Improved Regret Bounds for Linear Bandits with Heavy-Tailed Rewards

Artin Tajdini
University of Washington
artin@cs.washington.edu

Jonathan Scarlett
National University of Singapore
scarlett@comp.nus.edu.sg

Kevin Jamieson
University of Washington
jamieson@cs.washington.edu

Abstract

Assumptions of sub-Gaussian or bounded rewards are standard in the theoretical literature on sequential decision making, yet real environments often exhibit heavy-tailed distributions under which such algorithms may fail. We study stochastic linear bandits with heavy-tailed rewards, where the rewards have a finite $(1+\epsilon)$-absolute central moment bounded by υ for some $\epsilon \in (0,1]$. We improve both upper and lower bounds on the minimax regret compared to prior work. When $\upsilon = \mathcal{O}(1)$, the best prior known regret upper bound is $\tilde{\mathcal{O}}(dT^{\frac{1}{1+\epsilon}})$. While a lower with the same scaling has been given, it relies on a construction using $\upsilon = \mathcal{O}(d)$, and adapting the construction to the bounded-moment regime with $\upsilon = \mathcal{O}(1)$ yields only a $\Omega(d^{\frac{\epsilon}{1+\epsilon}}T^{\frac{1}{1+\epsilon}})$ lower bound. This matches the known rate for multi-armed bandits and is generally loose for linear bandits, in particular being $\sqrt{d}$ below the optimal rate in the finite-variance case ($\epsilon = 1$). We propose a new elimination-based algorithm guided by experimental design, which achieves regret $\tilde{\mathcal{O}}(d^{\frac{1+3\epsilon}{2(1+\epsilon)}}T^{\frac{1}{1+\epsilon}})$, thus improving the dependence on d for all $\epsilon \in (0,1)$ and recovering a known optimal result for $\epsilon = 1$. We also establish a lower bound of $\Omega(d^{\frac{2\epsilon}{1+\epsilon}}T^{\frac{1}{1+\epsilon}})$, which strictly improves upon the multi-armed bandit rate and highlights the hardness of heavy-tailed linear bandit problems. For finite action sets of size n, we derive upper and lower bounds of $\tilde{\mathcal{O}}(\sqrt{d}(\log n)^{\frac{\epsilon}{1+\epsilon}}T^{\frac{1}{1+\epsilon}})$ and $\tilde{\Omega}(d^{\frac{\epsilon}{1+\epsilon}}(\log n)^{\frac{\epsilon}{1+\epsilon}}T^{\frac{1}{1+\epsilon}})$, respectively. Finally, we provide action-set-dependent regret upper bounds, showing that (i) for some geometries, such as l_p-norm balls for $p \leq 1+\epsilon$, we can further reduce the dependence on d, and (ii) for RKHS functions with the Matérn kernel we can attain sublinear regret for all $\epsilon \in (0,1]$, thus substantially improving over the existing state-of-the-art.

1 Introduction

The stochastic linear bandit problem is a foundational setting of sequential decision-making under uncertainty, where the expected reward of each action is modeled as a linear function of known features. While most existing work assumes sub-Gaussian reward noise—enabling the use of concentration inequalities like Chernoff bounds—real-world noise often exhibits heavy tails, potentially with unbounded variance, violating these assumptions. Heavy-tailed noise naturally arises in diverse domains such as high-volatility asset returns in finance [CB00, Con01], conversion values in online advertising [CvdLG20, JSP21], cortical neural oscillations [RVHH15], and packet delays in communication networks [BTA02]. In such settings, reward distributions may be well-approximated by distributions such as Pareto, Student's t, or Weibull, all of which exhibit only polynomial tail decay.

2nd Workshop on Aligning Reinforcement Learning Experimentalists and Theorists (ARLET 2025).

The statistical literature has developed several robust estimation techniques for random variables with only bounded $(1+\epsilon)$-moments (for some $\epsilon \in (0,1]$), such as median-of-means estimators [DLLO16, LM19b] and Catoni M-estimators [Cat12, BJL15] in the univariate case, as well as robust least squares [AC11, HS14, HW19] and adaptive Huber regression [SZF20] for multivariate settings.

Robustness to heavy tails was first introduced into sequential decision-making by [BCBL13] in the context of multi-armed bandits. Subsequent work including [MY16, SYKL18, XWWZ20] extended these ideas to linear bandits, where each action is represented by a feature vector and the reward includes heavy-tailed noise. Generalizing robust estimators from the univariate to the multivariate setting is nontrivial, and many works have focused on designing such estimators and integrating them into familiar algorithmic frameworks like UCB. However, the relative unfamiliarity of heavy-tailed noise can make it difficult to judge the tightness of the regret bounds. As we discuss later, this has led to some degree of misinterpretation of existing lower bounds, with key problems prematurely considered "solved" despite persistent, unrecognized gaps.

