Does Sparsity Help in Learning Misspecified Linear Bandits?

Jialin Dong [1] **Lin F. Yang** [1]

Abstract

Recently, the study of linear misspecified bandits has generated intriguing implications of the hardness of learning in bandits and reinforcement learning (RL). In particular, Du et al. (2020) shows that even if a learner is given linear features in $\mathbb{R}^d$ that approximate the rewards in a bandit or RL with a uniform error of ε, searching for an $O(\varepsilon)$-optimal action requires pulling at least $\Omega(\exp(d))$ queries. Furthermore, Lattimore et al. (2020) show that a degraded $O(\varepsilon\sqrt{d})$-optimal solution can be learned within $\mathrm{poly}(d/\varepsilon)$ queries. Yet it is unknown whether a structural assumption on the ground-truth parameter, such as sparsity, could break $\varepsilon\sqrt{d}$ barrier. In this paper, we address this question by showing that algorithms can obtain $O(\varepsilon)$-optimal actions by querying $\tilde{O}(\exp(m\varepsilon))$ actions, where m is the sparsity parameter, removing the $\exp(d)$-dependence. We further show (with an information-theoretical lower bound) that this is the best possible if one demands an error $m^\delta\varepsilon$ for $0 < \delta < 1$. We further show that $\mathrm{poly}(m/\varepsilon)$ bounds are possible when the linear features are "good". These results provide a nearly complete picture of how sparsity can help in misspecified bandit learning and provide a deeper understanding of when linear features are "useful" for bandit and reinforcement learning with misspecification.

1. Introduction

Bandit and reinforcement learning problems in real-world applications, e.g., autonomous driving (Kiran et al., 2021), healthcare (Esteva et al., 2019), recommendation system (Bouneffouf et al., 2012), marketing and advertising (Schwartz et al., 2017), are challenging due to the magnificent state/action space. To address this challenge, a function approximation framework has been introduced, which first extracts feature vectors for state/action space and then approximates the value functions of all policies in RL (or the reward functions of all actions in bandit problems) with feature representations. In some real-world applications, feature representations may not have vanilla linear mapping. In these scenarios, a linear feature representation can approximate the value functions (or the reward functions) with a small uniform error known as misspecification. Unfortunately, Du et al. (2020) shows that searching for an $O(\varepsilon)$-optimal action in these scenarios requires pulling at least $\Omega(\exp(d))$ queries. However, if we relax the goal of finding $O(\varepsilon)$-optimal action, there is still a chance. Instead, Lattimore et al. (2020) find an action that is suboptimal with an error of at most $O(\varepsilon\sqrt{d})$ within $\mathrm{poly}(d/\varepsilon)$ queries, where d is the dimension of the feature vectors.

By scrutinizing the novel result proposed by Lattimore et al. (2020), the dependence on $\sqrt{d}$ raises concern regarding the potential blowup of the approximation error. We are modestly optimistic that some structural patterns, such as sparsity, in feature representation schemes are beneficial to break the $\varepsilon\sqrt{d}$ barrier. This idea comes from a vast literature that studies high-dimensional statistics in sparse linear regression (Bühlmann & Van De Geer, 2011; Wainwright, 2019) and successfully applies it to sparse linear bandits (Sivakumar et al., 2020; Abbasi-Yadkori et al., 2012; Bastani & Bayati, 2020; Wang et al., 2018; Su et al., 2020; Lattimore et al., 2015). Moreover, the sparsity-structure in linear bandits are meaningful and crucial to many practical applications where there are many potential features but no apparent evidence on which are relevant, such as personalized health care and online advertising (Carpentier & Munos, 2012; Abbasi-Yadkori et al., 2012). The essential difference in sparse linear bandits between our paper and state-of-the-art is the study of the possible model misspecification; i.e., the ground truth reward means might be an ε error away from a sparse linear representation for any action.

Model misspecification is widely seen in practice and has been widely studied only in the dense model (also known as misspecified linear bandits) (Bogunovic & Krause, 2021; Takemura et al., 2021; Zanette et al., 2020a; Wang et al., 2020a), where the best polynomial-sample algorithm suffers a $O(\varepsilon\sqrt{d})$ estimation error, which can be prominent

[1]Department of Electrical and Computer Engineering, University of California, Los Angeles, USA. Correspondence to: Jialin Dong, Lin F. Yang <jialind@g.ucla.edu, linyang@ee.ucla.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

when the feature dimension d is sufficiently large. However, it is unexplored whether a structural sparsity assumption on the ground-truth parameter could break the $\varepsilon\sqrt{d}$ barrier. Additionally, there is little understanding of the conditions when linear features are "useful" for bandit problems and reinforcement learning with misspecification.

Contribution.

- We establish novel algorithms that obtain $O(\varepsilon)$-optimal actions by querying $\tilde{O}(\varepsilon^{-m}d^m)$ actions, where m is the sparsity parameter. For fixed sparsity m, the algorithm finds an $O(\varepsilon)$-optimal action with $\mathrm{poly}(d/\varepsilon)$ queries, breaking the $O(\varepsilon\sqrt{d})$ barrier. The ε^{-m} dependence in the sample bound can be further improved to $O(1)$ if we allow an $O(\varepsilon\sqrt{m})$ suboptimality.
- We establish information-theoretical lower bounds to show that our upper bounds are nearly tight. In particular, we show that any algorithms that can obtain $O(\Delta)$-optimal actions need to query $\Omega(\exp(m\varepsilon/\Delta))$ samples from the bandit environment, where the approximate error Δ mentioned in Definition 1 satisfies $\Delta \geq \varepsilon$. Hence, for approximation error of the form $m^\delta\varepsilon$, for any $0 < \delta < 1$, $\exp(m)$-dependence in the sample complexity is not avoidable.
- We further break the $\exp(m)$ sample barrier by showing an algorithm that achieves $O(m\varepsilon)$ sub-optimal actions while only querying $\mathrm{poly}(dm/\varepsilon)$ samples in the regime the action features possess certain benign structures (hence "good" features).

In summary, our results provide a nearly complete picture of how sparsity can help in misspecified bandit learning and provide a deeper understanding of when linear features are "useful" for bandit and RL with misspecification.

2. Related work

This section summarizes the state-of-the-art in several areas of interest related to our work: function approximation, misspecified feature representation, and sparsity in bandits and reinforcement learning.

Function approximation in bandits and reinforcement learning Function approximation schemes that approximate value functions in RL (reward function in bandit problem) with feature representations are widely used for generalization across large state/action spaces. A recent line of work studies bandits (Ding et al., 2021; Russo & Van Roy, 2013; Dani et al., 2008; Chu et al., 2011) and RL with linear function approximation (Jin et al., 2020; Zanette et al., 2020a; Cai et al., 2020; Zanette et al., 2020b; Agarwal et al., 2020; Neu & Pike-Burke, 2020). Beyond the linear setting, there is a flurry line of research studying RL with general function approximation (Wang et al., 2020b; Osband & Van Roy, 2014; Jiang et al., 2017) and bandits with general function approximation (Li et al., 2017; Kveton et al., 2020; Filippi et al., 2010; Jun et al., 2017; Foster et al., 2021). The regret upper bound $O(\mathrm{poly}(d)\sqrt{n})$ can be achieved in the above papers, where d is the ambient dimension (or complexity measure such as eluder dimension) of the feature space and n is the number of rounds.

Misspecified bandits and reinforcement learning Recently, interest has been aroused to deal with the situation when the value function in RL (or the rewards functions in bandits) is approximated by a linear function where the approximation error is at most ε, also known as the misspecified linear bandit and reinforcement learning. The misspecification facilitates us to establish a more complicated reward function than a linear function. For instance, it enables the characterization of a reward function that may change over the rounds, which is common in real-world applications such as education, healthcare, and recommendation systems (Chu et al., 2011).

Du et al. (2020) showed that no matter whether value-based learning or model-based learning, the agent needs to sample an exponential number of trajectories to find an $O(\varepsilon)$-optimal policy for reinforcement learning with ε-misspecified linear features. This result shows that good features (e.g., linear features with small misspecification) are not sufficient for sample-efficient RL if the approximation error guarantee is close to the misspecification error. By relaxing the objective of achieving $O(\varepsilon)$-optimality, Lattimore et al. (2020) showed that $\mathrm{poly}(d/\varepsilon)$ samples are sufficient to obtain an $O(\varepsilon\sqrt{d})$-optimal policy (in the simulator model setting of RL), where d is the feature dimension, indicating the same features are "good" in a different requirement. The hard instances used in both papers are in fact bandit instances and hence provide understanding for misspecified linear bandit problems as well.

A number of works in the literature, such as (Foster et al., 2020; Vial et al., 2022; Takemura et al., 2021; Wei et al., 2022; Jin et al., 2020), can also deal with misspecification in linear bandits or RL with linear features. These algorithms can only achieve a $O(\varepsilon\sqrt{d})$ error guarantee at best (when their regret bounds are translated to PAC bounds) with $\mathrm{poly}(d/\varepsilon)$ samples.

Sparse linear bandits and reinforcement learning In this section, we briefly review the literature on the sparse linear bandits and RL, where no misspecification is considered. We also note that these results are stated in regret bounds, which can be easily converted to PAC bounds.

Abbasi-Yadkori et al. (2012) proposed an online-to-confidence-set conversion approach which achieves a regret upper bound of $O(\sqrt{mdn})$, where m is a known parameter on the sparsity. A matching lower bound is given in (Lattimore & Szepesvári, 2020)[Chapter 24.3], which shows that polynomial dependence on d is generally unavoidable without additional assumptions. To address this limitation, another line of literature (Kim & Paik, 2019; Bastani & Bayati, 2020; Wang et al., 2018) studied the sparse contextual linear bandits where the action set is different in each round and follows some context distribution. Kim & Paik (2019) developed a doubly-robust Lasso bandit approach with an $O(m\sqrt{n})$ upper bound. Bastani & Bayati (2020) considered the scenario where each arm has an underlying parameter and derived a $O(Km^2(\log(n))^2)$ upper bound which was improved to $O(Km^2\log(n))$ by Wang et al. (2018), where K is the number of arms. Sivakumar et al. (2020) proposed a structured greedy algorithm to achieve an $O(m\sqrt{n})$ upper bound. Hao et al. (2020) derived a $\Omega(n^{2/3})$ minimax regret lower bound for sparse linear bandits where the feature vectors lack a well-conditioned exploration distribution.

There are many previous works studying feature selection in reinforcement learning. Specifically, Kolter & Ng (2009); Geist & Scherrer (2012); Painter-Wakefield & Parr (2012); Liu et al. (2012) proposed algorithms with ℓ_1-regularization for temporal-difference (TD) learning. Ghavamzadeh et al. (2011) and Geist et al. (2012) proposed Lasso-TD to estimate the value function in sparse reinforcement learning and derived finite-sample MDP statistical analysis. Hao et al. (2021a) provided nearly optimal statistical analysis of high dimensional batch reinforcement learning (RL) using sparse linear function approximation. Ibrahimi et al. (2012) derived an $O(d\sqrt{n})$ regret bound in high-dimensional sparse linear quadratic systems where d is the dimension of the state space. The hardness of online reinforcement learning in fixed horizon has been studied by Hao et al. (2021b), which shows that linear regret is generally unavoidable in this case, even if there exists a policy that collects well-conditioned data.

