Robust Pure Exploration in Linear Bandits with Limited Budget

Ayya Alieva [1] Ashok Cutkosky [2] Abhimanyu Das [3]

Abstract

We consider the pure exploration problem in the fixed-budget linear bandit setting. We provide a new algorithm that identifies the best arm with high probability while being robust to unknown levels of observation noise as well as to moderate levels of misspecification in the linear model. Our technique combines prior approaches to pure exploration in the multi-armed bandit problem with optimal experimental design algorithms to obtain both problem dependent and problem independent bounds. Our success probability is never worse than that of an algorithm that ignores the linear structure, but seamlessly takes advantage of such structure when possible. Furthermore, we only need the number of samples to scale with the dimension of the problem rather than the number of arms. We complement our theoretical results with empirical validation.

1. Introduction

A variety of problems across disciplines involve making decisions based on noisy observations. For example, online ad companies need to decide which ads to show to users based on noisy estimates of click rates, and machine learning practitioners need to tune hyperparameters based on performance on evaluation sets. All of these situations call for an *exploration period*, after which a decision is made. In the ads setting, the exploration is a testing phase prior to deployment. In machine learning, it is a model development or tuning phase. Further, in all these cases, the exploration phase is limited by a fixed budget: there is only a limited amount of clicks ad companies can use for estimating click rates for ads, or a hyperparameter tuning job will only have a certain amount of resources. This limitation means that the exploration phase should be somehow *efficient* - we wish to make the best use of our limited budget in order to maximize the chance that our final decision is the best one.

We model this problem as a *pure exploration bandit* problem (Even-Dar et al., 2002; Bubeck et al., 2009). We consider a finite set of n possible actions[1], which we call "arms", such that each action x has an associated noisy reward $y \in \mathbb{R}$. We are allowed to take an action (or "pull an arm") and obtain an unbiased estimate of the corresponding reward. We will pull arms for a given exploratory period, observe the resulting reward estimates, and then output our best guess for the arm with the highest reward. The goal is to maximize the probability that our output is indeed the arm with the highest reward.

Often, we have some knowledge about relationships between different arms which we would like to take advantage of during the exploration phase. We model this prior knowledge by embedding our arms as $x \in \mathbb{R}^d$ in such a way that arms that seem qualitatively similar have similar embeddings. Note that this embedding is known - we either choose or are given the features for each arm. We then assume a linear model for the rewards $y = \langle \theta, x \rangle$ for some fixed (but unknown) $\theta \in \mathbb{R}^d$. However, in deference to the reality that such a model is almost certainly not perfect, we allow for a certain amount of misspecification: $y = \langle \theta, x \rangle + \gamma$, where we assume γ is unknown but not too large. This setting can be described as the *misspecified linear bandit*. We provide an algorithm that, after exploring T arms, outputs a suboptimal arm with probability at most:

$$\log_2(n) \exp\left(-\frac{T}{\sigma^2 \log_2(n) \tilde{H}_2}\right) \tag{1}$$

where $\tilde{H}_2$ is an instance-dependent quantity (defined precisely in Section 2) that is bounded by $\frac{d}{\max(\Delta_2 - \sqrt{d}\gamma, 0)^2}$, where Δ_2 indicates the gap in reward y between the second-best arm and the best arm, and σ^2 is the variance of the observational noise.[2] We also provide a lower bound that suggests that the above expression is tight up to logarithmic factors.

[1]Stanford University, Stanford, California, USA [2]Boston University, Boston, Massachussetts, USA [3]Google Research, Mountain View, California, USA. Correspondence to: Ayya Alieva <ayya@stanford.edu>, Ashok Cutkosky <ashok@cutkosky.com>, Abhimanyu Das <abhidas@google.com>.

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

[1] In Section 6 we provide a covering argument that extends to infinite actions

[2] The above expressions supress some additional constants that appear in our formal results.

The pure exploration bandit problem has been studied by several previous authors (Audibert & Bubeck, 2010a; Karnin et al., 2013; Gabillon et al., 2012; Jamieson et al., 2014). Classically, the literature distinguishes between two cases, one in which the goal is to minimize the expected number of samples of an exploration phase required to guarantee a certain success probability of identifying the best arm, and one in which one must maximize the probability of finding the best arm given a certain exploration budget. These are called the *fixed confidence* and *fixed budget* settings respectively. The former case has been studied in the linear bandit setting in a number of prior works (Karnin, 2016; Tao et al., 2018; Soare et al., 2014; Xu et al., 2018; Degenne et al., 2020). Notice that sample complexity bounds for fixed confidence algorithms do not, in general, translate to bounds for the fixed budget setting, and furthermore, the sampling strategies for many fixed confidence algorithms require knowledge of a target success probability, which is not available for fixed budget problems.

Indeed, surprisingly little work (e.g. (Hoffman et al., 2014; Katz-Samuels et al., 2020)) has focused on the fixed-budget setting in linear bandits, which is the focus of this paper. Furthermore, there appears to be a dearth of work on the problem of *misspecified* linear bandits in the pure exploration setting. Prior work on linear bandits (Karnin, 2016; Hoffman et al., 2014; Katz-Samuels et al., 2020) seems to rely heavily on the assumption that the linear model is correct, and it is unclear to what extent their results will degrade when this assumption is violated. Since, in reality, no linear model is likely to be completely correct, this limits the theoretical guarantees of these algorithms in practical situations. Nevertheless, it is intuitively the case that even a linear model with modest misspecification should provide *some* advantage over simply ignoring the feature vectors.

Our algorithm, in addition to obtaining the bound (1), guarantees a mistake probability that gracefully degrades with the level of misspecification. Although our mistake probability without misspecification does not quite match the recent bound in (Katz-Samuels et al., 2020) for non-misspecified settings, note that our algorithm does *not* need to know what the level of misspecification is in advance (i.e. it is adaptive), and moreover we provide an explicit polynomial-time algorithm, while (Katz-Samuels et al., 2020) involves running mirror descent on a subproblem that is only provably convex in particular scenarios. Beyond misspecification, we automatically adapt to the variance of the random observations σ^2, again without requiring this parameter as input. Additionally, our results can also be extended to provide guarantees for the case of problem independent bounds (independent of the Δ_i).