1.1 Problem Statement

We consider the problem of stochastic linear bandits with an action set $\mathcal{A} \subseteq \mathbb{R}^d$ and an unknown parameter $\theta^\star \in \mathbb{R}^d$. At each round $t = 1, 2, \ldots, T$, the learner chooses an action $x_t \in \mathcal{A}$ and observes the reward

$$y_t \;=\; \langle x_t, \theta^\star \rangle \;+\; \eta_t,$$

where η_t are independent noise terms that satisfy $\mathbb{E}[\eta_t] = 0$ and $\mathbb{E}\big[|\eta_t|^{1+\epsilon}\big] \le \upsilon$ for some $\epsilon \in (0,1]$ and finite $\upsilon > 0$. We adopt the standard assumption that the expected rewards and parameters are bounded, namely, $\sup_{x\in\mathcal{A}} |\langle x, \theta^\star\rangle| \le 1$ and $\|\theta^\star\|_2 \le 1$. Letting $x^\star \in \arg\max_{x\in\mathcal{A}} \langle x, \theta^\star\rangle$ be an optimal action, the cumulative expected regret after T rounds is

$$R_T \;=\; \sum_{t=1}^{T} \big(\langle x^\star, \theta^\star\rangle - \langle x_t, \theta^\star\rangle\big).$$

Given $(\mathcal{A}, \epsilon, \upsilon)$, the objective is to design a policy for sequentially selecting the points (i.e., x_t for $t = 1, \ldots, T$) in order to minimize R_T.

1.2 Contributions

We study the minimax regret of stochastic linear bandits under heavy-tailed noise and make several contributions that clarify and advance the current state of the art. Although valid lower bounds exist, we show that they have been misinterpreted as matching known upper bounds. After correcting this misconception, we provide improved upper and lower bounds in the following ways:

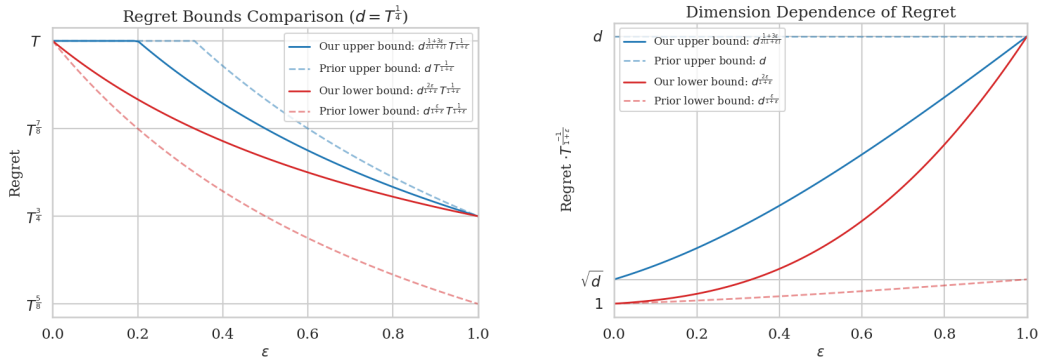
- **Novel estimator and analysis:** We introduce a new estimator inspired by [CJKS21] (who studied the finite-variance setting, $\epsilon = 1$), adapted to the heavy-tailed setting ($\epsilon \in (0,1]$). Its analysis leads to an experimental design problem that accounts for the geometry induced by the heavy-tailed noise, which is potentially of independent interest beyond linear bandits.
- **Improved upper bounds:** We use this estimator within a phased elimination algorithm to obtain state-of-the-art regret bounds for both finite- and infinite-arm settings. Additionally, we derive a geometry-dependent regret bound that emerges naturally from the estimator's experimental design.
- **Improved lower bounds:** We establish novel minimax lower bounds under heavy-tailed noise that are the first to reveal a dimension-dependent gap between multi-armed and linear bandit settings (e.g., when the arms lie on the unit sphere). We provide such results for both the finite-arm and infinite-arm settings.

Table 1 summarizes our quantitative improvements over prior work, while Figure 1 illustrates the degree of improvement obtained and what gaps still remain.

In addition to these results for heavy-tailed linear bandits, we show that our algorithm permits the kernel trick, and that this leads to regret bounds for the Matérn kernel (with heavy-tailed noise) that significantly improve on the best existing bounds, in particular being sublinear for all $\epsilon \in (0,1]$. See Section 3.1 for a summary, and Appendix C for the details.

Table 1: Comparison of regret bounds (in the $\widetilde{O}(\cdot)$ or $\widetilde{\Omega}(\cdot)$ sense) with heavy-tailed rewards for the model $y_t = \langle x_t, \theta_* \rangle + \eta_t$ where $\mathbb{E}[\eta_t] = 0$, $\mathbb{E}[|\eta_t|^{1+\epsilon}] \le 1$, $\|\theta\|_2 \le 1$, $|\langle x, \theta \rangle| \le 1$. The complexity measure $M(\mathcal{A})$ is defined in Theorem 3.