3. Preliminary

Throughout this paper, $f(n) = O(g(n))$ denotes that there exists a constant $c > 0$ such that $|f(n)| \le c|g(n)|$ whereas $f(n) = \Omega(g(n))$ means that there exists a constant $c > 0$ such that $|f(n)| \ge c|g(n)|$. In addition, the notation $f(n) = \Theta(g(n))$ means that there exists constants $c_1, c_2 > 0$ such that $c_1|g(n)| \le |f(n)| \le c_2|g(n)|$. For a given integer n, let $[n]$ denote the set $\{1, \cdots, n\}$. Let $C > 0$ denote a suitably universal large constant. For a matrix $A \in \mathbb{R}^{m\times n}$, the set of rows is denoted by $\mathrm{rows}(A)$. Define an index set $\mathcal{M} \subseteq [d]$ such that $|\mathcal{M}| = m$. Let $\Phi_{\mathcal{M}} \in \mathbb{R}^{k\times m}$ be the submatrix of $\Phi \in \mathbb{R}^{k\times d}$ and $\theta_{\mathcal{M}} \in \mathbb{R}^m$ be the sub-vector of $\theta \in \mathbb{R}^d$.

Consider a bandit problem where the expected rewards are nearly a linear function of their associated features. Let $\Phi \in \mathbb{R}^{k\times d}$ denote the feature matrix whose rows are feature vectors corresponding to k actions. In rounds $t \in [n]$, the agent chooses actions $(a_t)_{t=1}^n$ with $a_t \in \mathrm{rows}(\Phi)$ and receives a reward

$$r_{a_t} = \langle a_t, \theta^* \rangle + \nu_{a_t}\,, \tag{1}$$

where $\nu_{a_t} \in [-\varepsilon, \varepsilon]$, $\varepsilon > 0$ for $t \in [n]$ and $\theta^* \in \mathbb{R}^d$ is an unknown parameter vector. We here only consider *deterministic* rewards as small unbiased noises from rewards do not change the sample complexity analysis of this paper by much but complicate the presentation. In Appendix C, we provide additional discussion on the noisy setting of the rewards.

We make the mild boundedness assumption for each element of the feature matrix such that $\mathrm{rows}(\Phi) \in \mathbb{S}_B^{d-1}$. The parameter vector θ^* is assumed to be m-sparsity:

$$\|\theta^*\|_0 = \sum_{j=1}^{d} \mathbb{1}\{\theta_j^* \neq 0\} = m \;\text{ and }\; \|\theta^*\|_2 \le 1.$$

We also assume that $\forall x \in \mathrm{rows}(\Phi)$, there is $\|x\|_2 \le 1$.

4. Main Results

In this section, we first present an $O(\varepsilon)$-optimal algorithm that takes $O(\varepsilon^{-m}d^m)$ samples in Section 4.1 for ε-misspecified m-sparse linear bandit. Then we derive a nearly matching lower bound in Section 4.2.

4.1. An Algorithm that Breaks the $\Omega(\exp(d))$ Sample Barrier

The core idea of our algorithm is based on an elimination-type argument. In particular, we would guess an estimator $\hat{\theta}$ for θ^* and a index set $\mathcal{M} \subset [d]$. Then for each guess of $\hat{\theta}$ and $\mathcal{M}$, we check the actions that have similar features restricting to $\mathcal{M}$. Querying an action in this group allows us to rule out the guess of $\mathcal{M}$ and $\hat{\theta}$ if they were not correct. If the ground truth θ^* is dense, this algorithm would take $\Omega(\exp(d))$ queries. Fortunately, since $|\mathcal{M}| = m$, we can establish an $O(\varepsilon)$-net with a small size and eliminate the incorrect parameters in an efficient fashion. Below, we present the algorithm more formally.

Define an index set $\mathcal{M} \subseteq [d]$ such that $|\mathcal{M}| = m$. Let $\mathcal{M}^*$ denote the non-zero subset of θ^*. Denote $\mathcal{N}^m$ as a maximal $\varepsilon/2$-separated subset of the Euclidean sphere $\mathbb{S}^{m-1}$ with radius of 1 . The set $\mathcal{N}^m$ satisfies that $\|x - y\|_2 \ge \varepsilon/2$, for all $x, y \in \mathcal{N}^m$, and no subset of $\mathbb{S}^{m-1}$ containing $\mathcal{N}^m$

satisfies this condition. Thus, the size of $\mathcal{N}^m$ is

$$|\mathcal{N}^m| \leq \left(\frac{4}{\varepsilon}+1\right)^m. \tag{2}$$

For a set $\mathcal{M}$, we denote an estimator as $\hat{\theta}_{\mathcal{M}} \in \mathcal{N}^m$ to indicate the estimator which has only non-zero coordinates at $\mathcal{M}$.

For $\forall w \in \mathcal{N}^m$, we collect all $x \in \text{rows}(\Phi)$ close to w by the measurement $|\hat{\theta}_{\mathcal{M}}^\top(x_{\mathcal{M}} - w)|$ where $x_{\mathcal{M}} \in \mathbb{R}^m$ is the sub-vector of $x \in \mathbb{R}^d$ restricted to the index set $\mathcal{M}$ and define the set as

$$\mathcal{R}_{\mathcal{M}}^w(\hat{\theta}_{\mathcal{M}}) := \{x \in \text{rows}(\Phi) : |\hat{\theta}_{\mathcal{M}}^\top(x_{\mathcal{M}} - w)| \leq \frac{\varepsilon}{2}\}. \tag{3}$$

The above set is simply denoted as $\mathcal{R}_{\mathcal{M}}^w$ in the following proof if $\hat{\theta}_{\mathcal{M}}$ is clear from the context. In each round of the algorithm, we find $x \in \mathcal{R}_{\mathcal{M}}^w$ and a set $\mathcal{M}'$ ($\mathcal{M}' \neq \mathcal{M}$) such that $\hat{\theta}_{\mathcal{M}'}^\top x_{\mathcal{M}'}$ deviates from $\hat{\theta}_{\mathcal{M}}^\top w$ (y at least $\Omega(\varepsilon)$). Then, we query such x and receive the corresponding reward r_x. By comparing the difference between r_x and $\hat{\theta}_{\mathcal{M}}^\top w$, we can know whether the subset $\mathcal{M}$ or $\mathcal{M}'$ of x is more likely to determine the reward r_x and rule out the incorrect parameters. For $x \in \mathcal{R}_{\mathcal{M}}^w$, let $[x]_{\mathcal{N}^m}$ denote the vector $v = \arg\min_{w \in \mathcal{N}^m} \|w - x_{\mathcal{M}}\|_2$ where $x_{\mathcal{M}} \in \mathbb{R}^m$ is the sub-vector of x. Let $(\sim, \mathcal{M}, \hat{\theta}_{\mathcal{M}}) \in \mathcal{S}$ denote all of the elements involving the index set $\mathcal{M}$ and $\hat{\theta}_{\mathcal{M}} \in \mathcal{N}^m$. We present the full algorithm in Algorithm 1.

Algorithm 1 Parameter Elimination

1: **Input:** feature matrix $\Phi \in \mathbb{R}^{k \times d}$
2: **Initialize:** $\mathcal{S} := \{(w, \mathcal{M}, \hat{\theta}_{\mathcal{M}}) : w \in \mathcal{N}^m, \mathcal{M} \subseteq [d], |\mathcal{M}| = m, \hat{\theta}_{\mathcal{M}} \in \mathcal{N}^m\}$.
3: For each $(w, \mathcal{M}, \hat{\theta}_{\mathcal{M}}) \in \mathcal{S}$, establish $\mathcal{R}_{\mathcal{M}}^w$ as (3).
4: **while** there exit $(w, \mathcal{M}, \hat{\theta}_{\mathcal{M}}) \in \mathcal{S}$, $\mathcal{M}' \subseteq [d], |\mathcal{M}'| = m$, $\mathcal{M} \neq \mathcal{M}'$, and $x \in \mathcal{R}_{\mathcal{M}}^w$ such that $(\sim, \mathcal{M}', \hat{\theta}_{\mathcal{M}'}) \in \mathcal{S}$, $|\langle x_{\mathcal{M}'}, \hat{\theta}_{\mathcal{M}'}\rangle - \langle w, \hat{\theta}_{\mathcal{M}}\rangle| > 5\varepsilon/2$ **do**
5: Query the action x and receive a reward $r_x = \langle x, \theta^*\rangle + \nu_x$ where $\nu_x \in [-\varepsilon, \varepsilon]$.
6: If $|r_x - \langle w, \hat{\theta}_{\mathcal{M}}\rangle| > 3\varepsilon/2$ then $\mathcal{S} = \mathcal{S} \backslash (\sim, \mathcal{M}, \hat{\theta}_{\mathcal{M}})$, otherwise $\mathcal{S} = \mathcal{S} \backslash (\sim, \mathcal{M}', \hat{\theta}_{\mathcal{M}'})$.
7: **end while**
8: Find a certain set $\mathcal{L} \subseteq [d], |\mathcal{L}| = m$ and corresponding $\hat{\theta}_{\mathcal{L}} \in \mathcal{N}^m$ such that $(\sim, \mathcal{L}, \hat{\theta}_{\mathcal{L}}) \in \mathcal{S}$.
9: **Output:** $\hat{\theta}_{\mathcal{L}}$ and $\mathcal{L}$

Theorem 1. *After*

$$O\left(\left(\frac{1}{\varepsilon}\right)^m \cdot \binom{d}{m}\right)$$

number of queries, the outputs of Algorithm 1, $\hat{\theta}_{\mathcal{L}}$ and $\mathcal{L}$, satisfy $|y_i - \langle a_{\mathcal{L}}^i, \hat{\theta}_{\mathcal{L}}\rangle| \leq O(\varepsilon)$ for all $a^i \in \text{rows}(\Phi)$.

Proof. We first prove the correctness of the algorithm. Suppose for some $(w, \mathcal{M}, \hat{\theta}_{\mathcal{M}}) \in \mathcal{S}$, there is $x \in \mathcal{R}_{\mathcal{M}}^w$ such that $([x_{\mathcal{M}'}]_{\mathcal{N}^m}, \mathcal{M}', \hat{\theta}_{\mathcal{M}'}) \in \mathcal{S}$ and $|\langle x_{\mathcal{M}'}, \hat{\theta}_{\mathcal{M}'}\rangle - \langle w, \hat{\theta}_{\mathcal{M}}\rangle| > 5\varepsilon/2$ and $\mathcal{M}' \neq \mathcal{M}$. Consider two cases in Lines 4-7 in Algorithm 1.