A closely related problem to pure exploration in bandits is the problem of minimizing *regret*. In this setting, the algorithm attempts to minimize the total loss obtained over the set of pulled arms, thus mixing the "exploration" phase with an "exploitation" phase. This setting has been extensively studied in both the general (Auer & Ortner, 2010; Kaufmann et al., 2012; Audibert & Bubeck, 2010b) and linear bandit (Abbasi-Yadkori et al., 2011; Srinivas et al., 2009; Agrawal & Goyal, 2013) setting. Some work has also been done on the problem on misspecified linear models in the cumulative regret setting. (Ghosh et al., 2017; Gopalan et al., 2016) provide algorithms that consider misspecified linear bandits, and show how to obtain small regret when the vector of misspecifications $\gamma \in \mathbb{R}^n$ is bounded in 2-norm. This is a strong restriction on misspecification, but in fact (Ghosh et al., 2017) shows that for the case of minimizing regret, more moderate levels of misspecification may completely destroy any chance for improved performance. In contrast, in our pure exploration setting, we are able to handle γ bounded in ∞-norm, which is much less restrictive.

We organize this paper as follows: in Section 2 we formally describe our setting and notation. In Sections 3 and 4 we describe our algorithm and give its analysis. In Section 5 we provide our lower bound, and in Section 6 we extend our results to obtain a problem independent bound for the mistake probability. In Section 7 we provide a few experimental examples, and in Section 8 we provide some concluding remarks and open problems.

2. Problem Statement

In the pure exploration linear bandit setting, a player is given a set of n arms $\mathcal{A} = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$ with $\|x_i\| \leq 1$ for all i and $n \geq d$. Each x_i is associated with an expected reward y_i. Each y_i takes the form $y_i = \langle \theta, x_i \rangle + \gamma_i$, where θ is some vector in $\mathbb{R}^d$, and γ_i is the deviation from the linear model. Neither θ nor γ is known to the player. We define $\gamma_{\max} = \max_i |\gamma_i|$. At each step of the game, the player chooses one arm x_i of their choice and observes an independent sample $\hat{y}_i = y_i + \zeta$ where ζ is mean-zero σ-subgaussian random variable that is independent of the past history of the game, where σ is also not known to the player. The goal of the game is to query at most T arms $x_{i_1}, \dots, x_{i_T}$ and then output the arm with the highest expected reward.

For ease of notation, we assume that the arms are enumerated in order of the decreasing expected reward, i.e. $y_1 \geq y_2 \geq \dots \geq y_n$. We also assume that the set of arms $\{x_1, \dots, x_n\}$ spans $\mathbb{R}^d$. This is without loss of generality, as we may always change coordinates to a subspace spanned by the arms. We define the gap $\Delta_i = y_1 - y_i$ for $i > 1$, i.e. the difference between the expected rewards of the best arm and the ith best arm. We also define $N = \frac{T}{\log_2 n}$. Further, to simplify our presentation we assume that n, d and $\frac{T}{\log_2(n)}$

are powers of 2, and there is a unique best arm[3].

We will also use the following non-standard notation: given a symmetric positive semidefinite matrix M, and a vector x, we use $M^\ddagger x$ to indicate either ∞ if x is not in the range of M, or $\text{argmin}_{y, My=x} \|y\|_2$ otherwise. This notation is well-defined because the kernel and range of a symmetric positive semi-definition matrix are orthogonal subspaces. Intuitively, $M^\ddagger$ is similar to a pseudo-inverse, but instead of sending kernel elements to 0, we send them to ∞. We will also refer to the matrix norm of x with respect to M as $\|x\|_M := \sqrt{x^T M x}$

We will refer to the classical bandit problem in which the n arms are *not* associated with feature vectors $x_i \in \mathbb{R}^d$ as the multi-armed bandit problem. One quantity of interest in the multi-armed bandit setting is $H_2 = \max_{1<i\le n} \frac{i}{\Delta_i^2}$. This quantity frequently appears in error probabilities of algorithms. For example, (Karnin et al., 2013) obtains error probability

$$3\log_2(n)\exp\left(-\frac{T}{16\log_2(n)H_2}\right) \tag{2}$$

In our linear bandit setting, we define a similar quantity:

$$\tilde{d} = \inf_{v} \inf_{\substack{\|\pi\|_1 \le N \\ \pi \in [0,N]^{|\mathcal{A}|}}} \sup_{x_i \in \mathcal{A}} (x_i - v)^T \Big(\sum_{a \in \mathcal{A}} \pi(a) a a^T\Big)^\ddagger (x_i - v)$$

$$h_i = \begin{cases} \min(\frac{7}{4}\tilde{d}, 3i) & \text{if } i > \frac{T}{4\lceil \log_2(n) \rceil} \\ \min(\frac{7}{4}\tilde{d}, i) & \text{if } i \le \frac{T}{4\lceil \log_2(n) \rceil} \end{cases}$$

$$\tilde{H}_2 = \max_{1<i\le n} \frac{h_i}{\max(\Delta_i - (2 + 2\sqrt{2h_i})\gamma_{\max}, 0)^2}$$

We will show that $\tilde{d} \le d$, so that $\tilde{d}$ cannot be very large even in the worst-case. In the next section, we will also draw a connection between $\tilde{H}_2$ and the notion of "characteristic time" (Degenne et al., 2020) used in lower bounds for the fixed confidence problem setting.

In the absence of misspecification, $\tilde{H}_2$ can be bounded by $\min(3H_2, \frac{7\tilde{d}}{4\Delta_2^2})$. Our bounds will depend on $\tilde{H}_2$ in a way analogous to how bounds for the multi-armed bandit setting depend on H_2. Larger values of $\tilde{H}_2$ correspond to more difficult problems. As $\Delta_i - (2 + \sqrt{h_i})\gamma_{\max}$ decreases, the arms become harder to distinguish, and as $\tilde{d}$ increases, our ability to utilize linear structure to gain information about the arms also decreases.