Paper	Setting	Regret Upper Bound	Regret Lower Bound
[SYKL18]	general	$dT^{\frac{1}{1+\epsilon}}$	$d^{\frac{\epsilon}{1+\epsilon}} T^{\frac{1}{1+\epsilon}}$ [1]
[HZWY23]	$\mathbb{E}[\lvert\eta_t\rvert^{1+\epsilon}] \le v_t^{1+\epsilon}$	$d\sqrt{\sum_{t=1}^{T} v_t^2} T^{\frac{1-\epsilon}{2+2\epsilon}}$	
[XWWZ20]	$\lvert\mathcal{A}\rvert = n$	$\sqrt{d \log n} T^{\frac{1}{1+\epsilon}}$	$d^{\frac{\epsilon}{1+\epsilon}} T^{\frac{1}{1+\epsilon}}$
[CG19]	Matérn(ν, d)	$T^{\frac{2+\epsilon}{2(1+\epsilon)} + \frac{d}{2\nu+d}}$	$T^{\frac{\nu+d\epsilon}{\nu(1+\epsilon)+d\epsilon}}$
[BCBL13]	MAB$(\mathcal{A} = \Delta^d)$	$d^{\frac{\epsilon}{1+\epsilon}} T^{\frac{1}{1+\epsilon}}$	$d^{\frac{\epsilon}{1+\epsilon}} T^{\frac{1}{1+\epsilon}}$
Our Work	$\mathcal{A}$-dependent	$M(\mathcal{A})^{\frac{1}{1+\epsilon}} \min(d, \log \lvert\mathcal{A}\rvert)^{\frac{\epsilon}{1+\epsilon}} T^{\frac{1}{1+\epsilon}}$ (Theorem 3)	
	general	$d^{\frac{1+3\epsilon}{2(1+\epsilon)}} T^{\frac{1}{1+\epsilon}}$ (Corollary 1)	$d^{\frac{2\epsilon}{1+\epsilon}} T^{\frac{1}{1+\epsilon}}$ (Theorem 1)
	$\lvert\mathcal{A}\rvert = n$	$\sqrt{d}(\log n)^{\frac{\epsilon}{1+\epsilon}} T^{\frac{1}{1+\epsilon}}$ (Corollary 1)	$d^{\frac{\epsilon}{1+\epsilon}} (\log n)^{\frac{\epsilon}{1+\epsilon}} T^{\frac{1}{1+\epsilon}}$ (Theorem 2)
	Matérn(ν, d)	$T^{1-\frac{\epsilon}{1+\epsilon}\frac{2\nu}{2\nu+d}}$ (Corollary 4)	
	MAB$(\mathcal{A} = \Delta^d)$	$d^{\frac{\epsilon}{1+\epsilon}} T^{\frac{1}{1+\epsilon}}$ (Corollary 2)	



(a) Regret bounds comparison

(b) Dimension-dependence comparison

Figure 1: (a) Comparison of regret bounds across ϵ for $T = d^4$. (b) Scaling of the bounds in d.

1.3 Related Work

The first systematic study of heavy-tailed noise in bandits is due to [BCBL13], who replaced the empirical mean in UCB with robust mean estimators, and obtained a regret bound of $\widetilde{O}\big(n^{\frac{\epsilon}{1+\epsilon}} T^{1/(1+\epsilon)}\big)$ with n arms, along with a matching lower bound. A sequence of follow-up works [YNKL18, LWHZ19, LYLO20, WS21, HDH22, CHDH25] refined these ideas and extended them to best-arm identification, adversarial, parameter-free, and Lipschitz settings. The first extension of heavy-tailed analysis from MAB to linear bandits is due to [MY16], who proposed truncation- and MoM-based algorithms and proved an $\widetilde{O}\big(d\,T^{\frac{2+\epsilon}{2(1+\epsilon)}}\big)$ regret bound. Subsequently, [SYKL18, XWWZ20] improved the regret bounds for infinite and finite action sets, respectively (see Table 1). Huber-loss based estimators have emerged as another robustification strategy, for which [LS24, KK23, HZWY23, WZZZ25] provided moment-aware regret bounds. [ZHYW21] suggested median based estimators for symmetric error distributions without any bounded moments (e.g., Cauchy). Beyond linear bandits, [XWW+23] proved a similar $dT^{\frac{1}{1+\epsilon}}$ bound for generalized linear bandits, and [CG19] studied heavy-tailed kernel-based bandits, which we will cover in more detail in Appendix C. A summary of the best regret bounds of previous work and ours can be found in Table 1.

[1] We refer to this as the multi-armed bandit (MAB) rate because it matches that of a MAB problem with d arms. Note that that the $dT^{\frac{1}{1+\epsilon}}$ lower bound from [SYKL18] was only proved for an instance with $\mathbb{E}[|\eta_t|^{1+\epsilon}] = O(d)$ rather than $O(1)$; see Section 2 for further discussion.

2 Lower Bounds

Before describing our own lower bounds, we take a moment to clarify the state of lower bounds that exist in the literature, as there has been some apparent misinterpretation within the community. The regret lower bound construction presented in [SYKL18] leverages the reward distribution

$$y(x) = \begin{cases} (\frac{1}{\Delta})^{\frac{1}{\epsilon}} & \text{w.p. } \Delta^{\frac{1}{\epsilon}}\theta^\top x \\ 0 & \text{w.p. } 1 - \Delta^{\frac{1}{\epsilon}}\theta^\top x \end{cases}$$

under the choice $\Delta = \frac{1}{12}T^{-\frac{\epsilon}{1+\epsilon}}$, and with choices of θ and $\mathcal{A}$ that ensure $d\Delta \leq \theta^\top x \leq 2d\Delta$. A straightforward calculation shows that the reward distributions of this construction possesses a $(1+\epsilon)$-absolute moment of $\Delta^{-1}(\theta^\top x) \geq d$ for all actions. Recall that in our problem statement we consider the $(1+\epsilon)$-absolute moment to be a constant (that does not depend on the the dimension d or time horizon T). We can compare this with the canonical case of sub-Gaussian noise ($\epsilon = 1$) where it is assumed that the second moment is bounded by $\sigma^2 = \Omega(1)$, in which it is well-known that the optimal regret rate is on the order of $\sigma d\sqrt{T}$ [LS20]. If we were to set $\sigma^2 = \Theta(d)$, this would suggest a rate of $d^{3/2}\sqrt{T}$, but this only exceeds the usual $d\sqrt{T}$ because σ is artificially large. We stress that we are not claiming that the lower bound of [SYKL18] is in any way incorrect, and the authors even acknowledge that the bound on the moment scales with the dimension in the appendix of their work. We are simply pointing out that there has been some misinterpretation of the lower bound within the community.[2]