- Case 1: Suppose $|r_x - \langle w, \hat{\theta}_{\mathcal{M}}\rangle| \leq 3\varepsilon/2$, then we have that $|r_x - \langle x_{\mathcal{M}}, \hat{\theta}_{\mathcal{M}}\rangle| \leq 2\varepsilon$ and $|r_x - \langle x_{\mathcal{M}'}, \hat{\theta}_{\mathcal{M}'}\rangle| \geq |\langle x_{\mathcal{M}'}, \hat{\theta}_{\mathcal{M}'}\rangle - \langle w, \hat{\theta}_{\mathcal{M}}\rangle| - |r_x - \langle w, \hat{\theta}_{\mathcal{M}}\rangle| > \varepsilon$. Thus after the iterations, for some $(w, \mathcal{M}, \hat{\theta}_{\mathcal{M}}) \in S$ and $x \in \mathcal{R}_{\mathcal{M}}^w$, we have $|r_x - \langle x_{\mathcal{M}}, \hat{\theta}_{\mathcal{M}}\rangle| \leq 2\varepsilon$. We remove the elements $(\sim, \mathcal{M}', \hat{\theta}_{\mathcal{M}'})$ from $\mathcal{S}$ since there exists an $x \in \text{rows}(\Phi)$ such that $|r_x - \langle x_{\mathcal{M}'}, \hat{\theta}_{\mathcal{M}'}\rangle| > \varepsilon$.
- Case 2: Assume that $|r_x - \langle w, \hat{\theta}_{\mathcal{M}}\rangle| > 3\varepsilon/2$ for some $x \in \mathcal{R}_{\mathcal{M}}^w$. Then the elements $(\sim, \mathcal{M}, \hat{\theta}_{\mathcal{M}})$ get removed from $\mathcal{S}$ since there exists an $x \in \text{rows}(\Phi)$ such that $|r_x - \langle x_{\mathcal{M}}, \hat{\theta}_{\mathcal{M}}\rangle| \geq |r_x - \langle w, \hat{\theta}_{\mathcal{M}}\rangle| - |\langle x_{\mathcal{M}}, \hat{\theta}_{\mathcal{M}}\rangle - \langle w, \hat{\theta}_{\mathcal{M}}\rangle| > \varepsilon$.

Moreover, Algorithm 1 guarantees that

- The elements $(\sim, \mathcal{M}^*, [\theta^*_{\mathcal{M}^*}]_{\mathcal{N}^m})$ maintain in the set $\mathcal{S}$, which involves the ground-truth index set $\mathcal{M}^*$ and $[\theta^*_{\mathcal{M}^*}]_{\mathcal{N}^m} \in \mathcal{N}^m$ such that $|r_x - \langle x_{\mathcal{M}^*}, [\theta^*_{\mathcal{M}^*}]_{\mathcal{N}^m}\rangle| \leq \varepsilon$. Algorithm 1 only eliminates elements $(\sim, \mathcal{M}, \hat{\theta}_{\mathcal{M}})$ involving the index set $\mathcal{M}$ and $\hat{\theta}_{\mathcal{M}}$ such that $|r_x - \langle x_{\mathcal{M}}, \hat{\theta}_{\mathcal{M}}\rangle| > \varepsilon$ for some $x \in \text{rows}(\Phi)$.
- If no more pairs in the remaining set $\mathcal{S}$ satisfies the conditions on Line 4 in Algorithm 1, then it must be the case that, for all $(w, \mathcal{M}, \hat{\theta}_{\mathcal{M}}) \in \mathcal{S}$ with the remaining set $\mathcal{S}$ and $\forall\, x \in \mathcal{R}_{\mathcal{M}}^w$, $|\langle x_{\mathcal{M}^*}, [\theta^*_{\mathcal{M}^*}]_{\mathcal{N}^m}\rangle - \langle w, \hat{\theta}_{\mathcal{M}}\rangle| \leq 5\varepsilon/2$, and hence

$$\begin{aligned} &|r_x - \langle w, \hat{\theta}_{\mathcal{M}}\rangle| = |\langle x, \theta^*\rangle + \nu_x - \langle w, \hat{\theta}_{\mathcal{M}}\rangle| \\ \leq &|\langle x_{\mathcal{M}^*}, [\theta^*_{\mathcal{M}^*}]_{\mathcal{N}^m}\rangle - \langle w, \hat{\theta}_{\mathcal{M}}\rangle| + \varepsilon \leq 7\varepsilon/2, \end{aligned} \tag{4}$$

  Moreover,

$$|r_x - \langle x_{\mathcal{M}}, \hat{\theta}_{\mathcal{M}}\rangle| \leq |r_x - \langle w, \hat{\theta}_{\mathcal{M}}\rangle| + |\hat{\theta}_{\mathcal{M}}^\top(x_{\mathcal{M}} - w)| \leq 4\varepsilon.$$

In summary, for a set $\mathcal{L} \subseteq [d], |\mathcal{L}| = m$ and corresponding $\hat{\theta}_{\mathcal{L}} \in \mathcal{N}^m$ such that $(\sim, \mathcal{L}, \hat{\theta}_{\mathcal{L}})$ in the remaining set $\mathcal{S}$, we can guarantee that

$$|r_x - \langle x_{\mathcal{L}}, \hat{\theta}_{\mathcal{L}}\rangle| \leq 4\varepsilon,$$

for $\forall\, x \in \text{rows}(\Phi)$.

We arrive at the sample complexity analysis of the algorithm. If we find $(w, \mathcal{M}, \hat{\theta}_{\mathcal{M}}) \in \mathcal{S}$, $\mathcal{M}' \neq \mathcal{M}$, $x \in \mathcal{R}_{\mathcal{M}}^w$ satisfying the condition on Line 4 in Algorithm

1, we remove either the elements either $(\sim, \mathcal{M}, \hat{\theta}_{\mathcal{M}})$ or $(\sim, \mathcal{M}', \hat{\theta}_{\mathcal{M}'})$ after querying one action. The loop stops when the condition on Line 4 is not satisfied. Thus, at most $|\mathcal{N}^m|\binom{d}{m}$ queries are needed for the algorithm. Recall $|\mathcal{N}^m|$ (2), the number of queries in m-sparsity case can be bounded by

$$O\left(\left(\frac{1}{\varepsilon}\right)^m \cdot \binom{d}{m}\right).$$

□

When m is a fixed constant, the above theorem demonstrates that $\mathrm{poly}(d)$-samples are sufficient to learn an $O(\varepsilon)$-optimal action. This is in stark contrast to the $\Omega(\exp(d))$ lower-bound provided in Du et al. (2020) and Lattimore et al. (2020). When m is not fixed, the dependence on $\exp(m)$ is not desirable. One may ask, whether it is possible to achieve $\mathrm{poly}(m)$-dependence for some, e.g., relaxed error $m^\delta\varepsilon$ for some $\delta > 0$. Unfortunately, the next section provides a lower bound that rules out the possibility for $\delta < 1$.

4.2. Lower bound

In this section, we establish an information-theoretical lower bound to show that our upper bound is nearly tight. The basic idea is by reduction to the INDEX-QUERY problem (Du et al., 2020; Yao, 1977) using statistical analysis on sub-exponential random variables. More formally, it is shown (Du et al., 2020) that if one is given a vector of dimension n with only one non-zero entry, then it is necessary to query $\Omega(pn)$ entries of the vector to output the index of the entry with probability p. In what follows, we can show that for any algorithm that solves a m-sparse ε-misspecified linear bandit problem, we can use it to solve the INDEX-QUERY problem of size $\Omega(\exp(m))$. The idea is to establish a set of sparse vectors with sub-exponential random variables, such that the vector input to the INDEX-QUERY problem can be embedded into the bandit instance (without any queries to the vector).

The next lemma is the key tool that will be useful in our lower-bound arguments. It shows that there exists a sparse matrix $\Phi \in \mathbb{R}^{k\times d}$ with sufficiently large k where rows have unit norm and sparsity m, and all non-equal rows are almost orthogonal.

Lemma 1. *For $0 < \delta < 1$, $c > 1$ and $C' = \frac{2c^3}{(1+\tau)\sqrt{c^2-1}}$ with sufficiently small $0 \le \tau < 1$,*

- *if $0 < \varepsilon \le \frac{C'm}{d}$, by choosing $k \ge \sqrt{\delta}\exp\left(\frac{d(1+\tau)\varepsilon^2}{4C'}\right)$,*
- *if $\varepsilon > \frac{C'm}{d}$, by choosing $k \ge \sqrt{\delta}\exp\left(\frac{m(1+\tau)\varepsilon}{4}\right)$,*

there exists a feature matrix $\Phi \in \mathbb{R}^{k\times d}$ with rows such that for all $a, b \in \mathrm{rows}(\Phi)$ with $a \ne b$, $\|a\|_2 = 1$, $\|a\|_0 \le m$, and $|\langle a, \rangle| \le \varepsilon$.

Proof Sketch. The matrix is established by choosing each entry of the matrix Φ a small probability ($\sim m/d$) to be non-zero and if it is non-zero, the entry follows a Gaussian distribution. The formal proof is provided in Appendix A. □

As we will show shortly, the matrix in Lemma 1 can be used to *agnostically* embed an arbitrary INDEX-query problem to a sparse misspecified instance. To start with the formal reduction, we introduce the definition of (η, Δ)-sound algorithm for linear bandit problem, where the algorithm returns an estimated optimal action $\hat{a} \in \mathrm{rows}(\Phi)$ and an estimation vector $\hat{\theta} \in \mathbb{R}^d$.

Definition 1. *For any $0 < \eta < 1$ and $\Delta \ge \varepsilon$, an algorithm $\mathcal{A}$ solving linear bandit problem is called sound for (η, Δ) if with probability at least $1-\eta$, algorithm $\mathcal{A}$ returns the estimated optimal action $\hat{a}$ such that $r_{\hat{a}} \ge \max_x r_x - \Delta$.*

For any input vector v to the INDEX-QUERY problem (of dimension k) with some unknown index j to be non-zero, we can simply take Φ as the feature matrix, and the j-th row of Φ to be the ground-truth θ^*. Then we would have $\|v - \Phi\theta^*\|_\infty \le \varepsilon$. Thus any (η, Δ)-sound algorithm for some appropriate Δ would identify the non-zero index in v with good probability and thus inherits the lower bound of INDEX-QUERY. The formal lower bound is presented in the following theorem.

Theorem 2. *For any (η, Δ)-sound linear bandit algorithm $\mathcal{A}$, there exists a m-sparse ε-misspecified linear bandit instance such that algorithm $\mathcal{A}$ takes at least*

$$(1-\eta)\exp\left(c_0 d \cdot \left(\frac{\varepsilon}{\Delta}\right)^2\right), \text{ if } 0 < \frac{\varepsilon}{\Delta} \le \frac{C'm}{d}, \quad (5)$$

$$(1-\eta)\exp\left(\frac{c_1 m(1+\tau)\varepsilon}{\Delta}\right), \text{ if } \frac{\varepsilon}{\Delta} > \frac{C'm}{d}, \quad (6)$$

actions to halt, where c_0, c_1, C' are absolute constants.

Proof. We begin with the construction of the hard m-sparsity instances. Consider an INDEX-QUERY problem with dimension k. Suppose the input vector with the i^*-index (unknown to the algorithm) is non-zero, i.e., e_{i^*}. Here, e_i is the standard unit vector with the i-th coordinate equaling 1. In our hard instance, we choose reward $r_x = 2\Delta$ when $x = a_{i^*}$ with $i^* \in [k]$, otherwise is 0. Now we show that there exists a linear feature representation that approximates the reward vector $\Delta e_{i^*} \in \mathbb{R}^k$ with a uniform error. Based on Lemma 1, let Φ be the matrix $\mathrm{rows}(\Phi) = (a_i)_{i=1}^k$ such that for all $a_i, a_j \in \mathrm{rows}(\Phi)$ with $i \ne j$, $\|a_i\|_2 = 1$ and $|\langle a_i, a_j\rangle| \le \varepsilon/(2\Delta)$. With $\theta^* = 2\Delta a_{i^*}$, we

have $\Phi\theta^* = (2\Delta a_1^\top a_{i^*}, \ldots, 2\Delta a_{i^*}^\top a_{i^*}, \ldots, 2\Delta a_k^\top a_{i^*})^\top$. By choice of Φ, the i^*-th component of $\Phi\theta^*$ is Δ and the others are all less than ε in absolute value. Hence, we can represent the reward vector $2\Delta e_{i^*}$ by $2\Delta e_{i^*} = \Phi\theta^* + \nu$ for some $\nu \in [-\varepsilon, \varepsilon]^k$.