3. Algorithm

Our approach is based on the sequential halving algorithm of (Karnin et al., 2013). We construct a sequence of "candidate sets" $S_0, \dots, S_M$ for each of $M = \log_2(n)$ rounds such that $S_0 = \mathcal{A}$, and $|S_M| = 1$. To get S_{m+1} from S_m, we pick $N = \frac{T}{\log_2 n}$ points $x_{m,1}, \dots, x_{m,N}$ in $\mathcal{A}$, and receive rewards $y_{m,1}, \dots, y_{m,N}$ where $y_{m,i} = \langle \theta, x_{m,i} \rangle + \gamma_{x_{m,i}} + \zeta_{m,i}$. Then, we perform linear regression to get an estimate $\hat{\theta}_m$. Next, for each element $x \in S_m$, we compute $\hat{y} = \langle \hat{\theta}_m, x \rangle$. Finally, we order the elements of S_m according to the values of $\hat{y}$, and remove the bottom half of these elements to obtain S_{m+1}, breaking ties arbitrarily. The final output of our algorithm is the sole element of S_M.

A key step in this procedure is the choice of the N arms we pull in every round. We choose these arms using an *optimal experimental design* algorithm. Specifically, we deploy the method described by (Allen-Zhu et al., 2017), which we refer to as `OptDesign`. This algorithm provides a way to choose a discrete subset of arms whose covariance matrix approximately optimizes a certain objective, which arises organically in the analysis of the failure probability of our approach. In order to employ this result however, we will require $T \ge 45d\log_2(n)$ due to technical limitations on the algorithm of (Allen-Zhu et al., 2017). This requirement is not very restrictive, since T must be at least $\Omega(d)$ for us to be able to even sample all dimensions.

We provide the pseudocode for our pure exploration algorithm in Algorithms 1 and 2.

Algorithm 1 `LinearExploration`

```
Input: total budget $T$, set of arms $\mathcal{A} = \{x_1, \dots, x_n\}$.
Initialize: $S_0 \leftarrow \{x_1, \dots, x_n\} = \mathcal{A}$, $m \leftarrow 0$, $N \leftarrow \frac{T}{\lceil \log_2 n \rceil}$
while $|S_m| > 1$ do
    Pick $\mathcal{Z}_m = \{x_{m,1}, ..., x_{m,N}\} \leftarrow$ GetArms($S_m, N, \mathcal{A}$).
    Sample each arm $x_{m,i} \in \mathcal{Z}_m$ to obtain a reward estimate $\hat{y}_{m,i}$.
    Compute minimum norm OLS estimate
        $$\hat{\theta}_m = \underset{\theta \in span(\mathcal{Z}_m)}{\text{argmin}} \sum_{i=1}^{N} (\langle \theta, x_{m,i} \rangle - \hat{y}_{m,i})^2$$
    Set $S_{m+1}$ be the set of $|S_m|/2$ arms in $S_m$ with largest values of $\hat{y} = \langle \hat{\theta}, x \rangle$.
    $m \leftarrow m + 1$
end while
Return The unique element of $S_m$.
```

Theorem 1. *Let $T > 45d\log_2 n$ and $(2 + 2\sqrt{2h_i})\gamma_{\max} \le \Delta_i$ for all i. Then* `LinearExploration` *makes at most T arm pulls, and fails to return the optimal arm x_1 with probability at most:*

$$3\log_2(n)\exp\left(-\frac{T}{16\log_2(n)\tilde{H}_2\sigma^2}\right)$$

[3] These simplifying assumptions allow us to avoid $\lceil x \rceil$ operators in several places.

```
Algorithm 2 GetArms($S, N, \mathcal{A}$)
  Input: remaining set of arms $S$, accuracy $\varepsilon$, round budget $N$, set of arms to choose $\mathcal{A}$
  Define
      $$f_S(\mathcal{Z}) = \sup_{\substack{x_i \in S \\ x_j \in S}} (x_j - x_i)^T \left(\sum_{z \in \mathcal{Z}} zz^T\right)^{\ddagger} (x_j - x_i)$$
  $\hat{\mathcal{Z}} \leftarrow$ approximation to $\operatorname{argmin}_{\substack{\mathcal{Z} \subseteq \mathcal{A} \\ |\mathcal{Z}| \leq N}} f_S(\mathcal{Z})$ using OptDesign (Allen-Zhu et al., 2017).
  if $N > |S|$ then
    $\tilde{\mathcal{Z}} \leftarrow \{s \in S$ repeated $\frac{N}{|S|}$ times$\}$
    Set $\mathcal{Z} \leftarrow \operatorname{argmin}\{f_S(\hat{\mathcal{Z}}), f_S(\tilde{\mathcal{Z}})\}$
  end if
  return $\mathcal{Z}$
```

We also have the following important bounds on $\tilde{d}$, which is an immediate consequence of Lemma 11, using the Kiefer-Wolfowitz theorem (Kiefer & Wolfowitz, 1960).

