varying assumptions and theoretical guarantees. Two of the most common approaches are based on convex optimization [8, 9, 31, 15, 35, 36, 26, 24], or non-convex optimization [40, 22, 23]. Unlike [14, 6] that work with a fixed estimator, REAL-Bandit is designed to be estimator agnostic and works with any matrix estimation algorithm with theoretical guarantees (see §3 for details).

Other related literature. A class of decision-making problems with a large number of arms are assortment optimization problems; when a small subset from a potentially large set of items should be selected. Among the rich literature on this topic, [21, 3] are more related to our paper since they consider a dynamic allocation of assortments via multi-armed bandit ideas. [21] (like us) uses low-rank matrix estimation methods for the learning part. However, the problems and models they study are very different. Specifically, they assume that a decision-maker shows a subset of products to a user. Then the user selection is modeled via a multi-nomial logit (MNL) [32] where the parameters of the MNL model form a low-rank matrix with rows representing customer types and columns representing products.

Two other relevant papers are [42, 25] since they too tackle bandit problems with many actions. They introduce algorithms with regrets that scale with spectral dimension of the Laplacian of a graph that has arms as its vertices. These papers are in group (A) of the aforementioned class of bandit papers that are inherently different. Specifically, they assume actions have known feature vectors (with low spectral dimension) that, together with a single unknown parameter vector, define the linear reward functions. There is a reduction from this setting to our problem, only when action set is allowed to change (see [1]) which is not the case in [42, 25]. Another recent paper in this category is [28]. The main difference between all of these papers and ours, as discussed above, is that we consider i.i.d. context vectors which allows regret bounds that scale with $\log T$ instead of $\sqrt{T}$. Finally, recent paper of [20] studies a bandit problem where each action is a pair of arms and the reward function is a bilinear function of feature vectors of each arm, and has a low-rank parameter matrix.

Organization. We introduce additional notation in §2. Then the REAL-estimator and REAL-Bandit algorithm are introduced in §3, followed by simulations in §4. In §5 we present our assumptions, statement of the main theorem, as well as its proof. Proofs of lemmas and additional details are deferred to the extended version of the paper [17].

2 Setting and notation

Let $\mathbf{B}^\star$ be a $k \times d$ matrix with real-valued entries. We further assume that $\mathbf{B}^\star$ is of rank r with $r \ll \min(k, d)$. At time $t = 1, 2, \cdots$, a *context vector* $X_t \in \mathbb{R}^d$ is drawn from the given probability distribution $\mathcal{P}_X(\cdot)$, independently from X_s for $s < t$. Then, by choosing arm $1 \le \kappa \le k$, the reward $y_t := \langle B_\kappa, X_t \rangle + \varepsilon_{\kappa,t}$ is generated, where $B_\kappa^{\star\top}$ is the κ-th row of the matrix $\mathbf{B}^\star$, and $\varepsilon_{\kappa,t}$ are independent σ_ε^2-sub-Gaussian random variables. In addition, $\langle U, V \rangle$ refers to the inner product of vectors U and V, and $U^\top$ refers to the transpose of U. We also use notation $[n]$ for the set $\{1, 2, \ldots, n\}$, when n is an integer. For any two matrices $\mathbf{Y}_1$ and $\mathbf{Y}_2$ with d columns, by $\mathbf{Y}_1 \sqsubseteq \mathbf{Y}_2$, we mean that all rows of $\mathbf{Y}_1$ are also rows of $\mathbf{Y}_2$.

By policy π, we mean a sequential decision-making algorithm that, at each time t, chooses the arm $\pi_t \in \{1, 2, \cdots, k\}$ given previous observations and the revealed context X_t. We will evaluate the performance of a policy by its cumulative regret, defined as $R_T = \sum_{t=1}^T r_t$, where $r_t = \mathbb{E}\big[\max_{\kappa \in [k]} \langle X_t, B_\kappa^\star \rangle - \big\langle X_t, B_{\pi_t}^\star \big\rangle\big]$. Our goal is to find policies with low R_T.

In order to avoid dealing with unnecessary subscripts, for each context vector $X_t \in \mathbb{R}^d$, we define $\mathbf{X}_t^\pi$ to be a $k \times d$ matrix with all elements equal to zero, except for the π_t-th row which is defined to be equal to $X_t^\top$. Using this notation, we have that

$$y_t = \langle \mathbf{B}^\star, \mathbf{X}_t^\pi \rangle + \varepsilon_t, \tag{1}$$

where the inner product for matrices is defined as $\langle \mathbf{U}, \mathbf{V} \rangle := \operatorname{tr}\big(\mathbf{U}\mathbf{V}^\top\big)$. Note that ε_t is actually $\varepsilon_{\pi_t,t}$, but since all noise values are i.i.d., we will drop the dependence on π.

For a given subset $\mathcal{I} = \{t_1, \ldots, t_n\}$ of $[T]$, consider the set of corresponding context matrices $\{\mathbf{X}_{t_i}^\pi \mid i \in \mathcal{I}\}$. We define an associated *sampling* operator $\mathfrak{X}_\mathcal{I}^\pi : \mathbb{R}^{k \times d} \to \mathbb{R}^n$ to be defined as follows. For any matrix $\mathbf{B} \in \mathbb{R}^{k \times d}$, $\mathfrak{X}_\mathcal{I}^\pi(\mathbf{B})$ is a vector of length n where its ith entry ($i \in [n]$) is given by $[\mathfrak{X}_\mathcal{I}^\pi(\mathbf{B})]_i := \langle \mathbf{B}, \mathbf{X}_{t_i}^\pi \rangle$. Therefore, the vector form of (1) is

$$Y = \mathfrak{X}_\mathcal{I}^\pi(\mathbf{B}^\star) + E\,.$$

We also opt to use the simpler notation $\mathfrak{X}(\cdot)$ instead of $\mathfrak{X}_\mathcal{I}^\pi(\cdot)$ when $\mathcal{I}$ and π are implicitly clear.