Then an (η, Δ)-sound algorithm would identify an action a, such that with probability at least $1-\eta$, $a^\top\theta^* \geq 2\Delta - \Delta = \Delta$, which is only possible if $a = a_{i^*}$. Hence the algorithm would output i^* with probability at least $1-\eta$. By the lower bound of the INDEX query problem (e.g., Theorem A1 in (Du et al., 2020)), the algorithm takes at least $\Omega[(1-\eta)k]$ queries in the worst-case.

In the construction, we only need Lemma 1 to hold for k with the correct parameters. Hence we have

- if $0 < \varepsilon \leq \frac{C'm}{d}$, then $k \geq \sqrt{\delta}\exp\left(\frac{d(1+\tau)\varepsilon^2}{16C'\Delta^2}\right)$, and
- if $\varepsilon > \frac{C'm}{d}$, then $k \geq \sqrt{\delta}\exp\left(\frac{m(1+\tau)\varepsilon}{8\Delta}\right)$,

for constant τ, δ, and C', completing the proof.

□

5. Improvement on the ε^{-m} Dependence

Even though the dependence of d^m is unavoidable, we can improve the upper bound in Theorem 1 by eluding the dependence of ε. The fundamental idea of the improved algorithm is based on a mix of G-optimal design and elimination argument. Instead of guessing an estimator $\hat{\theta}$ for θ^*, we use G-optimal design to estimate $\hat{\theta}$ concerning an index set $\mathcal{M} \subset [d]$. Then for each estimator $\hat{\theta}$ and $\mathcal{M}$, we check the actions that have similar features restricting to $\mathcal{M}$. The rest of the elimination argument is similar to Section 4.1. Yet the optimal G-optimal design only gives an error guarantee of $O(\varepsilon\sqrt{m})$, which worsens our error guarantee. Below, we present the algorithm more formally.

We start with an essential theorem in G-optimal design which shows that there exists a near-optimal design with a small core set.

Theorem 3 (Todd (2016)). *Given a matrix $A \in \mathbb{R}^{k\times m}$ and a probability distribution $\rho : \text{rows}(A) \to [0,1]$, let $G(\rho) \in \mathbb{R}^{m\times m}$[1] and $g(\rho) \in \mathbb{R}$ be given by*

$$G(\rho) = \sum_{a\in\text{rows}(A)} \rho(a)aa^\top, \quad g(\rho) = \max_{a\in\text{rows}(A)} \|a\|^2_{G(\rho)^{-1}}.$$

There exists a probability distribution ρ such that $g(\rho) \leq 2m$ and the size of the support of ρ is at most $4m\log\log(m) + 16$.

[1]Without loss of generality, we assume $G(\rho)$ is invertible in the rest of the paper. If not, we can discard columns in Φ until the Φ is full column rank.

Remark 1. *The distribution satisfying the results in Theorem 3 can be computed by Frank Wolfe algorithm introduced in (Todd, 2016)[Chapter 3] after $O(km^2)$ computations.*

Let $\mathcal{S} \subset [d]^m$ be all the subsets of cardinality m. For each $\mathcal{M} \in \mathcal{S}$, suppose that $\rho_{\mathcal{M}}$ is a probability distribution over $\text{rows}(\Phi_{\mathcal{M}})$ satisfying the results of Theorem 3, where $\Phi_{\mathcal{M}} \in \mathbb{R}^{k\times m}$ is the sub-matrix of $\Phi \in \mathbb{R}^{k\times d}$. In the following, we use $G_{\mathcal{M}}(\rho_{\mathcal{M}})$ to present $G(\rho)$ defined in Theorem 3 with respect to $\mathcal{M}$. We begin with querying actions to estimate $\hat{\theta}_{\mathcal{M}}$ based on the support of $\rho_{\mathcal{M}}$ and obtain rewards,

$$\hat{\theta}_{\mathcal{M}} = G_{\mathcal{M}}(\rho_{\mathcal{M}})^{-1} \sum_{a\in\text{rows}(\Phi_{\mathcal{M}}),\rho_{\mathcal{M}}(a)\neq 0} \rho_{\mathcal{M}}(a) r_a a, \quad (7)$$

With Theorem 3, we can show that, for all $b \in \text{rows}(\Phi)$ and $\lceil 4m\log\log(m) + 16\rceil$ queries, we have

$$|\langle b_{\mathcal{M}^*}, \hat{\theta}_{\mathcal{M}^*}\rangle - \langle b, \theta^*\rangle| \leq \varepsilon\sqrt{2m}, \quad (8)$$

where $b_{\mathcal{M}^*} \in \mathbb{R}^m$ is the sub-vector of $b \in \mathbb{R}^d$. For $\mathcal{M}, \mathcal{M}' \in \mathcal{S}$, we try to find some $x \in \text{rows}(\Phi)$ making $\hat{\theta}_{\mathcal{M}'}^\top x_{\mathcal{M}'}$ deviate from $\hat{\theta}_{\mathcal{M}}^\top x_{\mathcal{M}}$. We query such x and receive the corresponding reward r_x. By comparing the difference between r_x and $\hat{\theta}_{\mathcal{M}}^\top x_{\mathcal{M}}, \hat{\theta}_{\mathcal{M}'}^\top x_{\mathcal{M}'}$, we can know whether the subset $\mathcal{M}$ or $\mathcal{M}'$ of x is more likely to determine the reward r_x, and hence eliminate the incorrect parameters set. The full algorithm is presented in Algorithm 2.

Algorithm 2 (ε^{-m})-free Algorithm

Input: feature matrix $\Phi \in \mathbb{R}^{k\times d}$
Initialize: $\mathcal{S} := \{\mathcal{M} : \mathcal{M} \subseteq [d], |\mathcal{M}| = m\}$.
For each $\mathcal{M} \in \mathcal{S}$, estimate $\hat{\theta}_{\mathcal{M}}$ based on (7).
while there exit $\mathcal{M}, \mathcal{M}' \in \mathcal{S}$, $\mathcal{M} \neq \mathcal{M}'$, and $x \in \text{rows}(\Phi)$ such that $|\langle x_{\mathcal{M}'}, \hat{\theta}_{\mathcal{M}'}\rangle - \langle x_{\mathcal{M}}, \hat{\theta}_{\mathcal{M}}\rangle| > 2\varepsilon(1+\sqrt{2m})$ **do**
  Query the action x and receive a reward $r_x = \langle x, \theta^*\rangle + \nu_x$ where $\nu_x \in [-\varepsilon, \varepsilon]$.
  If $|r_x - \langle x_{\mathcal{M}}, \hat{\theta}_{\mathcal{M}}\rangle| \leq \varepsilon(1+\sqrt{2m})$ then $\mathcal{S} = \mathcal{S}\backslash\mathcal{M}'$.
  Otherwise $\mathcal{S} = \mathcal{S}\backslash\mathcal{M}$, if $|r_x - \langle x_{\mathcal{M}'}, \hat{\theta}_{\mathcal{M}'}\rangle| > \varepsilon(1+\sqrt{2m})$ then $\mathcal{S} = \mathcal{S}\backslash\mathcal{M}'$.
end while
Find a certain set $\mathcal{L} \subseteq [d], |\mathcal{L}| = m$ such that $\mathcal{L} \in \mathcal{S}$ and estimation $\hat{\theta}_{\mathcal{L}} \in \mathbb{R}^m$.
Output: $\hat{\theta}_{\mathcal{L}}$ and $\mathcal{L}$

Theorem 4. *After $O\left(m\log m \cdot \binom{d}{m}\right)$ number of queries, the outputs of Algorithm 2, $\hat{\theta}_{\mathcal{L}}$ and $\mathcal{L}$, satisfy $|y_i - \langle a_{\mathcal{L}}^i, \hat{\theta}_{\mathcal{L}}\rangle| \leq O(\varepsilon\sqrt{m})$ for all $a^i \in \text{rows}(\Phi)$.*

Proof. We first prove the correctness of the algorithm. Suppose we find some $\mathcal{M}, \mathcal{M}' \in \mathcal{S}$, $\mathcal{M} \neq \mathcal{M}'$, and $x \in \text{rows}(\Phi)$ $|\langle x_{\mathcal{M}'}, \hat{\theta}_{\mathcal{M}'}\rangle - \langle x_{\mathcal{M}}, \hat{\theta}_{\mathcal{M}}\rangle| > 2\varepsilon(1+\sqrt{2m})$. Consider two cases in Lines 4-8 in Algorithm 2.

- Case 1: Suppose we have $|r_x - \langle x_{\mathcal{M}}, \hat{\theta}_{\mathcal{M}}\rangle| \leq \varepsilon(1+\sqrt{2m})$. We remove the element $\mathcal{M}'$ from $\mathcal{S}$ since there exists an $x \in \text{rows}(\Phi)$ such that $|r_x - \langle x_{\mathcal{M}'}, \hat{\theta}_{\mathcal{M}'}\rangle| \geq |\langle x_{\mathcal{M}'}, \hat{\theta}_{\mathcal{M}'}\rangle - \langle x_{\mathcal{M}}, \hat{\theta}_{\mathcal{M}}\rangle| - |r_x - \langle x_{\mathcal{M}}, \hat{\theta}_{\mathcal{M}}\rangle| > \varepsilon(1+\sqrt{2m})$.
- Case 2: Assume that $|r_x - \langle x_{\mathcal{M}}, \hat{\theta}_{\mathcal{M}}\rangle| > \varepsilon(1+\sqrt{2m})$, then the element $\mathcal{M}$ gets removed from $\mathcal{S}$. We can also remove the other index set $\mathcal{M}'$ from $\mathcal{S}$ if $|r_x - \langle x_{\mathcal{M}'}, \hat{\theta}_{\mathcal{M}'}\rangle| > \varepsilon(1+\sqrt{2m})$.

Moreover, Algorithm 2 guarantees that

- The ground-truth index set $\mathcal{M}^*$ maintains in the set $\mathcal{S}$. According to (8), for all $x \in \text{rows}(\Phi)$, we have $|r_x - \langle x_{\mathcal{M}^*}, \hat{\theta}_{\mathcal{M}^*}\rangle| \leq \varepsilon(1+\sqrt{2m})$. Algorithm 2 only eliminates $\mathcal{M}$ such that $|r_x - \langle x_{\mathcal{M}}, \hat{\theta}_{\mathcal{M}}\rangle| > \varepsilon(1+\sqrt{2m})$ for some $x \in \text{rows}(\Phi)$. After each query, Algorithm 2 removes at least one element from $\mathcal{S}$.
- If no more pair in the remaining set $\mathcal{S}$ satisfies the conditions on Line 4 in Algorithm 2, then it must be the case that, for all $\mathcal{M} \in \mathcal{S}$ with the remaining set $\mathcal{S}$ and $\forall\, x \in \text{rows}(\Phi)$, $|\langle x_{\mathcal{M}^*}, \hat{\theta}_{\mathcal{M}^*}\rangle - \langle x_{\mathcal{M}}, \hat{\theta}_{\mathcal{M}}\rangle| \leq 2\varepsilon(1+\sqrt{2m})$. According to (8), we have

$$\begin{aligned} &|r_x - \langle x_{\mathcal{M}}, \hat{\theta}_{\mathcal{M}}\rangle| \\ \leq &|r_x - \langle x_{\mathcal{M}^*}, \hat{\theta}_{\mathcal{M}^*}\rangle| + |\langle x_{\mathcal{M}^*}, \hat{\theta}_{\mathcal{M}^*}\rangle - \langle x_{\mathcal{M}}, \hat{\theta}_{\mathcal{M}}\rangle| \\ \leq &3\varepsilon(1+\sqrt{2m}). \end{aligned} \tag{9}$$

In summary, for a set $\mathcal{L} \subseteq [d], |\mathcal{L}| = m$ in the remaining set $\mathcal{S}$ and the estimation $\hat{\theta}_{\mathcal{L}} \in \mathbb{R}^m$, we can guarantee that

$$|r_x - \langle x_{\mathcal{L}}, \hat{\theta}_{\mathcal{L}}\rangle| \leq 3\varepsilon(1+\sqrt{2m}),$$

for $\forall\, x \in \text{rows}(\Phi)$.