Proposition 2. $\tilde{d} \leq d$

It is helpful to compare Theorem 1 to the result for ordinary multi-armed bandit exploration of (Karnin et al., 2013), which obtains an error probability of:

$$3 \log_2(n) \exp\left(-\frac{T}{16 \log_2(n) \sigma^2 \max_{i \leq n} i \Delta_i^{-2}}\right)$$

Notice that in case of a strictly linear model (for which $\gamma_i = 0$ for all i), our definition of $\tilde{H}_2$ simplifies to:

$$\tilde{H}_2 = \max_{1 < i \leq n} \frac{h_i}{\Delta_i^2} \leq \min\left(3H_2, \frac{7\tilde{d}}{4\Delta_2^2}\right)$$

Hence, our dependence on $\tilde{H}_2$ improves on the ordinary multi-armed bandit bound (2). Moreover, our algorithm's performance can be directly bounded in terms of the dimension of the space d rather than any other geometric properties of the arrangement of the arms. Even if the linear structure is essentially useless (i.e. $d \geq n$), we decay gracefully to near-optimal bounds for the pure-exploration multi-armed bandit problem:

For a tighter characterization of our algorithm's performance in terms of the geometric structure of the arms, it is also instructive to consider the following bound on $\tilde{H}_2$

$$\tilde{H}_2 \leq \frac{7 \inf\limits_{\substack{\|\pi\|_1 \leq N \\ \pi \in [0,N]^{|\mathcal{A}|}}} \sup\limits_{x_i \in \mathcal{A}} (x_i - x_1)^T (\sum\limits_{a \in \mathcal{A}} \pi(a) aa^T)^{\ddagger} (x_i - x_1)}{4\Delta_2^2}$$

This expression is somewhat similar in flavor to the characteristic time definition in (Degenne et al., 2020). In that work, it is shown that for the *fixed confidence* setting, if the target failure probability is δ and the (random) number of trials is T, then

$$\liminf_{\delta \to 0} \frac{\mathbb{E}[T]}{\log_2(1/\delta)} \geq \inf_{\substack{\|\pi\|_1 \leq 1 \\ \pi \in [0,1]^{|\mathcal{A}|}}} \max_{1 < i \leq n} \frac{(x_i - x_1)^T (\Sigma_\pi)^{\ddagger} (x_i - x_1)}{\Delta_i^2}$$

where we define $\Sigma_\pi = \sum_{a \in \mathcal{A}} \pi(a) aa^T$. Roughly speaking, this suggests that no fixed-confidence algorithm should expect to obtain T for which the error probability δ is less than:

$$\exp\left[-\frac{T}{\inf_{\substack{\|\pi\|_1 \leq 1 \\ \pi \in [0,1]^{|\mathcal{A}|}}} \max_{1 < i \leq n} \frac{(x_i - x_1)^T (\Sigma_\pi)^{\ddagger} (x_i - x_1)}{\Delta_i^2}}\right]$$

Although it is not clear how to convert this result into a lower bound for our fixed-budget setting, this at least provides intuitive evidence that our error bound is measuring the difficulty of the problem using the "right" geometric quantities.

Further, note that our algorithm *does not* need to sample all of the possible arms: it is possible for $T < |\mathcal{A}|$. This property highlights how our method is using the linear structure: by performing regression, we learn about arms that we have never pulled, and are able to infer whether they are worth considering or not.

Finally, comparing the output of OptDesign to a uniform sampling allows us to always at least match the error bound of successive halving. If we were to run OptDesign by itself, the variance of the estimates might be up to 3 times higher than the uniform sampling approach (due to error in optimization and rounding from fractional to integer solutions). In practice, this would rarely be an issue since it is rare for the best experiment design to be uniform sampling. Yet, for the sake of exposition we include the comparison in our algorithm.

We prove Theorem 1 by examining the probability of discarding the correct arm in each round of the loops in the algorithm. Then, we use union bound to bound the overall error probability, in a manner roughly analogous to the strategy followed by (Karnin et al., 2013).

To this end, we first need the following Lemma (whose proof is deffered to the appendix) that bounds the probability of a single arm being misordered when deciding which arms to remove in a round:

Lemma 3. *Assume that the best arm was not eliminated prior to round m. Let $[x]_+ = \max(x, 0)$. Then for any arm*

$x_i \in S_m$,

$$\mathbb{P}[\langle\hat{\theta}_m, x_i\rangle > \langle\hat{\theta}_m, x_1\rangle] \leq \\ \exp\left(-\frac{\left[\Delta_i - (2+2\sqrt{2h_{|S_m|/4}})\gamma_{\max}\right]_+^2 T}{16\log_2(n) h_{|S_m|/4}\sigma^2}\right)$$

We also require the following Lemma:

Lemma 4. *Assume that the best arm was not eliminated prior to round m, and let $[x]_+ = \max(0, x)$. Then the probability that the best arm is eliminated on round m is at most*

$$3\exp\left(-\frac{\left[\Delta_{\frac{1}{4}|S_m|} - (2+2\sqrt{2h_{\frac{1}{4}|S_m|}})\gamma_{\max}\right]_+^2 T}{16\log_2(n) h_{|S_m|/4}\sigma^2}\right)$$

Proof. If the best arm is thrown out at round m, there are at least $\frac{1}{2}|S_m|$ arms in S_m whose $\hat{y}$ estimates are higher than that of the best arm. Let $S'_m \subset S_m$ be the set of arms that excludes the $\frac{1}{4}|S_m|$ arms with the largest true means in S_m. If the best arm is thrown out, then at least $\frac{1}{3}$ of arms in S'_m must have higher $\hat{y}$ estimates than that of the best arm. Let N_m be the number of such arms. Define $D = \max\{(\Delta_{\frac{1}{4}|S_m|} - (2+2\sqrt{2h_{|S_m|/4}})\gamma_{\max})^2, 0\}$. Then using Lemma 3, the expected number of such arms is at most

$$\mathbb{E}[N_m] = \sum_{x_i \in S'_m} \mathbb{P}[\langle\hat{\theta}, x_i\rangle \geq \langle\hat{\theta}, x_1\rangle] \\ \leq |S'_m| \exp\left(-\frac{DT}{16\log_2(n) h_{|S_m|/4}\sigma^2}\right)$$

Then, by Markov inequality, the probability of the best arm being thrown out at round m is at most

$$\mathbb{P}\left[N_m > \frac{1}{3}|S'_m|\right] \leq \frac{\mathbb{E}[N_m]}{\frac{1}{3}|S'_m|} \\ \leq 3\exp\left(-\frac{DT}{16\log_2(n) h_{|S_m|/4}\sigma^2}\right)$$

□

Using the above Lemma, Theorem 1 follows by union bound:

Proof of Theorem 1. Define $D_m = \max\{(\Delta_{\frac{1}{4}|S_m|} - (2 + 2\sqrt{2h_{|S_m|/4}})\gamma_{\max})^2, 0\}$.
Then using Lemma 4 and the union bound, the probability of eliminating the best arm is at most

$$\sum_{m=0}^{\log_2(n)} 3\exp\left(-\frac{D_m T}{16\log_2(n) h_{|S_m|/4}\sigma^2}\right) \\ \leq 3\log_2(n)\exp\left(-\frac{T}{16\log_2(n)\sup_m \frac{h_{|S_m|/4}}{D_m}\sigma^2}\right) \\ \leq 3\log_2(n)\exp\left(-\frac{T}{16\log_2(n)\tilde{H}_2\sigma^2}\right)$$

□

4. Analysis of `GetArms`

In this section, we provide an analysis of the `GetArms` procedure that is used to choose which arms to pull in each round of `LinearExploration`. The heavy-lifting here is performed by the experimental design algorithm proposed by (Allen-Zhu et al., 2017), which we refer to as `OptDesign`. Given $S \subset \mathcal{A} \subset \mathbb{R}^d$, a number N and the objective $f_S : \mathcal{S}_+^d \to R$:

$$f_S(X) = \sup_{x_i, x_j \in S} (x_j - x_i)^T X^{\ddagger}(x_i - x_j)$$

`OptDesign` returns a set $\hat{\mathcal{Z}} = \{z_1, \ldots, z_n\} \subset \mathcal{A}$ such that $\hat{X} = \sum_{z \in \hat{\mathcal{Z}}} zz^\top$ nearly minimizes $f_S(X)$. We verify technical conditions required of f_S in order to employ `OptDesign` in Appendix C.

We use `OptDesign` to obtain a near-optimal set of arms to pull. `GetArms` then potentially improves this near-optimal set by comparing to a few special-case candidate sets of arms. By comparing to these special cases, we mitigate the risk that the approximation factor in the discrete optimization algorithm will cause worse performance than an algorithm that ignores the linear structure.

Our main result is the following, proved in Appendix B:

Theorem 5. *On inputs S, N and $\mathcal{A}$, `GetArms` runs in time polynomial in $|S|$, N, $|\mathcal{A}|$ and d and produces a set $\mathcal{Z} \subset \mathcal{A}$, $|\mathcal{Z}| \leq N$ that satisfies:*

$$\sup_{x_i - x_j \in S} (x_j - x_i)^T \left(\sum_{z \in \hat{\mathcal{Z}}} zz^T\right)^{\ddagger} (x_j - x_i) \leq 8\frac{h_{|S|/4}}{N}$$

Note that this theorem does not use any geometric properties of the arms (e.g. no condition numbers). The quantity $h_{|S_m|/4}$ depends only on the dimension and the relative gaps between the arms.

5. Lower Bound

We next provide a lower bound for the pure exploration linear bandit problem without misspecification, suggesting that $\tilde{H}_2$ is a good complexity measure for this problem. Our lower bound is a reduction to the ordinary multi-armed bandit lower bound (Audibert & Bubeck, 2010a). This result states that for any $p \in (0, 1/2)$, for any algorithm, there exists a d-armed bandit with true rewards in $[p, 1-p]$ and whose observed values are Bernoulli random variables such that the probability of successfully identifying the best arm is at most $1 - \exp\left(\frac{-(5+o(1))T}{p(1-p)H_2}\right)$, where $H_2 = \max_i i\Delta_i^{-2}$. Using this, we have the following:

Theorem 6. *Given a d-dimensional linear bandit pure-exploration algorithm, any $p \in (0, 1/2)$, and any $n \geq d$, there exists an problem instance on which the probability of identifying the best arm is at most $1 - \exp\left(\frac{-(15+o(1))T}{p(1-p)\tilde{H}_2}\right)$, where the $o(1)$ depends on n, T and p and goes to zero as $T \to \infty$.*

Note that this is not an instance dependent lower bound.

We also provide a lower bound that depends on the geometric structure of the problem. We first define some notation: for any vector $w \in \mathcal{R}^d$ and real number γ, we define the design matrix $V_w := \sum_{i=1}^d w_i x_i x_i^T$, and $V_{w,\gamma}$ to be $V_{w,\gamma} = V_w + \gamma I_d$, where I is the identity matrix in $\mathcal{R}^{dxd}$. We define D_d is the probability simplex of dimension d. Our lower bound for any problem instance with dimension d, arms $x_i \in A$ and parameter vector $\theta \in \mathcal{R}^d$ is in terms of the quantity $H_{LB} = \min_{w \in D_d} \max_{x \in \mathcal{A}, x \neq x_1} \min_{\gamma \in \mathcal{R}} \frac{\|x_1 - x\|^2_{V^{-1}_{w,\gamma}}}{(\max(\theta^T V^{-1}_{w,\gamma} V_w (x_1 - x), 0)^2}$.

Theorem 7. *Given a linear bandit pure-exploration algorithm, there exists a problem instance on which the probability of identifying the best arm is at most $1 - \exp\left(-T \cdot (1/\sqrt{H_{LB}} + 2\sin\frac{\pi}{n})^2\right)$.*

The H_{LB} term in the above geometry-dependent lower bound (whose proof is deferred to the appendix) plays a similar role as the $\tilde{H}_2$ quantity in our upper bound. We defer closing the gap between these two terms to future work.

6. Problem Independent Bound

Now we provide an analysis of our algorithm that obtains a *problem independent bound*. This means that the mistake probability will not depend on Δ_i. We do this by redefining the definition of a mistake: instead of bounding the probability of returning the *best* arm, we bound the probability of returning an arm with reward close to highest reward.