If we adjust the expected reward distributions such that $\Delta \leq \theta^\top x \leq 2\Delta$, so that the reward distribution maintains a constant $1+\epsilon$ absolute moment, the resulting regret lower bound turns out to scale as $d^{\frac{\epsilon}{1+\epsilon}}T^{\frac{1}{1+\epsilon}}$,[3] matching the known optimal lower bound for the Multi-Armed Bandit (MAB) setting with d arms. However, with a more precise analysis, we can prove a stronger lower bound on a similar instance (with modified parameters) having a constant $(1+\epsilon)$-central moment of rewards, as we will see below.

2.1 Infinite Arm Set

Given the context above, we are ready to present our own lower bound that builds on the construction introduced by [SYKL18] but is specifically tailored to improving the d dependence.

Theorem 1. *Fix the action set $\mathcal{A} = \{x \in [0,1]^{2d} : x_{2i-1} + x_{2i} = 1 \quad \forall i \in [d]\}$. There exists a reward distribution with a $(1+\epsilon)$-central moment bounded by 1 and a $\theta^* \in \mathbb{R}^{2d}$ with $\|\theta^*\|_2 \leq 1$ and $\sup_{x\in\mathcal{A}} |x^\top\theta^*| \leq 1$, such that for $T \geq 4^{\frac{1+\epsilon}{\epsilon}}d^2$, the regret incurred is $\Omega(d^{\frac{2\epsilon}{1+\epsilon}}T^{\frac{1}{1+\epsilon}})$.*

Proof. For a parameter $\Delta \leq \frac{1}{4d}$ to be specified later, we let the reward distribution be a Bernoulli random variable defined as follows:

$$y(x) = \begin{cases} (\frac{1}{\gamma})^{\frac{1}{\epsilon}} & \text{w.p. } \gamma^{\frac{1}{\epsilon}}\theta^\top x \\ 0 & \text{w.p. } 1 - \gamma^{\frac{1}{\epsilon}}\theta^\top x \end{cases}$$

with $\gamma := 2d\Delta$. We consider parameter vectors θ lying in the set $\Theta := \left\{\theta \in \{\Delta, 2\Delta\}^{2d} : \theta_{2i-1} + \theta_{2i} = 3\Delta\right\}$, from which the assumption $\Delta \leq \frac{1}{4d}$ readily implies $\|\theta\|_2 \leq 1$ and $\sup_{x\in\mathcal{A}} |x^\top\theta^*| \leq 1$. For any $\theta \in \Theta$, the $(1+\epsilon)$-raw moment of the reward distribution (and therefore the central moment, since the rewards are nonnegative) for each action is bounded by $\mathbb{E}[|y(x)|^{1+\epsilon}|x] = \gamma^{-(1+\epsilon)/\epsilon}\gamma^{1/\epsilon}\theta^\top x = \gamma^{-1}\theta^\top x \leq 1$, since $\gamma = 2d\Delta$ and $\theta^\top x \leq 2d\Delta$.

Let $R_T(\mathcal{A}, \theta)$ be the cumulative regret for arm set $\mathcal{A}$ and parameter θ, and let $\text{ind}_i(\theta) := \arg\max_{b\in\{0,1\}}(\theta_{2i-1+b})$ for $\theta \in \Theta$, and write $x_t = (x_{t,1}, \ldots, x_{t,d})$. We have

$$R_T(\mathcal{A}, \theta) = \sum_{t=1}^{T}\sum_{i=1}^{d}\left(\Delta - \Delta x_{t,2i-1+\text{ind}_i(\theta)}\right) = \Delta\sum_{t=1}^{T}\sum_{i=1}^{d}\left(\frac{1}{2} - \frac{1}{2}(-1)^{\text{ind}_i(\theta)}(x_{t,2i-1} - x_{t,2i})\right)$$

[2] Previous works that indicate the minimax optimality of this bound (with respect to T and d) include [XWWZ20, XWW+23, HZWY23, WZZZ25].

[3] This is obtained by optimizing Δ for the adjusted regret $\Delta T(\frac{1}{4} - \frac{3}{2}\sqrt{d^{-1}\Delta^{\frac{1+\epsilon}{\epsilon}}T})$

$$\geq \frac{\Delta}{2}\sum_{i=1}^{d}\mathbb{E}_\theta\left[\sum_{t=1}^{T}\mathbb{I}\{(-1)^{\mathrm{ind}_i(\theta)}(x_{t,2i-1}-x_{t,2i})\leq 0\}\right]$$
$$\geq \frac{\Delta T}{4}\sum_{i=1}^{d}\mathbb{P}_\theta\left[\sum_{t=1}^{T}\mathbb{I}\{(-1)^{\mathrm{ind}_i(\theta)}(x_{t,2i-1}-x_{t,2i})\leq 0\}\geq \frac{T}{2}\right],$$

where the second equality follows by using $x_{t,2i-1}+x_{t,2i}=1$ and checking the cases $\mathrm{ind}_i(\theta)=0$ and $\mathrm{ind}_i(\theta)=1$ separately.