We arrive at the sample complexity analysis of the algorithm. The estimation on Line 3 in Algorithm 2 takes $\lceil 4m\log\log(m)+16\rceil\binom{d}{m}$ queries. If we find $\mathcal{M}, \mathcal{M}' \in \mathcal{S}$, $\mathcal{M} \neq \mathcal{M}'$, and $x \in \text{rows}(\Phi)$ satisfying the condition on Line 4 in Algorithm 2, we remove at least one element from $\mathcal{M}, \mathcal{M}'$ after querying one action. The loop stops when the condition on Line 4 is not satisfied. Thus, the number of queries in the m-sparsity case can be bounded by $O\left(m\log m \cdot \binom{d}{m}\right)$.

□

6. A $\text{poly}(m)$-sample-complexity Algorithm for Benign Features

The lower bound derived in Section 4.2 does not rule out the possibility of $\exp(m)$-free bound when $\Delta = O(\varepsilon m)$, which we attempt to achieve in this section. The core idea of our algorithm is based on feature compression followed by action-elimination bandit learning. Specifically, we compress the feature vectors and the sparse parameter vector to a lower dimensional vector space, thus converting the sparse linear bandits to a dense case with a much lower dimensional features. Note that this compression is agnostic to the ground-truth parameters. Then we implement action-elimination learning in compressed linear bandits. The detailed algorithm is provided in the following.

We here consider the finite setting where the number of rows, k, in the feature matrix Φ is finite (recall the definition in (1)). This argument is without loss of generality as we can always find an ε-net to cover the actions if there are infinitely many. By Johnson-Linderstrauss lemma (Johnson & Lindenstrauss, 1984), we have that for some $p = \Theta(\log(k)/\upsilon^2)$, there is a function $f : \mathbb{R}^d \to \mathbb{R}^p$ that preserves inner product, i.e., for each $a \in \text{rows}(\Phi)$,

$$\langle f(a), f(\theta^*)\rangle = \langle a, \theta^*\rangle \pm 2\upsilon, \tag{10}$$

for some error $\upsilon > 0$. Such a function can be found efficiently using techniques in, e.g., Kane & Nelson (2014). Hence, we transform the previous sparse linear model $\langle a, \theta^*\rangle$ where $a, \theta^* \in \mathbb{R}^m$ to a new linear model $\langle f(a), f(\theta^*)\rangle$ where $f(a), f(\theta^*) \in \mathbb{R}^p$ with $p < d$. We apply G-optimal design mentioned in (7) to get an estimation of $f(\theta^*)$, i.e., $f(\hat{\theta})$. The detailed algorithm is illustrated in Algorithm 3 where $C > 0$ is a suitable large constant.

Algorithm 3 $\text{poly}(m)$-sample-complexity Algorithm

1: **Input:** feature matrix $\Phi \in \mathbb{R}^{k\times d}$, function $f : \mathbb{R}^d \to \mathbb{R}^p$ (10), the total time steps n
2: **Initialize:** $\mathcal{S} := \{f(a) \in \mathbb{R}^p : a \in \text{rows}(\Phi)\}$
3: **while** number of queries is no greater than n **do**
4: Compute the probability distribution $\rho : \mathcal{S} \to [0,1]$ satisfying the results of Theorem 3.
5: Compute $f(\hat{\theta}) = G(\rho)^{-1}\sum_{a\in\mathcal{S}} \rho(a) r_a a$
6: Update active action set:

$$\mathcal{S} \leftarrow \left\{a \in \mathcal{S} : \max_{b\in\mathcal{S}} \langle f(\hat{\theta}), b-a\rangle \leq C\cdot(\log(k))^{\frac{1}{4}}\sqrt{\varepsilon}\right\}$$

7: **end while**

Theorem 5. *Suppose there is a function $f : \mathbb{R}^d \to \mathbb{R}^p$ satisfied (10). After $n \geq O\left(\sqrt{\log k}/\varepsilon\right)$ number of queries, Algorithm 3 achieves $O((\log(k))^{\frac{1}{4}}\sqrt{\varepsilon})$-optimal actions.*

Proof Sketch. We start with the approximation error of

$f(\theta^*)$.

Similar to the G-optimal design in Section 5, we have

$$\begin{aligned}&|\langle f(a), f(\hat{\theta})\rangle - \langle a, \theta^*\rangle| \\ \leq &|\langle f(a), f(\hat{\theta})\rangle - \langle f(a), f(\theta^*)\rangle| + |\langle f(a), f(\theta^*)\rangle - \langle a, \theta^*\rangle|,\end{aligned} \tag{11}$$

for $\forall\, a \in \text{rows}(\Phi)$.

The first term in (11) can be termed as misspecified linear bandits in $\mathbb{R}^p$. Similar to (8), the first term in (11) can be bounded by

$$|\langle f(a), f(\hat{\theta})\rangle - \langle f(a), f(\theta^*)\rangle| \leq C\left(\varepsilon\sqrt{p}\right) \tag{12}$$

with $O(p\log(p))$ number of queries, where $C > 0$ is a suitably universal large constant. The second term in (11) can be bounded by 2υ. Hence, we have

$$|\langle f(a), f(\hat{\theta})\rangle - \langle a, \theta^*\rangle| \leq C\left(\varepsilon\sqrt{p} + \upsilon\right), \tag{13}$$

Recall that $p = \Theta(\log(k)/\upsilon^2)$, thus $C(\varepsilon\sqrt{p} + \upsilon)$ can be presented as an function with respect to υ, given by

$$g(\upsilon) = C(\varepsilon\sqrt{\log(k)/\upsilon^2} + \upsilon).$$

By optimizing $g(\upsilon)$ with respect to υ, we have the approximation error of $O((\log(k))^{\frac{1}{4}}\sqrt{\varepsilon})$ achieved by the number of queries $O(\sqrt{\log k}/\varepsilon)$.

We can derive the final approximate error as

$$|\langle f(a), f(\hat{\theta})\rangle - \langle a, \theta^*\rangle| \leq C\left((\log(k))^{\frac{1}{4}}\sqrt{\varepsilon}\right). \tag{14}$$

□

Corollary 1. *Based on the notations in Theorem 5, if setting $\upsilon = O(m^\delta\varepsilon)$ for $0 < \delta < 1$, the number of queries $O(m^{1+\delta})$ can be achieved whenever $\log(k) \leq \varepsilon^2 m^{2(1+\delta)}$. Additionally, the regret of Algorithm 3 is bounded by $O(m^\delta\varepsilon n\log(n))$.*

According to Corollary 1, when the coefficient $m^\delta < \sqrt{d}$, Algorithm 3 can break the $\varepsilon\sqrt{d}$ barrier with polynomial samples in all parameters if $\log(k)$ is small, which is achievable if the feature space possesses certain benign structures. E.g., the features are (close to) sparse as the instance in our lower bound construction. This also demonstrates that this result may not admit additional improvement as it resolves the lower bound instance.

All results above focus on the noiseless case. We further give a discussion on the noisy cases in Section C.

7. A $\log(d)$-dependent-sample-complexity Algorithm for General Features

Theorem 5 presents an algorithm with sample complexity dependent on $\log(k)$ where k is the number of actions. Corollary 1 shows that it is possible to achieve sample complexity of $\text{poly}(m)$ when k satisfies the condition $\log(k) \leq \varepsilon^2 m^{2(1+\delta)}$. However, to accommodate a wider range of scenarios, we aim for a sample complexity with a better dependence. In the following section, we will describe the method for achieving a sample complexity dependent on $\log(d)$.

The core idea of our algorithm is to select a submatrix $\Psi \in \mathbb{R}^{k'\times d}$ from the feature matrix $\Phi \in \mathbb{R}^{k\times d}$ where $k' < k$. The submatrix Ψ should contain enough representative actions, which we obtain by using G-optimal design with respect to $\mathcal{M}$ as (7) and collecting the corresponding actions $a \in \mathbb{R}^d$. Then, we apply the same compression process as in Section 6 to reduce the dimensionality of the feature vectors Ψ and the sparse parameter vector θ^*. Finally, we use action-elimination learning in compressed linear bandits. This method consists of two main steps:

1. **G-optimal design with respect to $\mathcal{M}$:** We first find a probability distribution $\rho_{\mathcal{M}}$ over $\text{rows}(\Phi_{\mathcal{M}})$ that meets the conditions of Theorem 3. We then use this distribution $\rho_{\mathcal{M}}$ to generate m^2 distinct feature vectors $a_{\mathcal{M}} \in \mathbb{R}^m$. We collect all the corresponding actions $a \in \mathbb{R}^d$ and denote them as $\Psi \in \mathbb{R}^{\binom{d}{m}\cdot m^2\times d}$.
2. **Compression:** By Johnson-Linderstrauss lemma (Johnson & Lindenstrauss, 1984), we have that for some $q = \Theta(\log(\binom{d}{m}\cdot m^2)/\upsilon^2)$, there is a function $h : \mathbb{R}^d \to \mathbb{R}^q$ that preserves inner product, i.e., for each $a \in \text{rows}(\Psi)$,
$$\langle h(a), h(\theta^*)\rangle = \langle a, \theta^*\rangle \pm 2\upsilon, \tag{15}$$
for some error $\upsilon > 0$.

After inputting the feature matrix Ψ and function h, we can use Algorithm 3 to select actions. The following theorem presents the sample complexity of our method.