To facilitate our discussion, we define the value $x_{1,m}$ to be the best arm remaining in S_m, and $y_{1,m}$ to be the expected reward of this arm. Thus the final output of the algorithm has expected reward $y_{1,\log_2(n)}$. We will show that the probability that $y_{1,m+1} < y_{1,m} - \epsilon$ is small for any given ϵ. Then, by union bound, we will obtain with high probablity $y_{1,\log_2(n)} \geq y_1 - \log_2(n)\epsilon$. Specifically, we have the following lemma:

Lemma 8. *The probability that $y_{1,m+1} < y_{1,m} - \epsilon$ is at most:*

$$3\exp\left(-\frac{\max\{(\epsilon - (2 + \sqrt{2h_{|S_m|/4}})\gamma_{\max})^2, 0\}T}{\log_2(n)h_{|S_m|/4}}\right)$$

Proof. Define $\Delta_{i,m} \leq \Delta_i$ to be the gap between the ith best arm and the best arm remaining in S_m. Then notice that the result of Lemma 3 still holds if we replace x_1 with the best arm remaining in S_m and Δ_i with $\Delta_{i,m}$

Let S_m^ϵ be the set of arms in S_m with y value less than $y_{1,m} - \epsilon$. Notice that in order for $y_{1,m+1}$ to be less than $y_{1,m} - \epsilon$, we must have $|S_m|/2$ elements of $|S_m^\epsilon|$ to have $\hat{y}$ values larger than those of the best arm left in S_m. Let $S_m^{\epsilon'}$ be the set of arms that excludes the $\frac{1}{4}|S_m|$ arms with highest true mean from S_m^ϵ. Then if $y_{1,m+1} < y_{1,m}$, we must have $\frac{1}{3}$ of the arms in $S_m^{\epsilon'}$ have higher $\hat{y}$ estimates than the best arm in S_m. Let N_m be the number of such arms. Define $D = \max\{(\epsilon - (2 + \sqrt{h_{|S_m|/4}})\gamma_{\max})^2, 0\}$.

Then, since all arms in S_m^ϵ have $\Delta_{i,m} \geq \epsilon$, we have by Lemma 3 that

$$\mathbb{E}[N_m] \leq |S_m^{\epsilon'}| \exp\left(-\frac{DT}{\log_2(n)h_{|S_m|/4}}\right)$$

So using Markov inequality in exactly the same way as in the proof of Lemma 3, the conclusion follows. □

This Lemma allows us to prove an analog of Theorem 1:

Theorem 9. *For any given $\epsilon > 0$, the probability that `LinearExploration` returns an arm with true mean more than than $\epsilon \log_2(n)$ lower than y_1 is at most:*

$$3\log_2(n)\exp\left(-\inf_i \frac{\max\{(\epsilon - (2 + \sqrt{2h_i})\gamma_{\max})^2, 0\}T}{\log_2(n)h_i}\right)$$

Note that the h_i in the bound do not depend on the problem instance. The proof is identical to the proof of Theorem 1, using Lemma 8 in the same way as Lemma 4.

In addition to verifying the natural intuition that our algorithm performs reasonably even with very small Δ_2, Theorem 9 lets us design an algorithm that allows for an *infinite* number of arms. For example, we might consider the situation in which $\mathcal{A} = \{x : \|x\| \leq 1\}$. The procedure is computationally inefficient, but straightforward. We require only one more assumption, namely that $\|\theta\| \leq 1$. For simplicity of exposition, also assume that our bandit problem is well-specified ($\gamma_{\max} = 0$), although this is not necessary. Given an error tolerance ϵ, we choose a subset of the arms, $\hat{\mathcal{A}}$, of cardinality at most ϵ^{-d} such that any arm in $\mathcal{A}$ is at most $O(\epsilon)$ far away in the 2-norm from an arm in $\hat{\mathcal{A}}$. Such a choice is possible because we assumed every arm in $\mathcal{A}$ has norm at most one. Since $\|\theta\| \leq 1$, This implies that the highest expected reward in $\hat{\mathcal{A}}$ is at most $O(\epsilon)$ away from the highest expected reward in $\mathcal{A}$.

Our approach is then to run `LinearExploration` on $\hat{\mathcal{A}}$ rather than $\mathcal{A}$. Since $\log_2(|\hat{\mathcal{A}}|) \leq d \log_2(1/\epsilon)$, by Theorem 9, with probability at least

$$1 - 3d \log_2(1/\epsilon) \exp\left(-\inf_i \frac{\epsilon^2 T}{d \log_2(1/\epsilon) h_i}\right)$$

the returned arm's expected reward is optimal to within $O(\epsilon d \log_2(1/\epsilon))$ in $\hat{\mathcal{A}}$, and so within $O(\epsilon d \log_2(1/\epsilon))$ of the best expected reward in $\mathcal{A}$.

Clearly, this method is slow computationally. If $\mathcal{A}$ has some structure, it may be possible to reduce the amount of computation. We leave this question for future work.

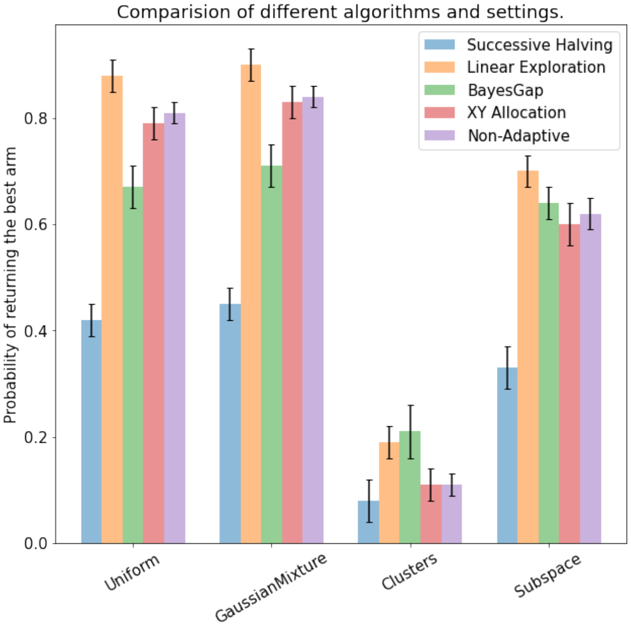

Figure 1. Probabilities of algorithms returning the best arm. Error bars are the standard error, averaged over 200 trials. Budget is $T = 800$, number of arms is $n = 50$, arms have dimensions $d = 3$, and there is no misspecification ($\gamma = 0$). We compared the performance of LinearExploration versus three baselines: BayesGap, Succesive Halving, Non-Adaptive and XY-Allocation. We consider four different settings: Uniform, when arms are sampled uniformly from a hypercube; GaussianMixture, when arms are sampled from a mixture of three gaussians; Clusters, when arms are sampled from two tight clusters; and Subspace, when arm features are close to forming a subspace in R^{d-1}.