We also use different norms in our algorithm and analysis in this paper. We use $\|\cdot\|_2$ to denote the regular ℓ^2 norm of a vector. The nuclear norm (or trace-norm) of a matrix is denoted by $\|\cdot\|_*$, and $\|\cdot\|_F$ and $\|\cdot\|_\infty$ refer to the Frobenius and infinity norm of a matrix. Also, for a given distribution $\mathcal{P}$ over $\mathbb{R}^{k \times d}$, we can define the following norm $\|\mathbf{B}\|_\mathcal{P} := \mathbb{E}\Big[\langle \mathbf{B}, \mathbf{Z} \rangle^2\Big]$ for all $\mathbf{B} \in \mathbb{R}^{k \times d}$ where $\mathbf{Z}$ is drawn from $\mathcal{P}$. Finally, $\|\mathbf{\Gamma}\|_{\infty,2}$ is the maximum of $\|\Gamma_\kappa\|_2$ for $\kappa \in [k]$ (recall that $\Gamma_\kappa^\top$ is row κ of $\mathbf{\Gamma}$). In fact, one of our assumptions that will be stated explicitly later is that the matrix $\mathbf{B}^\star$ belongs to the following set:

$$\mathcal{S} = \{\mathbf{B} \in \mathbb{R}^{k \times d} \mid \|\mathbf{B}\|_{\infty,2} \le b^\star\}\,,$$

for a positive constat $b^\star$. Also, for a k by d matrix $\mathbf{B}$ of rank r with singular value decomposition $\mathbf{B} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$, we define the *row-incoherence* parameter as

$$\mu(\mathbf{B}) = \sqrt{\frac{k}{r}} \cdot \frac{\|\mathbf{B}\|_{\infty,2}}{\mathbf{D}_{r,r}}.$$

3 Algorithm

In this section, we describe the Row-Enhanced and Low-Rank Bandit (REAL-Bandit) algorithm. The algorithm combines ideas from existing literature [14, 6] and a new *row-enhancement* procedure to obtain sharper convergence rate since k can be very large. REAL-Bandit algorithm has two disjoint phases for exploration and exploitation, similar to [14, 6]. In the exploration phase, all arms are given an equal chance to be explored to enable the algorithm to obtain an estimate of their corresponding vectors. These *forced-sampling* estimates are *not* sufficiently accurate to pick the best arm with high probability, however, they are accurate enough to rule out all the arms that are substantially inferior to the optimal arm. At each time t, these estimates are used as a proxy of the actual arm vectors to form a set of *candidate* arms. In order to choose one of these candidates, we need more accurate estimates and so, the algorithm uses the *all-sampling* estimates that are obtained from *all* the observations made thus far to pick the best arm.

However, unlike [14, 6], that estimate each of the arm parameters $\{B_\kappa^\star\}_{\kappa \in [k]}$ separately, our forced-sampling and all-sampling estimators utilize the low-rank assumption on matrix $\mathbf{B}^\star$ to estimate all parameters simultaneously (like in multi-task learning literature).

The Estimators. REAL-Bandit is designed to work with any matrix estimation method that has theoretical guarantees. Two such estimators (developed in the matrix completion literature) are: (1) estimators based on convex optimization and (2) estimators based on non-convex optimization. Before we present these two classes of estimators, we assume that a set of time periods $\mathcal{J} = \{t_1, \ldots, t_n\}$ and their associate observations $(\mathbf{X}^{\pi}_{t_1}, y_{t_1}), \ldots, (\mathbf{X}^{\pi}_{t_n}, y_{t_n})$, and a positive constant λ are available. We use notations $\bar{\mathbf{B}}(\mathcal{J}, \lambda)$ or $\widehat{\mathbf{B}}(\mathcal{J}, \lambda)$ for estimators of $\mathbf{B}^\star$, that use observations from time periods in $\mathcal{J}$ and regularization parameter λ. When $\mathcal{J}$ and λ are clear, we use simpler notations $\bar{\mathbf{B}}$ and $\widehat{\mathbf{B}}$.

(1) Convex optimization. In this approach, introduced by [8], the approximation to $\mathbf{B}^\star$ is the minimizer of the following convex program:

$$\text{minimize} \quad n^{-1}\|Y - \mathfrak{X}(\mathbf{B})\|^2 + \lambda\|\mathbf{B}\|_* \,.$$

In fact, as [16] shows, one just needs a feasible solution $\bar{\mathbf{B}} = \bar{\mathbf{B}}(\mathcal{J}, \lambda)$ that satisfies:

$$n^{-1}\left\|Y - \mathfrak{X}(\bar{\mathbf{B}})\right\|^2 + \lambda\left\|\bar{\mathbf{B}}\right\|_* \le n^{-1}\|Y - \mathfrak{X}(\mathbf{B}^\star)\|^2 + \lambda\|\mathbf{B}^\star\|_*. \tag{2}$$

This brings additional flexibility to choose the optimizer and has computational advantages.

(2) Non-convex optimization. Another approach is to explicitly impose the low-rank constraint by writing $\mathbf{B}$ as $\mathbf{U}\mathbf{V}^\top$ where $\mathbf{U} \in \mathbb{R}^{k \times r}$ and $\mathbf{V} \in \mathbb{R}^{d \times r}$. The optimization problem would be:

$$\text{minimize}_{\mathbf{U},\mathbf{V}} \quad n^{-1}\left\|Y - \mathfrak{X}(\mathbf{U}\mathbf{V}^\top)\right\|^2 + \lambda(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2)/2$$

One challenge is that this is not a convex program, but it has been shown that under certain conditions, *alternating minimization* can be an effective algorithm [18]. It can also be shown (e.g., see [22, 31]) that minimizing the above loss function is equivalent to solving the following optimization problem:

$$\begin{aligned} \text{minimize} \quad & n^{-1}\|Y - \mathfrak{X}(\mathbf{B})\|^2 + \lambda\|\mathbf{B}\|_* \\ \text{subject to} \quad & \text{rank}(\mathbf{B}) \le r. \end{aligned}$$

If $\bar{\mathbf{B}}$ is one such solution, since $\mathbf{B}^\star$ is also a feasible solution, (2) must hold.

REAL-estimator. The existing theory of matrix estimation provides error bounds for $\left\|\bar{\mathbf{B}} - \mathbf{B}^\star\right\|_F$ [9, 26, 33, 24, 16]. However, these results do not characterize how this error is distributed across different rows. On the other hand, in order to get a regret bound, we need to control $\left\|\bar{B}_\kappa - B^\star_\kappa\right\|_2$ for all $\kappa \in [k]$, and the trivial inequality $\left\|\bar{B}_\kappa - B^\star_\kappa\right\|_2 \le \left\|\bar{\mathbf{B}} - \mathbf{B}^\star\right\|_F$ would introduce an unnecessary $\sqrt{k}$ factor. Here, we introduce REAL-estimator that uses a set of *almost* independent observations to improve the row-wise error bound.