For any $\theta\in\Theta, i\in[d]$, we define $\theta'\in\Theta$ with entries $\theta'_j=\begin{cases}3\Delta-\theta_j & 2i-1\leq j\leq 2i\\ \theta_j & \text{otherwise}\end{cases}$, and let $p_{\theta,i}:=\mathbb{P}_\theta\left[\sum_{t=1}^{T}\mathbb{I}\{(-1)^{\mathrm{ind}_i(\theta)}(x_{t,2i-1}-x_{t,2i})\leq 0\}\geq\frac{T}{2}\right]$. We then have the following:

$$p_{\theta,i}+p_{\theta',i}\geq\frac{1}{2}\exp(-\mathrm{KL}(\mathbb{P}_\theta\|\mathbb{P}_{\theta'}))\qquad\text{(Bretagnolle–Huber inequality)}$$
$$=\frac{1}{2}\exp\left(-\mathbb{E}_\theta\left[\sum_{t=1}^{T}\mathrm{KL}\left(\mathrm{Ber}(\gamma^{\frac{1}{\epsilon}}\theta^\top x_t)\|\mathrm{Ber}(\gamma^{\frac{1}{\epsilon}}\theta'^\top x_t)\right)\right]\right).\qquad\text{(Chain rule)}$$

Now we set $\Delta:=\frac{1}{2}d^{\frac{\epsilon-1}{1+\epsilon}}T^{\frac{-\epsilon}{1+\epsilon}}$. Note that since $T\geq 4^{\frac{1+\epsilon}{\epsilon}}d^2$, the above-mentioned condition $\Delta\leq\frac{1}{4d}$ holds, ensuring the Bernoulli parameter is in $[0,1]$. Under this choice of Δ, we have

$$\mathrm{KL}\left(\mathrm{Ber}(\gamma^{\frac{1}{\epsilon}}\theta^\top x_t)\|\mathrm{Ber}(\gamma^{\frac{1}{\epsilon}}\theta'^\top x_t)\right)\leq\frac{2^{\frac{2}{\epsilon}}4\Delta^{\frac{2}{\epsilon}}d^{\frac{2}{\epsilon}}\Delta^2}{2^{\frac{1}{\epsilon}}\Delta^{\frac{1+\epsilon}{\epsilon}}d^{\frac{1+\epsilon}{\epsilon}}\cdot\frac{1}{2}}=2^{\frac{1}{\epsilon}}8\Delta^{\frac{1+\epsilon}{\epsilon}}d^{\frac{1-\epsilon}{\epsilon}}=4T^{-1},$$

where in the first inequality we used $\mathrm{KL}(\mathrm{Ber}(p)\|\mathrm{Ber}(q))\leq\frac{(p-q)^2}{q(1-q)}$; we get $|p-q|\leq 2\gamma^{\frac{1}{\epsilon}}\Delta=2(2d\Delta)^{\frac{1}{\epsilon}}\Delta$ because θ and θ' differ only via a single swap of $(\Delta,2\Delta)$ by $(2\Delta,\Delta)$, $q\geq\gamma^{\frac{1}{\epsilon}}\Delta d=(2d\Delta)^{\frac{1}{\epsilon}}\Delta d$ by construction, and $1-q\geq 1-\gamma^{\frac{1}{\epsilon}}2d\Delta\geq\frac{1}{2}$ via $\Delta\leq\frac{1}{4d}$.

Combining the preceding display equations gives $p_{\theta,i}+p_{\theta',i}\geq\frac{1}{2}\exp(-4)$, and averaging over all (θ,θ') (with $\theta'\neq\theta$) and summing over i, we obtain $\frac{1}{|\Theta|}\sum_{\theta\in\Theta}\sum_{i=1}^{d}p_{\theta,i}\geq\frac{1}{4}d\exp(-4)$. Hence, there exists $\theta^*\in\Theta$ such that $\sum_{i=1}^{d}p_{\theta^*,i}\geq\frac{1}{4}d\exp(-4)$, and substituting into our earlier lower bound on R_T gives $R_T(\mathcal{A},\theta^*)\geq\frac{1}{16}\exp(-4)\Delta dT=\frac{1}{32}\exp(-4)d^{\frac{2\epsilon}{1+\epsilon}}T^{\frac{1}{1+\epsilon}}$. □

The setting in Theorem 1 is not the only one that gives regret $\Omega(d^{\frac{2\epsilon}{1+\epsilon}}T^{\frac{1}{1+\epsilon}})$. In fact, the same lower bound turns out to hold for the unit ball action set with a slight change in reward distribution to avoid large KL divergences when $\theta^\top x$ is small. The details are given in Appendix B.

2.2 Finite Arm Set

The best known lower bound for finite arm sets matches the MAB lower bound of $d^{\frac{\epsilon}{1+\epsilon}}T^{\frac{1}{1+\epsilon}}$ with d arms (see [XWWZ20] and the summary in Table 1). We provide the first n-dependent lower bound (where $n:=|\mathcal{A}|$) by combining ideas from the MAB lower bound construction for m arms with the construction used in Theorem 1 for dimension $\frac{d}{m}$, where $m^{\frac{d}{m}}\approx n$. When $n=2^{\mathcal{O}(d)}$ or $n=T^{\mathcal{O}(d)}$, which arises naturally when finely quantizing in each dimension, our lower bound matches the infinite arm case (in the $\widetilde{\Omega}(\cdot)$ sense) as one might expect.