Theorem 6. *Suppose there is a function $g : \mathbb{R}^m \to \mathbb{R}^q$ satisfied (15). After $n \geq O\left(\sqrt{m\log(d)}/\varepsilon\right)$ number of queries, the method introduced in this section achieves $O((m\log(d))^{\frac{1}{4}}\sqrt{\varepsilon})$-optimal actions.*

Proof Sketch. We begin with the approximation error of $h(\theta^*)$. For $\forall b \in \text{rows}(\Phi)$, we have

$$\begin{aligned}&|\langle h(b), h(\hat{\theta})\rangle - \langle b, \theta^*\rangle| \\ \leq &|\langle h(b), h(\hat{\theta})\rangle - \langle h(b), h(\theta^*)\rangle| + |\langle h(b), h(\theta^*)\rangle - \langle b, \theta^*\rangle|,\end{aligned} \tag{16}$$

For any $b \in \text{rows}(\Phi)$, the first term can be bounded by

$$\begin{aligned}
&\langle h(b), h(\hat{\theta}) - h(\theta^*)\rangle \\
&\overset{(a)}{=} \left\langle h(b), G(\tilde{\rho})^{-1} \sum_{a\in\text{rows}(\Psi)} \tilde{\rho}(a)\nu_a h(a) \right\rangle \\
&= \sum_{a\in\text{rows}(\Psi)} \tilde{\rho}(a)(\nu_a)\langle h(b), G(\tilde{\rho})^{-1}h(a)\rangle \\
&\overset{(b)}{\leq} \varepsilon \sum_{a\in\text{rows}(\Psi)} \tilde{\rho}(a)|\langle h(b), G(\tilde{\rho})^{-1}h(a)\rangle| \\
&\overset{(c)}{\leq} \varepsilon \sqrt{\sum_{a\in\text{rows}(\Psi)} \tilde{\rho}(a)\langle h(b), G(\tilde{\rho})^{-1}h(a)\rangle^2} \\
&= \varepsilon \sqrt{\sum_{a\in\text{rows}(\Psi)} \tilde{\rho}(a)h(b)^\top G(\tilde{\rho})^{-1}h(a)h(a)^\top G(\tilde{\rho})^{-1}h(b)} \\
&= \varepsilon \sqrt{\|h(b)\|^2_{G(\tilde{\rho})^{-1}}} \\
&\leq \varepsilon \sqrt{\max_{b\in\text{rows}(\Phi)} \|h(b)\|^2_{G(\tilde{\rho})^{-1}}} \\
&\overset{(d)}{\leq} \varepsilon \sqrt{\max_{h(v)} \|h(v)\|^2_{G(\rho)^{-1}}} \leq \varepsilon\sqrt{g(\rho)} \leq \varepsilon\sqrt{2q}.
\end{aligned}$$

Here, step (a) depends on $\nu_a \in [-\varepsilon, \varepsilon]$ and $\tilde{\rho} := \rho \cdot \rho_{\mathcal{M}}$, where ρ is the probability distribution over $h(a)$ for all $a \in \text{rows}(\Psi)$ and $\rho_{\mathcal{M}}$ for all $\mathcal{M}$ determines Ψ. Step (b) derives from Holder's inequality and (c) comes from Jensen's inequality. At step (d), we have $v \in \text{rows}(\Psi)$ and the inequality is due to the construction of Ψ that includes sufficient representative actions. The last two inequalities are derived from the G-optimal design on $h(v)$ for all $v \in \text{rows}(\Psi)$.

We are left to bound the second term in (16). If b in (16) belongs to $\text{rows}(\Psi)$, the second term can be bounded by $2v$. If $b \in \text{rows}(\Phi\backslash\Psi)$, the bound is similar. Since Ψ contains enough representative rows of the feature matrix Φ, we can bound $|\langle h(b), h(\theta^*)\rangle - \langle b, \theta^*\rangle| \leq 2v$ for $b \in \text{rows}(\Phi)$. Hence, we have

$$|\langle h(b), h(\hat{\theta})\rangle - \langle b, \theta^*\rangle| \leq C\left(\varepsilon\sqrt{q} + v\right), \qquad (17)$$

Follow the same arguments in the proof of Theorem 5, we have the approximation error of $O((\log(\binom{d}{m}) \cdot m^2))^{\frac{1}{4}}\sqrt{\varepsilon})$ achieved by the number of queries $O(\sqrt{\log\binom{d}{m} \cdot m^2}/\varepsilon)$.

We can derive the final approximate error as

$$|\langle h(b), h(\hat{\theta})\rangle - \langle b, \theta^*\rangle| \leq C\left(m(\log(d))^{\frac{1}{4}}\sqrt{\varepsilon}\right). \qquad (18)$$

□

The results in Theorem 6 do not depend on the number of actions k, unlike Theorem 5. This is achieved by selecting representative actions and applying compression to get the submatrix Ψ. In other words, this method works for general features, not just benign ones introduced in Section 6. The following corollary restates Theorem 5 and shows the relaxed requirements on the sparse linear bandit model to achieve $O(m\varepsilon)$-optimal actions within $O(m^2)$ queries, which present more general results compared to Corollary 1.

Corollary 2. *Based on the notations in Theorem 6, if* $m > (\log(d)/\varepsilon^2)^{1/3}$, $O(m\varepsilon)$*-optimal actions can be achieved with the number of queries* $O(m^2)$.

8. Conclusions

We aim to utilize the sparsity in linear bandits to remove the $\varepsilon\sqrt{d}$ barrier in the approximation error in existing results (Lattimore et al., 2020) about the misspecified setting. We provide a thorough investigation of how sparsity helps in learning misspecified linear bandits.

We establish novel algorithms that obtain $O(\varepsilon)$-optimal actions by querying $\tilde{O}(\varepsilon^{-m}d^m)$ actions, where m is the sparsity parameter. For fixed sparsity m, the algorithm finds an $O(\varepsilon)$-optimal action with $\text{poly}(d/\varepsilon)$ queries, removing the dependence of $O(\varepsilon\sqrt{d})$. The ε^{-m} dependence in the sample bound can be further improved to $O(1)$ if we instead find an $O(\varepsilon\sqrt{m})$ suboptimal actions. We establish information-theoretical lower bounds to show that our upper bounds are nearly tight. In particular, we show that any algorithms that are able to obtain $O(\Delta)$-optimal actions need to query $\Omega(\exp(m\varepsilon/\Delta))$ samples from the bandit environment. We further break the $\exp(m)$ sample barrier by showing an algorithm that achieves $O(m\varepsilon)$ sub-optimal actions while only querying $\text{poly}(dm/\varepsilon)$ samples in the regime the action features possess certain benign structures (hence "good" features).

Starting from our results on the general bound in misspecified sparse linear bandits, it is interesting to explore results in different bandit learning settings, e.g., contextual bandit problems, RL problems, and distributed/federated-learning settings.

Acknowledgement

This work was supported in part by DARPA under agreement HR00112190130, NSF grant 2221871, and an Amazon Research Grant. We would like to thank Tor Lattimore for proposing the problem during LY's visit to Deepmind, and for the insightful discussions with Tor Lattimore and Botao Hao.

References

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

Agarwal, A., Kakade, S., Krishnamurthy, A., and Sun, W. Flambe: Structural complexity and representation learning of low rank MDPs. *Advances in neural information processing systems*, 33:20095–20107, 2020.

Bastani, H. and Bayati, M. Online decision making with high-dimensional covariates. *Operations Research*, 68 (1):276–294, 2020.

Bogunovic, I. and Krause, A. Misspecified gaussian process bandit optimization. *Advances in Neural Information Processing Systems*, 34:3004–3015, 2021.

Bouneffouf, D., Bouzeghoub, A., and Gançarski, A. L. A contextual-bandit algorithm for mobile context-aware recommender system. In *International conference on neural information processing*, pp. 324–331. Springer, 2012.

Bühlmann, P. and Van De Geer, S. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

Cai, Q., Yang, Z., Jin, C., and Wang, Z. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pp. 1283–1294. PMLR, 2020.

Carpentier, A. and Munos, R. Bandit theory meets compressed sensing for high dimensional stochastic linear bandit. In *Artificial Intelligence and Statistics*, pp. 190–198. PMLR, 2012.

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

Dani, V., Hayes, T. P., and Kakade, S. M. Stochastic linear optimization under bandit feedback. In *COLT*, pp. 355–366, 2008.

Ding, Q., Hsieh, C.-J., and Sharpnack, J. An efficient algorithm for generalized linear bandit: Online stochastic gradient descent and thompson sampling. In *International Conference on Artificial Intelligence and Statistics*, pp. 1585–1593. PMLR, 2021.

Du, S. S., Kakade, S. M., Wang, R., and Yang, L. F. Is a good representation sufficient for sample efficient reinforcement learning? In *International Conference on Learning Representations*, 2020.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., and Dean, J. A guide to deep learning in healthcare. *Nature medicine*, 25(1):24–29, 2019.

Filippi, S., Cappe, O., Garivier, A., and Szepesvári, C. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, pp. 586–594, 2010.

Foster, D., Rakhlin, A., Simchi-Levi, D., and Xu, Y. Instance-dependent complexity of contextual bandits and reinforcement learning: A disagreement-based perspective. In *Conference on Learning Theory*, pp. 2059–2059. PMLR, 2021.

Foster, D. J., Gentile, C., Mohri, M., and Zimmert, J. Adapting to misspecification in contextual bandits. *Advances in Neural Information Processing Systems*, 33: 11478–11489, 2020.

Geist, M. and Scherrer, B. ℓ_1-penalized projected bellman residual. In *European Workshop on Reinforcement Learning*, pp. 89–101. Springer, 2012.

Geist, M., Scherrer, B., Lazaric, A., and Ghavamzadeh, M. A dantzig selector approach to temporal difference learning. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pp. 347–354, 2012.

Ghavamzadeh, M., Lazaric, A., Munos, R., and Hoffman, M. Finite-sample analysis of lasso-td. In *International Conference on Machine Learning*, 2011.

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

Hao, B., Duan, Y., Lattimore, T., Szepesvári, C., and Wang, M. Sparse feature selection makes batch reinforcement learning more sample efficient. In *International Conference on Machine Learning*, pp. 4063–4073. PMLR, 2021a.

Hao, B., Lattimore, T., Szepesvári, C., and Wang, M. Online sparse reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 316–324. PMLR, 2021b.

Ibrahimi, M., Javanmard, A., and Roy, B. Efficient reinforcement learning for high dimensional linear quadratic systems. *Advances in Neural Information Processing Systems*, 25, 2012.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. Contextual decision processes

with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pp. 1704–1713. PMLR, 2017.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pp. 2137–2143. PMLR, 2020.

Johnson, W. B. and Lindenstrauss. Extensions of lipschitz mappings into a hilbert space. *Contemp. Math.*, 26:189–206, 1984.

Jun, K.-S., Bhargava, A., Nowak, R., and Willett, R. Scalable generalized linear bandits: Online computation and hashing. In *Advances in Neural Information Processing Systems*, pp. 99–109, 2017.

Kane, D. M. and Nelson, J. Sparser johnson-lindenstrauss transforms. *Journal of the ACM (JACM)*, 61(1):1–23, 2014.

Kim, G.-S. and Paik, M. C. Doubly-robust lasso bandit. *Advances in Neural Information Processing Systems*, 32, 2019.

Kiran, B. R., Sobh, I., Talpaert, V., Mannion, P., Al Sallab, A. A., Yogamani, S., and Pérez, P. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 2021.

Kolter, J. Z. and Ng, A. Y. Regularization and feature selection in least-squares temporal difference learning. In *Proceedings of the 26th annual international conference on machine learning*, pp. 521–528, 2009.

Kveton, B., Zaheer, M., Szepesvari, C., Li, L., Ghavamzadeh, M., and Boutilier, C. Randomized exploration in generalized linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 2066–2076, 2020.

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

Lattimore, T., Crammer, K., and Szepesvári, C. Linear multi-resource allocation with semi-bandit feedback. *Advances in Neural Information Processing Systems*, 28, 2015.

Lattimore, T., Szepesvari, C., and Weisz, G. Learning with good feature representations in bandits and in RL with a generative model. In *International Conference on Machine Learning*, pp. 5662–5670. PMLR, 2020.

Li, L., Lu, Y., and Zhou, D. Provably optimal algorithms for generalized linear contextual bandits. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 2071–2080. JMLR. org, 2017.