7. Experiments

In this section, we empirically validate our approach via synthetic experiments. We tested our algorithm on both truly linear problems (misspecification of 0) and problems with misspecification of order $O(1/d)$.

In Figure 1, we tested four different distributions of arms: a *Uniform* distribution where each arm was sampled uniformly from a hypercube $\{x : \|x\|_\infty \leq 1\}$; a *Gaussian-Mixture* distribution where each arm was sampled from a mixture of three gaussians with constant variance and three different basis vectors chosen as the means; a *Clusters* distribution where arms are sampled uniformly from one of two hypercubes of size $[-\epsilon, \epsilon]^d$ for a small $\epsilon = 0.05$, one centered at θ and one at $-\theta$; and *Subspace*, where arms were sampled uniformly from $[-1, 1]^{d-1} \times [-\epsilon, \epsilon]$, resulting in arms being close to a subspace of R^{d-1}. The experiments used $d = 3$, $n = 50$, $T = 800$, and are averaged over 200 trials.

We compared `LinearExploration` to two baseline fixed-budget algorithms: the successive halving approach from (Karnin et al., 2013), which ignores the linear structure, and the BayesGap algorithm from (Hoffman et al., 2014). In addition, we also compare against the XY-Allocation fixed confidence algorithm from (Soare et al., 2014), where we ignore the stopping criteria and run for the full time budget. We report results with the static version of the XY-Allocation - the adaptive version (and many other adaptive fixed confidence algorithms) requires knowledge of a target success probability for their arm-selection strategy, which is not available in the fixed budget setting. Finally, we also compare `LinearExploration` against a non-adaptive variant (denoted as `Non-Adaptive`), where we just run a single round of `GetArms` and spend the entire budget in a single experimental design call.

Except for the hard *Clusters* instance (where all the algorithms perform relatively poorly due to the small gap between arms), `LinearExploration` performed better than the other baselines in most instances. Note that unlike `LinearExploration`, BayesGap algorithm (Hoffman et al., 2014) requires hyperparameters σ, the variance of the noise, and η, a prior on the variance of θ. In Figure 1, we used the true values for both η and σ. However, these values are unlikely to be known apriori in most applications, which would degrade the performance of BayesGap. All algorithms perform better on simpler arm distributions, but `LinearExploration` seems to exploit the linear structure significantly better than the baselines. In particular, while XY-Allocation also utilizes linear structure, it is outperformed by `LinearExploration` , possibly because fixed-confidence algorithms do not explicitly utilize knowledge of the time horizon that is available in the fixed-budget setting. `LinearExploration` also outperforms

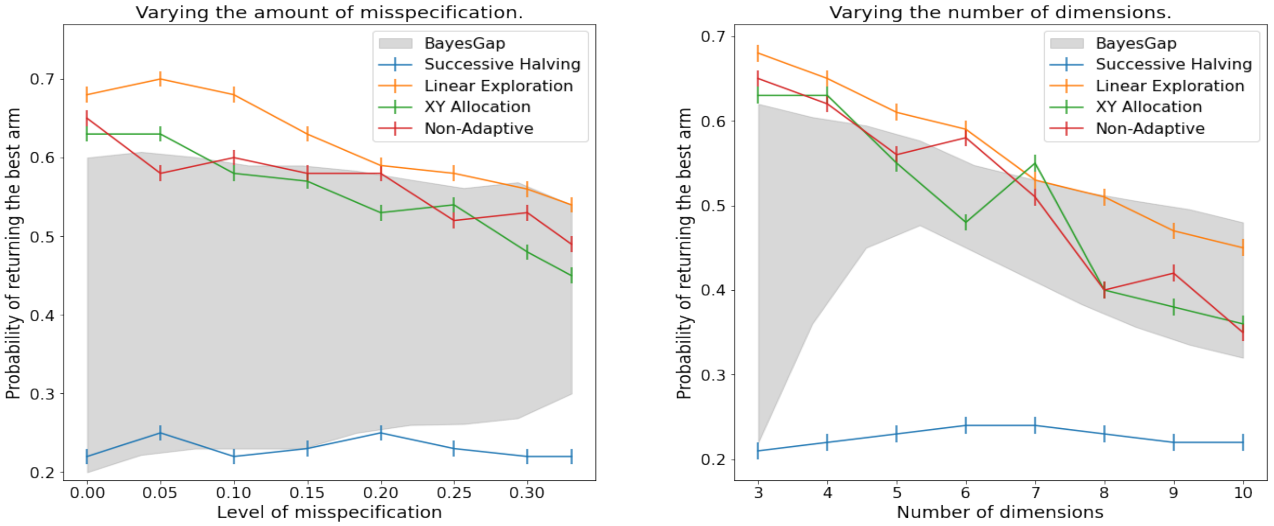

Figure 2. Probabilities of algorithms returning the best arm under varying level of misspecification and number of dimensions. Arms are sampled iid uniformly from a bounded hypercube. Error bars are the standard error, averaged over 200 trials. Budget is $T = 80$, number of arms is $n = 20$. In the figure on the right, misspecification is 0, and dimensions vary from $d = 3$ to $d = 10$. In the figure on the left, arms have dimensions $d = 3$ and the misspecification is sampled uniformly iid for each arm, varying $\|\gamma\|_\infty$ from 0 to $\frac{1}{d} = \frac{1}{3}$.