Theorem 2. *For each $n\in[d,2^{\lfloor\frac{d}{4}\rfloor}]$, there exists an action set $\mathcal{A}$ with $|\mathcal{A}|\leq n$, a reward distribution with a $(1+\epsilon)$-central moment bounded by 1, and a $\theta^*\in\mathbb{R}^d$ with $\|\theta^*\|_2\leq 1$ and $\sup_{x\in\mathcal{A}}|x^\top\theta^*|\leq 1$, such that for $T\geq 4^{\frac{1+\epsilon}{\epsilon}}d^{\frac{1+\epsilon}{\epsilon}}$, the regret incurred is $\Omega\big(T^{\frac{1}{1+\epsilon}}d^{\frac{\epsilon}{1+\epsilon}}\big(\frac{\log n}{\log d}\big)^{\frac{\epsilon}{1+\epsilon}}\big)$.*

Proof Sketch. We outline the main proof steps here, and defer the full details to Appendix E.

Consider $\log(\cdot)$ with base 2, and define m to be the smallest integer such that $\frac{m}{\log m}\geq\frac{d}{\log n}$. From the assumption $n\in[d,2^{\lfloor\frac{d}{4}\rfloor}]$ we can readily verify that $d>4$ and $m\in[4,d]$. For convenience, we assume that d is a multiple of m, since otherwise we can form the construction of the lower bound

with $d' = d - (d \bmod m)$ and pad the action vectors with zeros. Letting $d_i := (i-1)m$, we define the action set and the parameter set as follows for some $\Delta \le \frac{1}{4d}$ to be specified later:

$$\mathcal{A} := \left\{ a \in \{0,1\}^d : \sum_{j=d_i+1}^{d_{i+1}} a_j = 1, \ \ \forall i \in [d/m] \right\}$$

$$\theta^* \in \Theta := \left\{ \theta \in \{\Delta, 2\Delta\}^d : \sum_{j=d_i+1}^{d_{i+1}} \theta_j = (m+1)\Delta, \ \ \forall i \in [d/m] \right\}.$$

In simple terms, the d-dimensional vectors are arranged in d/m groups of size m; each block in $a \in \mathcal{A}$ has a single entry of 1 (with 0 elsewhere), and each block in θ^* has a single entry of 2Δ (with Δ elsewhere). The condition $\Delta \le \frac{1}{4d}$ readily implies $\|\theta^*\|_2 \le 1$ and $x^\top \theta^* \le 1$ as required. Moreover, we have $|\mathcal{A}| = m^{\frac{d}{m}}$, and thus $\log |\mathcal{A}| = \frac{d}{m} \log m \le \log n$ by the definition of m.

Similar to Theorem 1, we let the rewards distribution be $y(x) = \begin{cases} (\frac{1}{\gamma})^{\frac{1}{\epsilon}} & \text{w.p. } \gamma^{\frac{1}{\epsilon}} \theta^\top x \\ 0 & \text{w.p. } 1 - \gamma^{\frac{1}{\epsilon}} \theta^\top x \end{cases}$ with $\gamma := 2\frac{d}{m}\Delta$. The choices of $\mathcal{A}$ and Θ give $\theta^\top x \le 2\Delta \frac{d}{m}$, so by the same reasoning as in Theorem 1, the $(1+\epsilon)$-moment of the reward distribution is bounded by 1.

Let $\text{ind}_i(x) := \arg\max_{b \in [m]} (x_{d_i+b})$ for any $x \in \mathcal{A} \cup \Theta$, and define $T_{i,b} := |\{t : x_{t,d_i+b} = 1\}|$. Moreover, define t_{U} to be a random integer drawn uniformly from $[T]$, which immediately implies that $\mathbb{P}_\theta[x_{t_{\text{U}},d_i+b} = 1] = \frac{\mathbb{E}_\theta[T_{i,b}]}{T}$. Then we can rewrite regret as $R_T(\mathcal{A}, \theta) = \Delta T \sum_{i=1}^{d/m} \big(1 - \mathbb{P}_\theta[x_{t_{\text{U}},d_i+\text{ind}_i(\theta)} = 1]\big)$. For any $\theta \in \Theta$ and $i \in [\frac{d}{m}]$, and any $b \in [m]$, we define $\theta^{(b)} \in \Theta$ to have entries given by $\theta_j^{(b)} = \begin{cases} \Delta + \Delta \mathbb{I}\{j = d_i + b\} & j \in [d_i+1, d_{i+1}] \\ \theta_j & \text{otherwise} \end{cases}$; and define the base parameter $\theta^{(0)}$ with entries $\theta_j^{(0)} = \begin{cases} \Delta & j \in [d_i+1, d_{i+1}] \\ \theta_j & \text{otherwise} \end{cases}$. Note that $\theta^{(\text{ind}_i(\theta))} = \theta$. Moreover, similar to Theorem 1, the KL divergence of reward distribution of $\theta^{(0)}$ and $\theta^{(b)}$ problem instances at action x_t can be bounded by $2^{\frac{1+\epsilon}{\epsilon}} \Delta^{\frac{1+\epsilon}{\epsilon}} \left(\frac{d}{m}\right)^{\frac{1-\epsilon}{\epsilon}} \mathbb{I}\{x_{t,d_i+b} = 1\}$. We set $\Delta := \frac{1}{8} \left(\frac{d}{m}\right)^{\frac{\epsilon-1}{1+\epsilon}} \left(\frac{T}{m}\right)^{\frac{-\epsilon}{1+\epsilon}}$, from which the condition $T \ge 4^{\frac{1+\epsilon}{\epsilon}} d^{\frac{1+\epsilon}{\epsilon}}$ readily yields $\gamma^{\frac{1}{\epsilon}} (\frac{2d}{m}\Delta) \le \frac{1}{2}$.