As before, let $\mathcal{J} = \{t_1, \ldots, t_n\}$ be a set of n time periods with $t_1 < t_2 < \cdots < t_n$. We split these to two halves $\mathcal{J}_1 := \{t_1, \ldots, t_{\frac{n}{2}}\}$ and $\mathcal{J}_2 := \{t_{\frac{n}{2}+1}, \ldots, t_n\}$. For any $\mathcal{K} \subseteq [k]$, let $\mathcal{J}^{\mathcal{K}}$ be the subset of $\mathcal{J}$ such that an arm in $\mathcal{K}$ is pulled, i.e. $\pi_{t_i} \in \mathcal{K}$. Moreover, for $\ell \in \{1, 2\}$ and $\mathcal{K} \subseteq [k]$, let $\mathcal{J}^{\mathcal{K}}_\ell := \mathcal{J}^{\mathcal{K}} \bigcap \mathcal{J}_\ell$. For $\kappa \in [k]$, when $\mathcal{K} = \{\kappa\}$, we use superscript κ rather than $\{\kappa\}$ for simplicity. Next, for *any* low-rank matrix estimator $\bar{\mathbf{B}}$, Algorithm 1 performs the row-enhancement procedure. We call the output of this algorithm REAL-estimator and denote it by $\widehat{\mathbf{B}}(\mathcal{J})$. The difficulty of

Algorithm 1 Row-enhancement procedure

Input: Low-rank matrix estimator $\bar{\mathbf{B}} \in \mathbb{R}^{k \times d}$, $\mathcal{J} = \{t_1, \ldots, t_n\}$, and observations $(\mathbf{X}^{\pi}_{t_1}, y_{t_1}), \ldots, (\mathbf{X}^{\pi}_{t_n}, y_{t_n})$.

Initialize, $\widehat{\mathbf{B}} \in \mathbb{R}^{k \times d}$,
Split $\mathcal{J}$ into $\mathcal{J}_1 := \{t_1, \ldots, t_{\frac{n}{2}}\}$ and $\mathcal{J}_2 := \{t_{\frac{n}{2}+1}, \ldots, t_n\}$,
Compute SVD $\bar{\mathbf{B}}(\mathcal{J}_1) = \mathbf{U}\mathbf{D}\mathbf{V}^\top$,
Let $\mathbf{V}_r$ be the matrix containing first r columns of $\mathbf{V}$,
for $\kappa = 1, 2, \cdots, k$ **do**
    Let $\hat{\vartheta}_\kappa = \arg\min_{\vartheta \in \mathbb{R}^r} \sum_{t_i \in \mathcal{J}^\kappa_2} (y_{t_i} - \langle \mathbf{V}_r\vartheta, X_{t_i}\rangle)^2$,
    Set row κ of $\widehat{\mathbf{B}}$ to $(\mathbf{V}_r\hat{\vartheta}_\kappa)^\top$.
end for
Return $\widehat{\mathbf{B}}$.

analyzing this estimator arises from the fact that the observations are generated in an adaptive fashion, and thus, the results that require independence assumption are not applicable in our case. However, in the analysis we will show that these observations can be *approximated* by i.i.d. samples, and as a result theoretical guarantees can be obtained. In the following, we will state the assumptions formally, and then, we will verify that they hold in the course of the analysis.

Before we define the notion of approximately independent, we need a few more notations to make the definition cleaner. For $\kappa \in [k]$, we denote by $\mathbf{X}^\kappa$ the set of context vectors of observations for arm κ stacked horizontally. We define $\mathbf{X}_\ell^\kappa$ for $\ell \in \{1, 2\}$ for $\mathcal{J}_\ell$ similarly. Recall that, for matrices $\mathbf{Y}_1$ and $\mathbf{Y}_2$, by $\mathbf{Y}_1 \sqsubseteq \mathbf{Y}_2$ with d columns, we mean that all rows of $\mathbf{Y}_1$ are also rows of $\mathbf{Y}_2$.

Definition 3.1 (Approximately independence)**.** Let $\mathcal{J}$ be a given set of n time periods and $\mathcal{P}$, $\mathcal{P}_\mathcal{U}$, and $\mathcal{P}_\mathcal{V}$ be three distributions. Then, for $\kappa \in [k]$, we say that $\mathcal{J}^\kappa$ is a $(n_\mathcal{U}, n_\mathcal{V})$*-approximately independent* set of observations if there exists random matrices $\mathbf{X}_\mathcal{U}^\kappa$ and $\mathbf{X}_\mathcal{V}^\kappa$ such that

1. $\mathbf{X}_\mathcal{U}^\kappa \sqsubseteq \mathbf{X}^\kappa$ and $\mathbf{X}^\kappa \sqsubseteq \mathbf{X}_\mathcal{V}^\kappa$,
2. All rows of $\mathbf{X}_\mathcal{U}^\kappa$ are independent samples of $\mathcal{P}_\mathcal{U}$,
3. All rows of $\mathbf{X}_\mathcal{V}^\kappa$ are independent samples from either $\mathcal{P}$ or $\mathcal{P}_\mathcal{V}$,
4. $\mathbf{X}_\mathcal{U}^\kappa$ and $\mathbf{X}_\mathcal{V}^\kappa$ have $n_\mathcal{U}$ and $n_\mathcal{V}$ rows respectively.

This definition requires the observations for a row κ to lie between two sets of i.i.d. samples. This notion becomes extremely useful whenever one can prove that $n_\mathcal{U}$ and $n_\mathcal{V}$ are of the same order.

Next, we specify the conditions that $\mathcal{P}$, $\mathcal{P}_\mathcal{U}$, and $\mathcal{P}_\mathcal{V}$ need to meet so that we can prove error bounds.

Definition 3.2. We say that a distribution $\mathcal{P}(\cdot)$ on $\mathbb{R}^d$ is $(\gamma_{\min}, \gamma_{\max}, \sigma_X)$-diverse if

$$\gamma_{\min} \leq \lambda_{\min}(\mathbf{\Sigma}) \leq \lambda_{\max}(\mathbf{\Sigma}) \leq \gamma_{\max},$$

where $\mathbf{\Sigma} = \mathbb{E}\left[XX^\top\right]$, and $\mathbf{\Sigma}^{-\frac{1}{2}}X$ is σ_X^2-sub-Gaussian (i.e., for any deterministic unit vector $u \in \mathbb{R}^d$, the real-valued random variable $u^\top \mathbf{\Sigma}^{-\frac{1}{2}} X$ is σ_X^2-sub-Gaussian).