its `Non-Adaptive` counterpart, confirming that our algorithm's use of multiple adaptive rounds of experiment design are valuable

Figure 2 compares the performance of our algorithm as a function of varying misspecification and number of dimensions d, while fixing $n = 20$ and $T = 80$ for the *Uniform* distribution instance. For BayesGap, we use the shaded grey area to illustrate its dependence on the hyperparameters σ and η. The shaded part shows the one standard error confidence bound on performance of BayesGap under different reasonable choices of the two parameters. For instance, while the real σ and real η are both equal to 1.0 in our setting, the grey area was sampled from 0.5 to 3.0 for σ, and from 0.5 to 2.0 for η. `LinearExploration`, `Non-Adaptive`, Successive Halving, and XY Allocation (Soare et al., 2014) do not require any prior knowledge, so those experiments are plotted as lines with standard error bars.

Since Successive Halving algorithm does not rely on linear structure, its performance does not vary with the number of dimensions or misspecification levels. Similarly, as BayesGap heavily relies on its Bayesian prior, its performance is affected by the specification of σ and η, rather than by misspecification and number of features. While this makes these two baselines seem more robust to a misspecified linear model, this also implies that neither of the two algorithms can successfully leverage the information encoded in the linear structure, as showcased by Figure 2. XY-Allocation and `Non-Adaptive` do rely on linear structure, but not as effectively as `LinearExploration`, as seen from the figures. Thus, even though `LinearExploration` degrades with the level of misspecification and the number of features, the algorithm still outperforms the other baselines as long as the linear structure is able to provide some information.

8. Conclusion

We have introduced a algorithm for pure exploration in the linear bandits setting. Our algorithm is robust in the sense that it does not require any input relating to the distribution of the observation noise, and is even able to handle modest amounts of misspecification. We further demonstrate that our mistake probabilty is nearly tight in the well-specified case. We also show that we can bound method enjoys a *problem independent* mistake probability - that is, a mistake probability that does not depend on the gaps in the values of the arms. We leverage this observation to design an algorithm for the case of infinitely many potential arms. We support our theoretical contributions with an empirical study verifying the desirable properties of our algorithm.

While our work represents a step towards robustifying pure exploration problems to misspecification, we hope for further improvements. For example, we require the misspecification to be a factor of $\sqrt{h_i} \leq \sqrt{d}$ smaller than the gap between the ith best and the best arm. It is unclear if this is tight in general. It might be possible to relax this requirement in some settings and do well so long as the misspecification does not actually reorder the linear model.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pp. 2312–2320, 2011.

Agrawal, S. and Goyal, N. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pp. 127–135, 2013.

Allen-Zhu, Z., Li, Y., Singh, A., and Wang, Y. Near-optimal design of experiments via regret minimization. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 126–135. JMLR. org, 2017.

Audibert, J.-Y. and Bubeck, S. Best arm identification in multi-armed bandits. In *International conference on Algorithmic learning theory*, 2010a.

Audibert, J.-Y. and Bubeck, S. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11(Oct):2785–2836, 2010b.

Auer, P. and Ortner, R. Ucb revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2):55–65, 2010.

Bubeck, S., Munos, R., and Stoltz, G. Pure exploration in multi-armed bandits problems. In *International conference on Algorithmic learning theory*, pp. 23–37. Springer, 2009.

Degenne, R., Menard, P., Shang, X., and Valko, M. Gamification of pure exploration for linear bandits. In *Proceedings of the 37th International Conference on Machine Learning*, Proceedings of Machine Learning Research, 2020.

Even-Dar, E., Mannor, S., and Mansour, Y. Pac bounds for multi-armed bandit and markov decision processes. In *International Conference on Computational Learning Theory*, pp. 255–270. Springer, 2002.

Gabillon, V., Ghavamzadeh, M., and Lazaric, A. Best arm identification: A unified approach to fixed budget and fixed confidence. In *Advances in Neural Information Processing Systems*, pp. 3212–3220, 2012.

Ghosh, A., Chowdhury, S. R., and Gopalan, A. Misspecified linear bandits. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

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

Hoffman, M., Shahriari, B., and Freitas, N. On correlation and budget constraints in model-based bandit optimization with application to automatic machine learning. In *Artificial Intelligence and Statistics*, pp. 365–374, 2014.

Jamieson, K., Malloy, M., Nowak, R., and Bubeck, S. lil'ucb: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*, pp. 423–439, 2014.

Karnin, Z., Koren, T., and Somekh, O. Almost optimal exploration in multi-armed bandits. In *International Conference on Machine Learning*, pp. 1238–1246, 2013.

Karnin, Z. S. Verification based solution for structured mab problems. In *Advances in Neural Information Processing Systems*, pp. 145–153, 2016.

Katz-Samuels, J., Jain, L., Jamieson, K. G., et al. An empirical process approach to the union bound: Practical algorithms for combinatorial and linear bandits. *Advances in Neural Information Processing Systems*, 33, 2020.

Kaufmann, E., Korda, N., and Munos, R. Thompson sampling: An asymptotically optimal finite-time analysis. In *International conference on algorithmic learning theory*, pp. 199–213. Springer, 2012.

Kiefer, J. and Wolfowitz, J. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12: 363–366, 1960.

Soare, M., Lazaric, A., and Munos, R. Best-arm identification in linear bandits. In *Advances in Neural Information Processing Systems*, pp. 828–836, 2014.

Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.

Tao, C., Blanco, S., and Zhou, Y. Best arm identification in linear bandits with linear dimension dependency. In *International Conference on Machine Learning*, pp. 4877–4886, 2018.

Xu, L., Honda, J., and Sugiyama, M. Fully adaptive algorithm for pure exploration in linear bandits. In *International Conference on Artificial Intelligence and Statistics, AISTATS*, 2018.