Next, using Pinsker's inequality along with averaging over $b \in m$, we can show that $\frac{1}{m} \sum_b \mathbb{P}_{\theta^{(b)}}[x_{t,d_i+b} = 1] \le \frac{1}{m} + \frac{1}{2}$. Averaging over all $\theta \in \Theta$, summing over $i \in [d/m]$, and recalling that $m \ge 4$, we obtain

$$\frac{1}{|\Theta|} \sum_{\theta \in \Theta} \sum_{i=1}^{d/m} \big(1 - \mathbb{P}_\theta[x_{t,d_i+\text{ind}_i(\theta)} = 1]\big) \ge \frac{d}{m}\Big(1 - \frac{1}{m} - \frac{1}{2}\Big) \ge \frac{d}{4m}.$$

Hence, there exists $\theta^* \in \Theta$ such that $\sum_{i=1}^{d/m} \big(1 - \mathbb{P}_{\theta^*}[x_{t,d_i+\text{ind}_i(\theta^*)} = 1]\big) \ge \frac{d}{4m}$, substituting into our earlier lower bound on R_T along with our choice of Δ, we obtain

$$R_T(\mathcal{A}, \theta^*) \ge \frac{d}{4m} \Delta T = \frac{1}{32} d^{\frac{\epsilon}{1+\epsilon}} \left(\frac{d}{m}\right)^{\frac{\epsilon}{1+\epsilon}} T^{\frac{1}{1+\epsilon}}.$$

Since $f(x) = \frac{x}{\log x}$ is increasing for $x \ge e$, and $m \in [4, d]$, the definition of m gives the $\frac{d}{\log n} > \frac{m-1}{\log(m-1)} > \frac{m-1}{\log m} \ge \frac{m-1}{\log d}$. Rearranging the terms and bounding m, we obtain $\frac{d}{m} \ge \frac{\log n}{2 \log d}$, which gives the desired result. □

3 Proposed Algorithm and Upper Bounds

In this section, we propose a phased elimination–style algorithm called *MED-PE* that achieves the best known minimax regret upper bound for linear bandits with noise that has bounded $(1+\epsilon)$-moments. In each phase ℓ, the algorithm operates as follows:

1. Design a sampling distribution over the currently active arms that minimizes the $(1+\epsilon)$-absolute moment of a certain estimator of θ^* in the worst-case direction among all active arms (see Lemma 1), along with a suitable regularization term.

Algorithm 1 Moment-based Experimental Design Phased Elimination (*MED-PE*)

Input: $\mathcal{A}$, $\gamma > 0$,$\beta \geq 0$, $\epsilon \in (0, 1]$, υ, T, robust mean estimator $\widehat{\mu}(S, \delta)$
Initialization $\ell \leftarrow 1$, $t \leftarrow 0$, $\mathcal{A}_1 \leftarrow \mathcal{A}$
while $t < T$ *and* $|\mathcal{A}_\ell| > 1$ **do**

$$M_{1+\epsilon}(\lambda; \mathcal{A}_\ell, \gamma, \beta) \leftarrow \max_{a \in \mathcal{A}_\ell} \mathbb{E}_{x\sim\lambda}\left[\left|a^\top A^{(\gamma)}(\lambda)^{-1} x\right|^{1+\epsilon}\right] + \beta^{1+\epsilon}\|a\|^{1+\epsilon}_{A^{(\gamma)}(\lambda)^{-1}}$$

$$(A^{(\gamma)}(\lambda) := \gamma I + \mathbb{E}_{x\sim\lambda}[xx^\top])$$

$$\lambda^*_\ell \leftarrow \operatorname*{argmin}_{\lambda \in \Delta_{\mathcal{A}_\ell}} M_{1+\epsilon}(\lambda; \mathcal{A}_\ell, \gamma, \beta)$$

$\varepsilon_\ell \leftarrow 2^{-\ell}$, $\tau_\ell \leftarrow 32^{\frac{1+\epsilon}{\epsilon}}(1+\upsilon)^{\frac{1}{\epsilon}}\varepsilon_\ell^{-\frac{1+\epsilon}{\epsilon}} M_{1+\epsilon}(\lambda^*_\ell; \mathcal{A}_\ell, \gamma, \beta)^{\frac{1}{\epsilon}} \log(2\ell^2|\mathcal{A}_\ell|T)$
for $s \leftarrow 1$ **to** τ_ℓ **do**
  Draw $x_s \sim \lambda^*_\ell$, observe reward y_s
$W^{(a)} \leftarrow \widehat{\mu}\left(\{a^\top A^{(\gamma)}(\lambda^*_\ell)^{-1} x_s\, y_s\}_{s=1}^{\tau_\ell}, \dfrac{1}{2\ell^2 T|\mathcal{A}_\ell|}\right)$
$\widehat{\theta}_\ell \leftarrow \arg\min_\theta \max_{a \in \mathcal{A}_\ell} |\theta^\top a - W^{(a)}|$
$\mathcal{A}_{\ell+1} \leftarrow \left\{ a \in \mathcal{A}_\ell : \widehat{\theta}_\ell^\top a \geq \max_{a' \in \mathcal{A}_\ell} \widehat{\theta}_\ell^\top a' \; - \; 4\varepsilon_\ell \right\}$, $\ell \leftarrow \ell + 1$, $t \leftarrow t + \tau_\ell$

2. Pull a budgeted number of samples (scaled by $2^{\ell \cdot \frac{1+\epsilon}{\epsilon}}$) from that distribution, and estimate the reward for each active arm separately using a robust mean estimator.
3. Fit a parameter $\widehat{\theta}$ that minimizes the maximum distance of $\widehat{\theta}^\top a$ to the estimated reward of a over all active arms.
4. Eliminate suboptimal arms from the active set.