We will treat σ_X as a constant. Note that, for instance, when X follows a multivariate Gaussian distribution, then $\sigma_X = 1$.

In our proofs, we will show that whenever $\mathcal{J}$ can be split into two *almost* independent halves, then row-enhancement procedure gives us sharper per-row guarantees than the raw matrix estimator $\bar{\mathbf{B}}$.

The REAL-Bandit algorithm. In this section, we describe REAL-Bandit algorithm presented in Algorithm 2. As mentioned earlier, this algorithm has disjoint exploration and exploitation phases which are specified by a *force-sampling rule* $f : \mathbb{N} \to [k] \cup \{\varnothing\}$. At time t, the force-sampling rule decides between *forcing* the arm $f_t \in [k]$ to be pulled or *exploiting* the past data by $f_t = \varnothing$. By $\mathcal{F}_t$, we denote the times that an arm was forced to be pulled, i.e. $\mathcal{F}_t := \{\tau \leq t : f_\tau \in [k]\}$. For simplicity, we also use $\mathcal{A}_t := [t]$ to refer to the all time periods up to time t. The force-sampling rule that we use in this algorithm is a randomized function that picks an arm $\kappa \in [k]$ with probability

$$\mathbb{P}(f_t = \kappa) = \begin{cases} \frac{1}{k} & \text{if } t \leq 2\rho\log(\rho)\,, \\ \frac{\rho}{k(t-\rho\log(\rho)+1)} & \text{if } t > 2\rho\log(\rho)\,, \end{cases} \tag{3}$$

and $f_t = \varnothing$ otherwise. We will specify the hyper-parameter ρ in §5. As we will see in the analysis of the algorithm, this force-sampling rule ensures that $\mathcal{F}_t^\kappa$ grows as $\mathcal{O}(\log t)$ for all $\kappa \in [k]$. One can alternatively use any force-sampling rule that has this rate of exploration.

Remark 1. *The algorithm proposed in [14, 6] are similar in nature to the REAL-Bandit. They, however, use a deterministic force-sampling rule (that can be used here as well). However, our randomized rule brings practical advantages in exchange for a slightly more complex theoretical analysis.*

Now, let $\bar{\mathbf{B}}^F$ and $\bar{\mathbf{B}}^A$ be two low-rank matrix estimators and denote by $\widehat{\mathbf{B}}^F$ and $\widehat{\mathbf{B}}^A$ their corresponding REAL-estimators, introduced above. These estimators serve different purposes in our algorithm. As we will show in the analysis of the algorithm, the forced-samples estimator $\widehat{\mathbf{B}}^F$ satisfies $\left\|\widehat{B}_\kappa^F - B_\kappa^\star\right\|_2 \leq \mathcal{O}(1)$ with probability at least $1 - \mathcal{O}(1/t)$ for *all* arms $\kappa \in [k]$. The key idea is that $\mathcal{O}(\log t)$ i.i.d. samples are enough to get such a guarantee. These estimates are then only used to rule out some arms that are *very* far from the optimal arm. The threshold for eliminating sub-optimal arms is determined by a hyper-parameter h that is given to the algorithm. This parameter can be thought of as the *gap* of the problem.

The remaining arms are *candidates* of being the optimal arm. Then, the all-samples estimator $\widehat{\mathbf{B}}^A$ comes into play. This estimator is used to pick the *best* arm among these candidate arms. We will show that $\widehat{\mathbf{B}}^A$ enjoys the sharper bound $\left\|\widehat{B}_\kappa^A - B_\kappa^\star\right\|_2 \leq \mathcal{O}\big(1/\sqrt{t}\big)$ for *all optimal* arms $\kappa \in \mathcal{K}_{\text{opt}}$ with probability at least $1 - \mathcal{O}(1/t)$. This sharper rate improves the accuracy of the decisions made by the algorithm significantly.

Finally, note that the theory on low-rank matrix estimation requires positive probability of observing all rows the matrix $\mathbf{B}^\star$ to work. This assumption holds for the forced-samples, as we explore arms with the same rate. Nevertheless, this definitely fails to hold for the rest of the samples. In fact, only *optimal* arms (i.e. the arms that are optimal with positive probability) receive $\mathcal{O}(t)$ observations. This results in the breakdown of the guarantees for $\bar{\mathbf{B}}^A$ and thus, for $\bar{\mathbf{B}}^A$. We address this issue by detecting the set of optimal arms as the time goes on. We pass the set $\widehat{\mathcal{K}}_{\text{opt}}$ to the all-samples estimator to indicate the set of optimal arms.

Algorithm 2 REAL-Bandit algorithm

Input: Force-sampling rule f, gap h.

Initialize the vector $last \in \mathbb{R}^k$ by $last[\kappa] \leftarrow 0$ for all $\kappa \in [k]$
for $t = 1, 2, \cdots$ **do**
  Observe $X_t \sim \mathcal{P}_X$,
  if $f_t \neq 0$ **then**
    $\pi_t \leftarrow f_t$
  else
    $\mathcal{C} = \left\{\kappa \in [k] \mid \left\langle X_t, \widehat{B}_\kappa^F(\mathcal{F}_{t-1})\right\rangle \geq \max_{\ell \in [k]} \left\langle X_t, \widehat{B}_\ell^F(\mathcal{F}_{t-1})\right\rangle - \frac{h}{2} \cdot \|X_t\|_2\right\}$
    for $\kappa \in \mathcal{C}$ **do**
      $last[\kappa] \leftarrow t$
    end for
    $\widehat{\mathcal{K}}_{\text{opt}} \leftarrow \{\kappa \in [k] \mid last[\kappa] \geq \frac{t}{2}\}$
    $\pi_t \leftarrow \arg\max_{\kappa \in \mathcal{C}} \left\langle X_t, \widehat{B}_\kappa^A(\mathcal{A}_{t-1}, \widehat{\mathcal{K}}_{\text{opt}})\right\rangle$
  end if
end for

4 Simulations

We compared REAL-Bandit with four other algorithms: OLS-Bandit of [14], but we use the improved version from [6] that filters sub-optimal arms, LASSO-Bandit of [6], OFUL of [2] which is based on the Upper Confidence Bound (UCB) ideas, and Thompson sampling (the version from [37]). Taking $k = 201, d = 200$, and $r = 3$, we generated matrix $\mathbf{B}^\star$ as $\mathbf{U}\mathbf{V}^\top$ where rows of $\mathbf{U} \in \mathbb{R}^{201\times 3}$ and $\mathbf{V} \in \mathbb{R}^{200\times 3}$ are drawn independently and uniformly from the unit sphere in $\mathbb{R}^3$. Noise variance is 1 and features are i.i.d. $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. We gave Thompson sampling the true prior mean and variance of arm parameters, and true noise variance. Similarly, OFUL had access to the true noise variance. Other parameters of OLS-Bandit, LASSO-Bandit, and OFUL are selected as in [6]. We generated 10 data sets, ran all algorithms for a time horizon of length $T = 40,000$. Figure 1 shows average cumulative regret (with 1 SE error bars) for all algorithms. The results of this simulation back our theoretical analysis, that REAL-Bandit takes advantage of the low-rank structure of the problem parameters and significantly outperforms the state-of-the-art methods that do not leverage the structure.