Liu, B., Mahadevan, S., and Liu, J. Regularized off-policy td-learning. *Advances in Neural Information Processing Systems*, 25, 2012.

Neu, G. and Pike-Burke, C. A unifying view of optimism in episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1392–1403, 2020.

Osband, I. and Van Roy, B. Model-based reinforcement learning and the eluder dimension. *Advances in Neural Information Processing Systems*, 27, 2014.

Painter-Wakefield, C. and Parr, R. Greedy algorithms for sparse reinforcement learning. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pp. 867–874, 2012.

Ross, S. M. *Introduction to probability models*. Academic press, 2014.

Russo, D. and Van Roy, B. Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems*, 26, 2013.

Schwartz, E. M., Bradlow, E. T., and Fader, P. S. Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science*, 36(4): 500–522, 2017.

Sivakumar, V., Wu, S., and Banerjee, A. Structured linear contextual bandits: A sharp and geometric smoothed analysis. In *International Conference on Machine Learning*, pp. 9026–9035. PMLR, 2020.

Su, Y., Dimakopoulou, M., Krishnamurthy, A., and Dudík, M. Doubly robust off-policy evaluation with shrinkage. In *International Conference on Machine Learning*, pp. 9167–9176. PMLR, 2020.

Takemura, K., Ito, S., Hatano, D., Sumita, H., Fukunaga, T., Kakimura, N., and Kawarabayashi, K.-i. A parameter-free algorithm for misspecified linear contextual bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 3367–3375. PMLR, 2021.

Todd, M. J. *Minimum-volume ellipsoids: Theory and algorithms*. SIAM, 2016.

Vial, D., Parulekar, A., Shakkottai, S., and Srikant, R. Improved algorithms for misspecified linear markov decision processes. In *International Conference on Artificial Intelligence and Statistics*, pp. 4723–4746. PMLR, 2022.

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Wang, R., Salakhutdinov, R. R., and Yang, L. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. *Advances in Neural Information Processing Systems*, 33:6123–6135, 2020a.

Wang, X., Wei, M., and Yao, T. Minimax concave penalized multi-armed bandit model with high-dimensional covariates. In *International Conference on Machine Learning*, pp. 5200–5208. PMLR, 2018.

Wang, Y., Wang, R., Du, S. S., and Krishnamurthy, A. Optimism in reinforcement learning with generalized linear function approximation. In *International Conference on Learning Representations*, 2020b.

Wei, C.-Y., Dann, C., and Zimmert, J. A model selection approach for corruption robust reinforcement learning. In *International Conference on Algorithmic Learning Theory*, pp. 1043–1096. PMLR, 2022.

Yao, A. C.-C. Probabilistic computations: Toward a unified measure of complexity. In *18th Annual Symposium on Foundations of Computer Science (sfcs 1977)*, pp. 222–227. IEEE Computer Society, 1977.

Zanette, A., Lazaric, A., Kochenderfer, M., and Brunskill, E. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pp. 10978–10989. PMLR, 2020a.

Zanette, A., Lazaric, A., Kochenderfer, M. J., and Brunskill, E. Provably efficient reward-agnostic navigation with linear value iteration. *Advances in Neural Information Processing Systems*, 33:11756–11766, 2020b.

A. Proof of Lemma 1

Let $\mathcal{P} = \{a_1, a_2, \cdots, a_k\}$ be a set of k independent random vectors in $\mathbb{R}^d$. For $\forall i \in [k]$, $a_i = [a_{i1}, a_{i2}, \cdots, a_{id}]^\top \in \mathbb{R}^d$, we have

$$a_{i\ell} := \begin{cases} a_{i\ell} \sim \mathcal{N}(0, \frac{1}{m}) & \text{with probability } \frac{m}{d}, \\ 0 & \text{otherwise.} \end{cases} \tag{19}$$

Thus, we have the following properties

$$\begin{aligned} \mathbb{E}[\langle a_i, a_i \rangle] &= 1, \ \forall i \in [k], \\ \mathbb{E}[\langle a_i, a_j \rangle] &= 0, \ \forall i, j \in [k], i \neq j, \\ \mathbb{E}[\|a_i\|_0] &= m, \ \forall i \in [k]. \end{aligned}$$

Based on the above definitions, three steps achieve the proof of Lemma 1:

1. Prove that under certain condition, for any $i, j \in [k]$ with $i \neq j$, with probability at least $1 - \frac{2\delta}{k^2}$, we have $|\langle a_i, a_j \rangle| \leq \varepsilon$. With probability at least $1 - \frac{\delta}{k}$, we have $|\, \|a_i\|_2^2 - 1| \leq \tau$ and $\|a_i\|_0 \leq m + \tau$ for any $i \in [k]$. This is provided in Lemma 2.
2. By a union bound over all the $\binom{k}{2} = k(k-1)/2$ possible pairs of (i, j) mentioned in Step 1, it concludes that for all $i, j \in [k]$ with $i \neq j$, we have $|\langle a_i, a_j \rangle| \leq \varepsilon$ with probability at least $1 - \delta$. We also have $|\, \|a_i\|_2^2 - 1| \leq \tau$ and $\|a_i\|_0 \leq m + \tau$ for all $i \in [k]$ with probability at least $1 - \delta$ by a union bound over all $i \in [k]$.
3. We normalize $\forall\ a_i \in \mathcal{P}$ and get $\tilde{\mathcal{P}} = \{\tilde{a}_1, \tilde{a}_2, \cdots, \tilde{a}_k\}$ where $\|\tilde{a}_i\|_2 = 1$ with $i \in [k]$. From $\|a_i\|_0 \leq m + \tau$ and $0 \leq \tau < 1$ mentioned in Step 2, we can bound $\|\tilde{a}_i\|_0 \leq m$ with $m \in [k]$. Based on Lemma 2 and normalized set $\tilde{\mathcal{P}}$, Theorem 1 presents the condition where the feature matrix $\Phi \in \mathbb{R}^{k \times d}$ in Lemma 1 can be constructed by setting $\text{rows}(\Phi) = (\tilde{a}_i)_{i=1}^k$.

Lemma 2. *Let $0 < \delta < 1$. Consider the set $\mathcal{P} = \{a_1, a_2, \cdots, a_k\}$ described in (19).*

If $0 < \varepsilon \leq \frac{C^2 m}{d}$, by choosing $k \geq \sqrt{\delta} \exp\left(\frac{d\varepsilon^2}{4C^2}\right)$, we have

$$\text{for any } i, j \in [k],\ i \neq j, \quad |\langle a_i, a_j \rangle| \leq \varepsilon \ \text{ with probability at least } 1 - 2\delta/k^2. \tag{20}$$

If $\varepsilon > \frac{C^2 m}{d}$, by choosing $k \geq \sqrt{\delta} \exp\left(\frac{m\varepsilon}{4}\right)$, we have

$$\text{for any } i, j \in [k],\ i \neq j, \quad |\langle a_i, a_j \rangle| \leq \varepsilon \ \text{ with probability at least } 1 - 2\delta/k^2. \tag{21}$$

For sufficiently small τ, $0 \leq \tau < 1$, by choosing $k \geq \frac{\delta}{2} e^{\tau^2/8}$, we have

$$\text{for any } i \in [k], \quad \left| \|a_i\|_2^2 - 1 \right| \leq \tau \ \text{ with probability at least } 1 - \delta/k. \tag{22}$$

Moreover, by choosing $k \geq \delta e^{2\tau^2/d}$, we have

$$\text{for any } i \in [k], \quad \|a_i\|_0 \leq m + \tau \ \text{ with probability at least } 1 - \delta/k. \tag{23}$$

Proof. Please refer to Section B for detailed proof. □

Proposition 1. *Let $0 < \delta < 1$, $0 \leq \tau < 1$, $c > 1$ and $C' = \frac{2c^3}{(1+\tau)\sqrt{c^2-1}}$. Consider the normalized set $\tilde{\mathcal{P}} = \{\tilde{a}_1, \tilde{a}_2, \cdots, \tilde{a}_k\}$ derived from $\mathcal{P}$ (19). For sufficiently small τ, we have*

$$\text{for all } i, j \in [k]\ i \neq j, \quad |\langle \tilde{a}_i, \tilde{a}_j \rangle| \leq \varepsilon, \quad \|\tilde{a}_i\|_0 \leq m \ \text{ with probability at least } 1 - \delta, \tag{24}$$

by choosing $k \geq \sqrt{\delta}\exp\left(\frac{d(1+\tau)\varepsilon^2}{4C'}\right)$ *if* $0 < \varepsilon \leq \frac{C'm}{d}$. *If* $\varepsilon > \frac{C'm}{d}$, *we choose* $k \geq \sqrt{\delta}\exp\left(\frac{m(1+\tau)\varepsilon}{4}\right)$ *to achieve (24).*

Therefore, with probability at least $1-\delta$, *the normalized set* $\tilde{\mathcal{P}}$ *satisfies that for all* $i, j \in [k]$, $i \neq j$, $\langle \tilde{a}_i, \tilde{a}_j \rangle \leq \varepsilon$, $\|\tilde{a}_i\|_0 \leq m$. *Hence, the feature matrix* $\Phi \in \mathbb{R}^{k\times d}$ *in Lemma 1 can be established by choosing* $\mathrm{rows}(\Phi) = (\tilde{a}_i)_{i=1}^k$ *where* $\tilde{a}_i \in \tilde{\mathcal{P}}$ *when* k *is sufficiently large according to Proposition 1.*

B. Proof of Lemma 2

We first introduce some existential definitions and propositions which are helpful to our proof.

Definition 2. *A random variable* X *with mean* $\mu = \mathbb{E}[X]$ *is sub-exponential if there are non-negative parameters* (v, α) *such that*

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\frac{v^2\lambda^2}{2}}, \quad \forall\, |\lambda| < \frac{1}{\alpha}.$$

Proposition 2 (Sub-exponential tail bound)**.** *Assume that* X *is sub-exponential with parameters* (v, α). *Then*

$$\mathbb{P}[|X-\mu| \geq t] \leq \begin{cases} 2e^{-\frac{t^2}{2v^2}}, & 0 \leq t \leq \frac{v^2}{\alpha}, \\ 2e^{-\frac{t}{2\alpha}}, & t > \frac{v^2}{\alpha}. \end{cases}$$