5 Analysis

This section is dedicated to the analysis of REAL-Bandit. We will first state the assumptions underlying the analysis and then state the main theorem of this section. A discussion of the following assumptions can be found in [6].

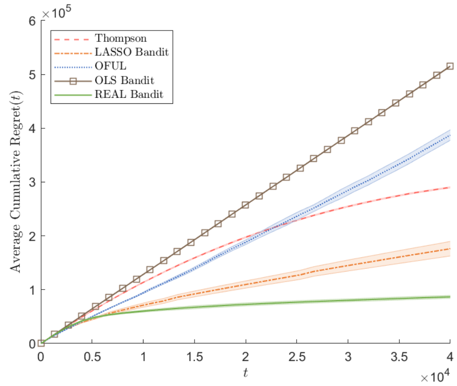

Figure 1: Cumulative regret of REAL-Bandit versus LASSO-Bandit, OFUL, OLS-Bandit, and Thompson sampling for $(k, d, r) = (201, 200, 3)$.

Assumption 1 (Parameter set)**.** Assume the rank of $\mathbf{B}^\star$ is r, $\|\mathbf{B}^\star\|_{\infty,2} \leq b^\star$, and $\mu^\star := \mu(\mathbf{B}^\star)$.

Assumption 2 (Margin condition)**.** For any $a > 0$, there is a constant $c_0 > 0$ such that $\mathbb{E}[N_a] \leq kc_0 a$, where the random variable N_a is defined by $N_a := \sum_{\kappa=1}^{k} \mathbb{I}\left(\langle X, B^\star_{\kappa^*} - B^\star_\kappa\rangle \leq a \cdot b^\star \cdot \|X\|_2\right)$.

Assumption 3 (Arm optimality)**.** Let $\mathcal{K}_{opt}$ and $\mathcal{K}_{sub}$ be a partitioning of $[k]$. Then, for some $h > 0$, the following conditions hold:
1) For any sub-optimal arm $\kappa \in \mathcal{K}_{sub}$, $\langle X, B^\star_\kappa\rangle \leq \max_{\kappa'}\langle X, B^\star_{\kappa'}\rangle - h \cdot \|X\|_2$ for any context X,
2) For each arm $\kappa \in \mathcal{K}_{opt}$, assume that $\mathbb{P}(X \in \mathcal{U}_k) \geq \frac{p^*}{|\mathcal{K}_{opt}|}$, where $\mathcal{U}_\kappa$ is defined by

$$\mathcal{U}_\kappa := \left\{X \in \mathbb{R}^d \mid \langle X, B^\star_\kappa\rangle > \max_{\kappa' \neq \kappa}\langle X, B^\star_{\kappa'}\rangle + h \cdot \|X\|_2\right\},$$

3) For each arm $\kappa \in \mathcal{K}_{opt}$, there exists constant $q^* > 0$ such that $\max_{\kappa \in \mathcal{K}_{opt}} \mathbb{P}(X \in \mathcal{V}_\kappa) \leq \frac{q^*}{|\mathcal{K}_{opt}|}$ where the set $\mathcal{V}_\kappa$ is defined as

$$\mathcal{V}_\kappa := \left\{X \in \mathbb{R}^d \mid \langle X, B^\star_\kappa\rangle > \max_{\kappa' \neq \kappa}\langle X, B^\star_{\kappa'}\rangle - h \cdot \|X\|_2\right\}.$$

Assumption 4 (Diversity)**.** For all $\kappa \in \mathcal{K}_{opt}$, $\mathcal{P}$, $\mathcal{P}_{\mathcal{U}_\kappa}$, and $\mathcal{P}_{\mathcal{V}_\kappa}$ are $(\gamma_{\min}, \gamma_{\max}, \sigma_X)$-diverse.

Assumption 5 (Low-rank estimators)**.** Assume the following tail bounds hold for $\bar{\mathbf{B}}^F$ and $\bar{\mathbf{B}}^A$:
1) Let $\mathcal{J}$ be a set of n time periods such that for each arm $\kappa \in [k]$, the matrix $\mathbf{X}^\kappa$ of the context vectors associated to the arm κ has i.i.d. rows sampled from $\mathcal{P}$. Then, if $n \geq c_2(k+d)\log(kdt)$ and $n \geq c_3 \frac{\sigma_\varepsilon^2 \log(kdt)}{\delta^2}$, we have

$$\mathbb{P}\left(\left\|\bar{\mathbf{B}}^F(\mathcal{J}) - \mathbf{B}^\star\right\|_F \geq \delta\right) \leq \exp\left(-\frac{c_1\delta^2 n}{(k+d)r}\right),$$

where δ is defined to be $\sqrt{\frac{\gamma_{\min}}{\gamma_{\max}}} \cdot \sqrt{\frac{k}{r}} \cdot \frac{h}{64\mu^\star}$, and c_1, c_2, and c_3 are positive constants.

2) Let $\mathcal{J}$ be a set of n observations, such that for all $\kappa \in \mathcal{K}_{\text{opt}}$, $\mathcal{J}^\kappa$ is a set of $(\frac{np^*}{2}, 2nq^*)$-approximately independent observations. Then, we get

$$\mathbb{P}\left(\left\|\bar{\mathbf{B}}^A(\mathcal{J}) - \mathbf{B}^\star\right\|_F \geq \left(\frac{\sqrt{d}\sigma_\varepsilon \vee b^\star}{\sqrt{d}\gamma_{\min}}\right) \cdot \sqrt{\frac{c_5(k+d)r\log(n)}{n}}\right) \leq \frac{1}{n},$$

provided that $n \geq c_4(k+d)\log(kdt)$ for some constants $c_4, c_5 > 0$.

Now, we are prepared to state our main theoretical result.