For $\forall\, a \in \mathcal{P}$, each element of a can be taken as the product of two independent random variables, i.e., one is from the Bernoulli distribution and the other is from the Gaussian distribution. Hence, the individual term, i.e., $a_{i\ell}a_{j\ell}$, of $\langle a_i, a_j\rangle = \sum_{\ell=1}^d a_{i\ell}a_{j\ell}$ with $\forall a_i, a_j \in \mathcal{P}$, $i \neq j$ can be represented by a random variable Z_ℓ. Specifically, $Z_\ell = P_\ell X_\ell Q_\ell Y_\ell$ where $\ell \in [d]$ is the product of independent random variables. Herein, P_ℓ and Q_ℓ are independent Bernoulli random variables which take the value 1 with probability m/d and the value 0 with probability $1-m/d$. X_ℓ and Y_ℓ are independent Gaussian random variables drawn from $\mathcal{N}(0, 1/m)$. For $|\lambda| < m$, we have

$$\begin{aligned}
\mathbb{E}[e^{\lambda Z_\ell}] &= \sum_{pq\in\{0,1\}} \mathbb{P}[P_\ell Q_\ell = pq] \cdot \frac{m}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{\lambda(pq)xy}\cdot e^{-m(x^2+y^2)/2}dxdy \\
&= \frac{m}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{\lambda xy}\cdot e^{-m(x^2+y^2)/2}dxdy \cdot \left(\frac{m}{d}\right)^2 \\
&\quad + \frac{m}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-m(x^2+y^2)/2}dxdy \cdot \left(1-\left(\frac{m}{d}\right)^2\right) \\
&\overset{(i)}{\leq} \frac{m}{2\pi}\cdot\frac{2\pi}{\sqrt{m^2-\lambda^2}}\cdot\left(\frac{m}{d}\right)^2 + \frac{m}{2\pi}\cdot\frac{2\pi}{m}\left(1-\left(\frac{m}{d}\right)^2\right) \\
&\leq \frac{m^3}{d^2\sqrt{m^2-\lambda^2}} + 1 \\
&\overset{(ii)}{=} \frac{c^3\lambda^2}{d^2\sqrt{c^2-1}} + 1 \\
&\overset{(iii)}{\leq} e^{\frac{c^3\lambda^2}{d^2\sqrt{c^2-1}}}
\end{aligned} \tag{25}$$

where step (i) comes from

$$\begin{aligned}
&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{\lambda(xy)}e^{-m(x^2+y^2)/2}dxdy \\
=&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-m(x-\frac{\lambda}{m}y)^2/2}e^{\lambda^2y^2/(2m)}e^{-my^2/2}dxdy \\
=&\sqrt{\frac{2\pi}{m}}\int_{-\infty}^{\infty} e^{\lambda^2y^2/(2m)}e^{-m^2y^2/(2m)}dy \\
=&\sqrt{\frac{2\pi}{m}}\int_{-\infty}^{\infty} e^{-y^2(m^2-\lambda^2)/(2m)}dy \\
=&\frac{2\pi}{\sqrt{m^2-\lambda^2}},
\end{aligned} \tag{26}$$

step (ii) is derived by choosing $m = c|\lambda|$, $c > 1$, and step (iii) is due to the fact $x + 1 \leq e^x$.

Following (25) and Definition 2, we find that

$$\mathbb{E}[e^{\lambda Z_\ell}] \leq e^{\frac{c^3\lambda^2}{d^2\sqrt{c^2-1}}} = e^{\frac{v^2\lambda^2}{2}}, \quad \text{for all } |\lambda| < m \text{ and } v^2 = \frac{2c^3}{d^2\sqrt{c^2-1}}, c > 1, \tag{27}$$

which shows that Z_ℓ is sub-exponential with parameters $(v_\ell, \alpha_\ell) = (C/d, 1/m)$ where $C = \sqrt{\frac{2c^3}{\sqrt{c^2-1}}}$ and $c > 1$. Furthermore, the variable $\sum_{\ell=1}^{d}(Z_\ell - \mathbb{E}[Z_\ell])$ is sub-exponential with the parameters (v_*, α_*), where

$$\alpha_* := \max_{\ell=1,\ldots,n} \alpha_\ell = \frac{1}{m} \quad \text{and} \quad v_* := \sqrt{\sum_{\ell=1}^{d} v_\ell^2}.$$

Based on the fact $\mathbb{E}[Z_\ell] = 0$, the tail bound can be derived from Proposition 2,

$$\mathbb{P}\left[\left|\sum_{\ell=1}^{d} Z_\ell\right| \geq t\right] \leq \begin{cases} 2e^{-\frac{t^2}{2v_*^2}}, & 0 \leq t \leq \frac{v_*^2}{\alpha_*}, \\ 2e^{-\frac{t}{2\alpha_*}}, & t > \frac{v_*^2}{\alpha_*}. \end{cases} \tag{28}$$

Thus, we have for two vectors $a_i, a_j \in \mathcal{P}$ and $i \neq j$,

$$\mathbb{P}\left[|\langle a_i, a_j\rangle| \geq t\right] \leq \begin{cases} 2e^{-\frac{dt^2}{2C^2}}, & 0 \leq t \leq \frac{C^2m}{d}, \\ 2e^{-\frac{mt}{2}}, & t > \frac{C^2m}{d}, \end{cases} \tag{29}$$

where $C = \sqrt{\frac{2c^3}{\sqrt{c^2-1}}}$ and $c > 1$. By setting $2e^{-\frac{dt^2}{2C^2}} = 2\delta/k^2$, we have $t = \sqrt{\frac{2C^2}{d}\log(\frac{k^2}{\delta})}$. We choose $k \geq \sqrt{\delta}\exp\left(\frac{d\varepsilon^2}{4C^2}\right)$ such that $t \geq \varepsilon$. Hence, we conclude $\mathbb{P}\left[|\langle a_i, a_j\rangle| \geq \varepsilon\right] \leq 2\delta/k^2$, which implies the statement (20) when $0 < \varepsilon \leq \frac{C^2m}{d}$ in Lemma 2. Similar arguments can be applied to the proof of the statement (21) when $\varepsilon > \frac{C^2m}{d}$ in Lemma 2. The proof of the statement (22) can also be completed by following similar but simpler arguments of proving the statement (20) and (21).

We are left to the proof of statement (23). For $\forall\ a \in \mathcal{P}$, the random variable $\|a\|_0$ obeys the binomial distribution with parameters d and m/d, i.e., $\mathcal{B}(d, m/d)$. It is the discrete probability distribution of the number of d independent Bernoulli trials which return Boolean-valued outcome: the ℓ-th ($\ell \in [d]$) element of a is non-zero (with probability m/d) or zero (with probability $1 - m/d$).

According to the book by Ross (Ross, 2014), we first introduce several properties of the binomial distribution. The cumulative distribution function of binomial distribution $\mathcal{B}(n, p)$ can be represented by

$$\mathbb{F}(k;n,p) = \mathbb{P}[X \le k] = \sum_{i=0}^{\lfloor k \rfloor} \binom{n}{i} p^i (1-p)^{n-i},$$

where we also have $\mathbb{F}(n-k;n,1-p) = 1 - \mathbb{F}(k;n,p)$. Based on Hoeffding's inequality, $F(k;n,p)$ can be bounded by

$$\mathbb{F}(k;n,p) \le \exp\left(-2n\left(p - \frac{k}{n}\right)^2\right).$$

Hence, the upper tail bound for the random variable $\|a\|_0$ is given by

$$\mathbb{P}[\|a\|_0 \ge m + \tau] = \mathbb{F}(d - m - \tau; d, 1 - \frac{m}{d}) \le \exp\left(\frac{-2\tau^2}{d}\right), \tag{30}$$

where $0 \le \tau < 1$. By choosing $k \ge \delta \exp(2\tau^2/d)$, it yields $\mathbb{P}[\|a\|_0 \ge m + \tau] \le \frac{\delta}{k} \le \exp\left(\frac{-2\tau^2}{d}\right)$. Thus, we completed the proof of statement (23).

C. $\mathrm{poly}(m)$-sample-complexity Algorithm for m-sparsity Case with Noise

All results above focus on the noiseless case. We briefly give a discussion on the noisy cases. Consider the stochastic misspecified sparse linear bandits where a feature matrix $\Phi \in \mathbb{R}^{k\times d}$, $x_t \in \mathrm{rows}(\Phi)$, and the reward

$$r_{x_t} = \langle x_t, \theta^* \rangle + \nu_{x_t} + \eta_t \tag{31}$$

where $\nu_{x_t} \in [-\varepsilon, \varepsilon]$ and $\{\eta_t\}$ is a sequence of independent 1-subgaussian random variables.

Based on the reward function (31) and the notation in Algorithm 3, we start with the approximation error of $f(\theta^*)$:

$$\begin{aligned}
&|\langle f(a), f(\hat{\theta})\rangle - \langle a, \theta^*\rangle| \\
\le &|\langle f(a), f(\hat{\theta})\rangle - \langle f(a), f(\theta^*)\rangle| + |\langle f(a), f(\theta^*)\rangle - \langle a, \theta^*\rangle|, \\
\le &\left| f(a)^\top G(\rho)^{-1} \sum_{b_t \in \mathcal{S}} \rho(b_t)\nu_{b_t} b_t + f(a)^\top G(\rho)^{-1} \sum_{b_t \in \mathcal{S}} \rho(b_t) b_t \eta_t \right| + 2\upsilon \\
\le &\left| f(a)^\top G(\rho)^{-1} \sum_{b_t \in \mathcal{S}} \rho(b_t)\nu_{b_t} b_t \right| + \left| f(a)^\top G(\rho)^{-1} \sum_{b_t \in \mathcal{S}} \rho(b_t) b_t \eta_t \right| + 2\upsilon
\end{aligned} \tag{32}$$

for $\forall\, a \in \mathrm{rows}(\Phi)$.

The first term in (32) can be bounded as

$$\begin{aligned}
&\left| f(a)^\top G(\rho)^{-1} \sum_{b_t \in \mathcal{S}} \rho(b_t)\nu_{b_t} b_t \right| \le \varepsilon \sum_{b_t \in \mathcal{S}} \rho(b_t) \left| f(a)^\top G(\rho)^{-1} b_t \right| \\
\le &\varepsilon \sqrt{\left(\sum_{b_t \in \mathcal{S}} \rho(b_t)\right) f(a)^\top \sum_{b_t \in \mathcal{S}} \rho(b_t) G(\rho)^{-1} b_t b_t^\top G(\rho)^{-1} f(a)} \\
= &\varepsilon \sqrt{\sum_{b_t \in \mathcal{S}} \rho(b_t) \|f(a)\|^2_{G(\rho)^{-1}}} \\
\le &2\varepsilon\sqrt{p},
\end{aligned} \tag{33}$$

where is derived from Jensen's inequality and the fact that $\|f(a)\|^2_{G^{-1}} \le 2p/t$ for t-th time step in Algorithm 3. The

second term in 32 can be bounded by standard concentration bounds: with probability at least $1 - 2/(kn)$,

$$\left| f(a)^\top G(\rho)^{-1} \sum_{b_t \in \mathcal{S}} \rho(b_t) b_t \eta_t \right| \leq \|f(a)\|_{G^{-1}} \sqrt{2 \log(kn)} \leq \sqrt{\frac{4p}{t} \log(kn)}. \tag{34}$$

Combining (32), (33), (34), we have

$$|\langle f(a), f(\hat{\theta}) \rangle - \langle a, \theta^* \rangle| \leq 2\varepsilon\sqrt{p} + \sqrt{\frac{4p}{t} \log(kn)} + 2\upsilon. \tag{35}$$

Similarly to the analysis in Section 6, we can derive the final approximate error as

$$\begin{aligned} &|\langle f(a), f(\hat{\theta}) \rangle - \langle a, \theta^* \rangle| \\ \leq & C \left((\log(k))^{\frac{1}{4}} \sqrt{\varepsilon} + \sqrt{\frac{p}{t} \log(kn)} \right). \end{aligned} \tag{36}$$

Based on (36), the active action set in Algorithm 3 in the noise case should be

$$\mathcal{S} \leftarrow \left\{ a \in \mathcal{S} : \max_{b \in \mathcal{S}} \langle f(\hat{\theta}), b - a \rangle \leq C \left((\log(k))^{\frac{1}{4}} \sqrt{\varepsilon} + \sqrt{\frac{p}{t} \log(kn)} \right) \right\}.$$