Theorem 1. *If Assumptions 1-5 hold, then the cumulative regret of Algorithm 2 is bounded above by*

$$\frac{R_T}{b^\star \cdot x_{\mathrm{n}}} \leq C\left\{C_1 \frac{r\log(kdT)^2}{kh^2} + C_2 r(k+d)\log(kdT)^2\right\}$$

where $C > 0$ is a numerical constant and

$$\rho := C_0(k+d)\,, \quad C_0 := \max\left\{\frac{1}{p^*}, c_1, \frac{\sigma_\varepsilon^2}{h^2\gamma_{\min}}\cdot\frac{\gamma_{\max}}{\gamma_{\min}}\right\}, \quad C_1 := c_3\frac{\sigma_\varepsilon^2 h^2}{\delta^2}\,,$$

$$C_2 := c_0c_3\mu^{\star 2}\cdot\frac{\gamma_{\max}\sqrt{q^*}}{\gamma_{\min}\sqrt{p^*}}\cdot\left(\frac{\sqrt{d}\sigma_\varepsilon \vee b^\star}{\sqrt{d\gamma_{\min}}}\right)+1+C_0\,, \qquad x_{\mathfrak{n}} := \sup_{\|V\|_2=1}\mathbb{E}\big[\|X\|_2 \,\big|\, X = \|X\|_2 V\big].$$

Before describing the proof of Theorem 1, we state four lemmas that will be used in the proof. Due to space limitations, we defer proof of the lemmas to the extended version of this paper [17].

Lemma 1. *The force-sampling sets created by the force-sampling rule* (3) *satisfy the following inequalities, for all $t \geq 2\rho\log(\rho)$ provided that $\rho \geq 24$,*

$$\mathbb{P}\Big(|\mathcal{F}_t| \geq 6\rho\log(t)\Big) \leq \frac{1}{t} \qquad \textit{and} \qquad \mathbb{P}\Big(|\mathcal{F}_t| \leq \frac{\rho\log(t)}{2}\Big) \leq \frac{1}{t^3}\,.$$

Lemma 2. *Let $\mathcal{I}$ be a (deterministic) subset of forced-sampling observations and by $\mathcal{I}_\kappa \subseteq \mathcal{I}$, we denote the observations corresponding to arm $\kappa \in [k]$. Then, we have the following inequality*

$$\mathbb{P}\left(\frac{|\mathcal{I}_\kappa|}{|\mathcal{I}|} \leq \frac{1}{2k} \,\middle|\, |\mathcal{I}|\right) \leq e^{-\frac{|\mathcal{I}|}{8k}}\,.$$

Lemma 3. *The tail-bound inequality*

$$\mathbb{P}\left(\exists\kappa \,:\, \big\|\widehat{B}^F_\kappa - B^\star_\kappa\big\|_2 \geq \frac{h}{4}\right) \leq \frac{9}{t^3}$$

holds for all $t \geq 10\max\left\{c_2(k+d)\log(kdt), c_3\frac{\sigma_\varepsilon^2\log(kdt)}{\delta^2}\right\}$.

Lemma 4. *For all $t > 10c_4(k+d)\log(kd)^2$, with probability at least $1-\frac{100k}{t}$ for all $t > c_4(k+d)\log(kd)^2$, we have for $\kappa \in \mathcal{K}_{opt}$,*

$$\big\|\widehat{B}^A_\kappa - B^\star_\kappa\big\|_2 \leq 10c_3r\mu^{\star 2}\cdot\frac{\gamma_{\max}\sqrt{q^*}}{\gamma_{\min}\sqrt{p^*}}\cdot\left(\frac{\sqrt{d}\sigma_\varepsilon \vee b^\star}{\sqrt{d\gamma_{\min}}}\right)\cdot\sqrt{\frac{(k+d)\log(n)}{kn}}\,.$$

Proof of Theorem 1. Following the lines of the proof of Theorem 1 in [6], we define $G(\cdot)$ as

$$G(\mathcal{F}_t) := \begin{cases} 1 & \text{if } \|\widehat{B}^F_\kappa(\mathcal{F}_{t-1}) - B^\star_\kappa\|_2 \leq \frac{h}{4} \text{ for all } \kappa \in [k], \\ 0 & \text{otherwise.}\end{cases}$$

We split the regret of the algorithm into the following three cases and bound each case separately.
(a) Initialization $t \leq c_4$ and forced-sampling phase.
(b) When $t > c_4$ and $G(\mathcal{F}_{t-1}) = 0$.
(c) When $t > c_4$ and $G(\mathcal{F}_{t-1}) = 1$, but a suboptimal arm is chosen due to errors in all-sampling estimates.

By $R_T^{(a)}$, $R_T^{(b)}$ and $R_T^{(c)}$, we denote the regret incurred in the items above, respectively. Clearly, we have that $R_T = R_T^{(a)} + R_T^{(b)} + R_T^{(c)}$.

Before obtaining the upper bounds, note that, for each suboptimal choice, the regret incurred at each step is at most $\langle X, (B^\star_\kappa - B^\star_{\kappa'})\rangle$ for some κ, κ' which in turn is bounded above by

$$|\langle X, (B^\star_\kappa - B^\star_{\kappa'})\rangle| \leq \|X\|_2\cdot\|B^\star_\kappa - B^\star_{\kappa'}\|_2 \leq 2b^\star\cdot\|X\|_2.$$

This fact can be used to obtain regret bounds by bounding the number of times that each suboptimal arm is pulled. Clearly, for part (a), this number is less than or equal to $c_4 + |\mathcal{F}_T|$. Using Chernoff bound, we have that

$$\mathbb{E}\Big[R_T^{(a)}\Big] \leq 2b^\star x_{\mathfrak{n}}\mathbb{E}[c_4 + |\mathcal{F}_T|] \leq 2b^\star x_{\mathfrak{n}}\left(1 + c_4 + 6\rho\log(T)\right)\,.$$

Next, it follows from the definition of $G(\mathcal{F}_{t-1})$ and Lemma 3 that the number of times that $G(\mathcal{F}_{t-1}) = 0$ is controlled by

$$\mathbb{E}\left[\sum_{t=c_4+1}^{T} (1 - G(\mathcal{F}_{t-1}))\right] = \sum_{t=c_4+1}^{T} \mathbb{P}(G(\mathcal{F}_{t-1}) = 0) \leq \sum_{t=c_4+1}^{T} \frac{c_5}{t} \leq c_5 \log(T)\,.$$

It thus entails that $R_T^{(b)} \leq 2b^\star \mathbb{E}\Big[\sum_{t=c_4+1}^{T} \|X_t\|_2 \cdot (1 - G(\mathcal{F}_{t-1}))\Big]$ which leads to

$$R_T^{(b)} \leq 2b^\star x_{\mathfrak{n}} \mathbb{E}\left[\sum_{t=c_4+1}^{T} (1 - G(\mathcal{F}_{t-1}))\right] \leq 2b^\star x_{\mathfrak{n}} c_5 \log(T)\,.$$

Finally, we need to find an upper bound for $R_T^{(c)}$. It follows from a slightly modified version of Lemma EC.18 in [6] that, whenever $G(\mathcal{F}_{t-1}) = 1$, the set $\hat{\mathcal{K}}_t$ contains the optimal arm and no suboptimal arm. In particular, if the best arm is κ^*, we get the following inequality for all $\kappa \in \hat{\mathcal{K}}$

$$0 < \langle X_t, B^\star_{\kappa^*} - B^\star_\kappa \rangle < h \cdot \|X_t\|_2. \tag{4}$$

Therefore, whenever $G(\mathcal{F}_{t-1}) = 1$, for any $\kappa \in \hat{\mathcal{K}}_t$, we have $X \in \mathcal{V}_\kappa$ and if $X \in \mathcal{U}_\kappa$, then $\hat{\mathcal{K}}_t = \{\kappa\}$. Now, we are ready to use Lemma 4 to bound the probability of pulling an incorrect arm. Letting $\mathcal{E}_t^A := \Big\{\exists \kappa \in \mathcal{K}_{opt} : \|\widehat{B}_\kappa^A(\mathcal{A}_t) - B^\star_\kappa\|_2 > c_7 b^\star \sqrt{\tfrac{1}{t}}\Big\}$, and using Lemma 4, we have for all $t > c_4$,

$$\mathbb{P}\big(\mathcal{E}_t^A\big) \leq \frac{100k}{t}\,. \tag{5}$$

Now, let κ^* denote the optimal arm and the arm π_t is the pulled arm. For (random variable) $\kappa \in [k]$, define $\mathcal{D}_\kappa := \big\{\langle X_t, B^\star_{\kappa^*} - B^\star_\kappa\rangle \geq 2\sqrt{\tfrac{c_7}{t}} \cdot b^\star \cdot \|X_t\|_2\big\}$. It follows from the definition of r_t that

$$\begin{aligned}
r_t &= \mathbb{E}\big[\langle X_t, B^\star_{\kappa^*} - B^\star_{\pi_t}\rangle\big] \leq \mathbb{E}\big[\langle X_t, B^\star_{\kappa^*} - B^\star_{\pi_t}\rangle \mathbb{I}\left(\mathcal{D}_{\pi_t} \cup \mathcal{E}_t^A\right)\big] + \mathbb{E}\big[\langle X_t, B^\star_{\kappa^*} - B^\star_{\pi_t}\rangle \mathbb{I}\left(\mathcal{D}^c_{\pi_t} \cap \mathcal{E}_t^{Ac}\right)\big] \\
&\leq \mathbb{E}\Big[\langle X_t, B^\star_{\kappa^*} - B^\star_{\pi_t}\rangle \mathbb{I}\left(\langle X_t, \widehat{B}^A_{\pi_t}\rangle \geq \langle X_t, \widehat{B}^A_{\kappa^*}\rangle, \left(\mathcal{D}_{\pi_t} \cup \mathcal{E}_t^A\right)\right)\Big] + \mathbb{E}\big[\langle X_t, B^\star_{\kappa^*} - B^\star_{\pi_t}\rangle \mathbb{I}\left(\mathcal{D}^c_{\pi_t} \cap \mathcal{E}_t^{Ac}\right)\big] \\
&\leq 2x_{\mathfrak{n}} b^\star \left(\mathbb{P}\Big(\Big\{\langle X_t, \widehat{B}^A_{\pi_t}\rangle \geq \langle X_t, \widehat{B}^A_{\kappa^*}\rangle\Big\} \cap \left(\mathcal{D}_{\pi_t} \cup \mathcal{E}_t^A\right)\Big) + \sqrt{c_7/t}\,\mathbb{P}\big(\mathcal{D}^c_{\pi_t} \cap \mathcal{E}_t^{Ac}\big)\right).
\end{aligned} \tag{6}$$

Note that $\langle X_t, \widehat{B}^A_{\pi_t}\rangle \geq \langle X_t, \widehat{B}^A_{\kappa^*}\rangle$ in combination with the definition of $\mathcal{D}_{\pi_t}$ implies that

$$0 \geq \Big\langle X_t, \widehat{B}^A_{\kappa^*} - B^\star_{\kappa^*}\Big\rangle + \Big\langle X_t, B^\star_{\pi_t} - \widehat{B}^A_{\pi_t}\Big\rangle + 2\sqrt{\frac{c_7}{t}} \cdot b^\star \cdot \|X_t\|_2.$$

And this entails that at least one of the following inequalities hold:

$$\|X_t\|_2 \cdot \|B^\star_{\kappa^*} - \widehat{B}^A_{\kappa^*}\|_2 \geq \Big|\langle X_t, B^\star_{\kappa^*} - \widehat{B}^A_{\kappa^*}\rangle\Big| \geq \sqrt{\frac{c_7}{t}} \cdot b^\star \cdot \|X_t\|_2$$

$$\|X_t\|_2 \cdot \|\widehat{B}^A_{\pi_t} - B^\star_{\pi_t}\|_2 \geq \Big|\langle X_t, \widehat{B}^A_{\pi_t} - B^\star_{\pi_t}\rangle\Big| \geq \sqrt{\frac{c_7}{t}} \cdot b^\star \cdot \|X_t\|_2\,.$$

Since $\hat{\mathcal{K}}_t$ does not contain any suboptimal arm, we have that $\Big\{\langle X_t, \widehat{B}^A_{\pi_t}\rangle \geq \langle X_t, \widehat{B}^A_{\kappa^*}\rangle\Big\} \bigcap D_{\pi_t} \subseteq \mathcal{E}_t^A$. We can use this fact in addition to Lemma 4 to get $\mathbb{P}\Big(\Big\{\langle X_t, \widehat{B}^A_{\pi_t}\rangle \geq \langle X_t, \widehat{B}^A_{\kappa^*}\rangle\Big\} \cap \left(D_{\pi_t} \cup \mathcal{E}_t^A\right)\Big) \leq \frac{1}{t}$ for all $t > c_4$.

Finally, by using the margin condition, we get that

$$\begin{aligned}
&\mathbb{P}\big(\mathcal{D}^c_{\pi_t} \cap \mathcal{E}_t^{Ac}\big) \leq \mathbb{E}\big[\mathbb{I}(\mathcal{D}^c_{\pi_t}) \cdot \mathbb{I}(\mathcal{E}_t^{Ac})\big] = \mathbb{E}\big[\mathbb{E}\big[\mathbb{I}(\mathcal{D}^c_{\pi_t}) \mid \kappa^*, \mathcal{E}_t^A\big] \cdot \mathbb{I}(\mathcal{E}_t^{Ac})\big] \\
&\quad = \mathbb{E}\big[\mathbb{E}\big[N \mid \kappa^*, \mathcal{E}_t^A\big] \cdot \mathbb{I}(\mathcal{E}_t^{Ac})\big] = \mathbb{E}\big[\mathbb{E}\big[N \mid \mathcal{E}_t^A\big] \cdot \mathbb{I}(\mathcal{E}_t^{Ac})\big] \leq \mathbb{E}\left[c_0\sqrt{\frac{c_7}{t}} \cdot \mathbb{I}(\mathcal{E}_t^{Ac})\right] \leq c_0\sqrt{\frac{c_7}{t}}\,.
\end{aligned}$$

Using (5) and (6), we have $r_t \leq 2x_{\mathfrak{n}} \cdot b^\star \cdot \left(\frac{100k + c_0 c_7}{t}\right)$ for all $t > c_4$ which leads to

$$R_T^{(c)} \leq \sum_{t=c_4+1}^{T} 2x_{\mathfrak{n}} \cdot b^\star \cdot \left(\frac{100k + c_0 c_7 |\mathcal{K}_{opt}|}{t}\right) \leq 2x_{\mathfrak{n}} \cdot b^\star \cdot (100k + c_0 c_7) \log(T)\,.$$

$\square$

Acknowledgments

The authors gratefully acknowledge support of the National Science Foundation (CAREER award CMMI: 1554140), Stanford Data Science Initiative, and Human-Centered AI Initiative.

References

[1] Yasin Abbasi-Yadkori. Online Learning for Linearly Parametrized Control Problems. PhD. Thesis, 2012.

[2] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

[3] S. Agrawal, V. Avadhanula, V. Goyal, and A. Zeevi. MNL-Bandit: A Dynamic Learning Approach to Assortment Selection. *ArXiv e-prints*, June 2017.

[4] Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *ICML (3)*, pages 127–135, 2013.

[5] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397–422, 2003.

[6] Hamsa Bastani and Mohsen Bayati. Online decision-making with high-dimensional covariates, 2015. `https://papers.ssrn.com/sol3/papers.cfm?abstract\_id=2661896`.

[7] Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

[8] Emmanuel Candes and Benjamin Recht. Exact matrix completion via convex optimization. *Communications of the ACM*, 55(6):111–119, 2009.

[9] Emmanuel J Candes and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

[10] Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

[11] Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma, and Yuling Yan. Noisy Matrix Completion: Understanding Statistical Guarantees for Convex Relaxation via Nonconvex Optimization. *arXiv e-prints*, page arXiv:1902.07698, Feb 2019.

[12] Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214, 2011.

[13] Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. 2008.

[14] Alexander Goldenshluger and Assaf Zeevi. A linear response bandit problem. *Stochastic Systems*, 3(1):230–261, 2013.

[15] David Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Trans. Information Theory*, 57(3):1548–1566, 2011.

[16] Nima Hamidi and Mohsen Bayati. On low-rank trace regression under general sampling distribution. *arXiv preprint arXiv:1904.08576*, 2019.

[17] Nima Hamidi, Mohsen Bayati, and Kapil Gupta. Personalizing many decisions with high-dimensional covariates. *arXiv preprint*, 2019. Extended Version.

[18] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 665–674. ACM, 2013.

[19] Ramesh Johari, Pete Koomen, Leonid Pekelis, and David Walsh. Peeking at a/b tests: Why it matters, and what to do about it. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1517–1525, New York, NY, USA, 2017. ACM.

[20] Kwang-Sung Jun, Rebecca Willett, Stephen Wright, and Robert Nowak. Bilinear Bandits with Low-rank Structure. *arXiv e-prints*, page arXiv:1901.02470, January 2019.

[21] N. Kallus and M. Udell. Dynamic Assortment Personalization in High Dimensions. *ArXiv e-prints*, October 2016.

[22] Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2009.

[23] Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from noisy entries. *Journal of Machine Learning Research*, 11(Jul):2057–2078, 2010.

[24] Olga Klopp et al. Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, 20(1):282–303, 2014.

[25] Tomáš Kocák, Michal Valko, Rémi Munos, and Shipra Agrawal. Spectral thompson sampling. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, AAAI'14, pages 1911–1917. AAAI Press, 2014.

[26] Vladimir Koltchinskii, Karim Lounici, Alexandre B Tsybakov, et al. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.

[27] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[28] Sahin Lale, Kamyar Azizzadenesheli, Anima Anandkumar, and Babak Hassibi. Stochastic Linear Bandits with Hidden Low Rank Structure. *arXiv e-prints*, page arXiv:1901.09490, Jan 2019.

[29] John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 817–824. Curran Associates, Inc., 2008.

[30] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM, 2010.

[31] Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of machine learning research*, 11(Aug):2287–2322, 2010.

[32] Daniel McFadden. Econometric models for probabilistic choice among products. *The Journal of Business*, 53(3):S13–S29, 1980.

[33] Sahand Negahban and Martin J Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, 13(May):1665–1697, 2012.

[34] Ian Osband, Dan Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, pages 3003–3011, 2013.

[35] Benjamin Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12(Dec):3413–3430, 2011.

[36] Angelika Rohde, Alexandre B Tsybakov, et al. Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2):887–930, 2011.

[37] Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

[38] Steven L Scott. A modern bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26(6):639–658, 2010.

[39] Steven L. Scott. Multi-armed bandit experiments in the online service economy. *Appl. Stoch. Model. Bus. Ind.*, 31(1):37–45, 2015.

[40] Nathan Srebro, Noga Alon, and Tommi S. Jaakkola. Generalization error bounds for collaborative prediction with low-rank matrices. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 1321–1328. 2005.

[41] William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

[42] Michal Valko, Rémi Munos, Branislav Kveton, and Tomáš Kocák. Spectral bandits for smooth graph functions. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pages II–46–II–54. JMLR.org, 2014.