This process is repeated with progressively tighter accuracy until the time horizon is reached or a single arm remains. In the latter case, the remaining arm is pulled for all remaining rounds.

MED-PE is a generalization of Robust Inverse Propensity Score estimator in [CJKS21] which assumes a bounded variance for the rewards. We first find an experimental design that minimizes the $(1+\epsilon)$-absolute moment of $a^\top A^{(\gamma)}(\lambda)^{-1}x$, with suitable regularization, for all a that are active (and therefore the confidence interval of the robust estimator). Note that $A^{(\gamma)}(\lambda)^{-1}x_s y_s$ (with $A^{(\gamma)}(\lambda) := \gamma I + \mathbb{E}_{x\sim\lambda}[xx^\top]$) can be interpreted as a single-sample regularized least squares estimator, which is then robustified through a robust mean estimation subroutine $\widehat{\mu}$ for each arm. The overall accuracy guarantee of this estimator turns out to depend directly on $M_{1+\epsilon}(\lambda; \mathcal{A}, \gamma, \beta)$ (see Lemma 1 below), which is why we seek to minimize this quantity in our design λ^*_ℓ. Moreover, we include a regularization term for design optimization to mitigate the estimator's bias, as $A^{(\gamma)}(\lambda)^{-1}x_s y_s$ is biased for $\gamma \neq 0$.

Any robust mean estimator such as truncated (trimmed) mean, median-of-means, or Catoni's M estimator [LM19a, Cat12], can be used as the subroutine $\widehat{\mu}$ of *MED-PE* . We adopt the truncated mean for concreteness and simplicity. The following lemma provides our main confidence interval for our regression estimator.

Lemma 1. *Consider $(x_i, y_i)_{i=1}^n$, where $x_i \sim \lambda(\mathcal{A})$ are i.i.d. vectors from distribution λ over $\mathcal{A}$, and suppose that $y_i = \langle \theta^*, x_i \rangle + \eta_i$, where η_i are independent zero-mean noise terms such that $\mathbb{E}[|\eta_i|^{1+\epsilon}] \leq \upsilon$, and $\max_{a\in\mathcal{A}} |\langle \theta^*, a \rangle| \leq 1$. The estimator $\widehat{\theta}(\gamma)$ with a robust mean estimator $\widehat{\mu}$ as a subroutine is defined as follows:*

$$\widehat{\theta}(\gamma) := \arg\min_\theta \max_{a\in\mathcal{A}} \left| \theta^\top a - \widehat{\mu}\left(\{a^\top A^{(\gamma)}(\lambda)^{-1} x_i\, y_i\}_{i=1}^n, \frac{\delta}{|\mathcal{A}|}\right)\right|,$$

where $A^{(\gamma)}(\lambda) := \gamma I + \mathbb{E}_{x\sim\lambda}[xx^\top]$. For any $\beta \geq 0$, $\widehat{\theta}(\gamma)$ with the truncated empirical mean $\widehat{\mu}(\{X_i\}_{i=1}^n, \delta) := \frac{1}{n}\sum X_i \mathbb{I}\{|X_i| \leq \left(\frac{\upsilon t}{\log(\delta^{-1})}\right)^{\frac{1}{1+\epsilon}}\}$ as a subroutine, satisfies the following with

probability at least $1-\delta$*:*

$$\max_{a\in\mathcal{A}} |\langle \widehat{\theta}-\theta^*, a\rangle| \le \left(2\gamma^{1/2}\|\theta^*\|_2\beta^{-1} + 32(1+\upsilon)^{\frac{1}{1+\epsilon}}\left(\frac{\log(|\mathcal{A}|/\delta)}{n}\right)^{\frac{\epsilon}{1+\epsilon}}\right) M_{1+\epsilon}(\lambda;\mathcal{A},\gamma,\beta)^{\frac{1}{1+\epsilon}},$$

where $M_{1+\epsilon}(\lambda;\mathcal{A},\gamma,\beta) := \max_{a\in\mathcal{A}} \mathbb{E}_{x\sim\lambda}\left[\left|a^\top A^{(\gamma)}(\lambda)^{-1}x\right|^{1+\epsilon}\right] + \beta^{1+\epsilon}\|a\|^{1+\epsilon}_{A^{(\gamma)}(\lambda)^{-1}}$.

Proof Sketch. In order to use the robust mean estimator guaranties, we bound the $(1+\epsilon)$-absolute moment of our samples $a^\top A^{(\gamma)}(\lambda)^{-1}x\,y$ for $x\sim\lambda$. Using the boundedness of the expected rewards and the $(1+\epsilon)$-absolute moment of the noise η, we show that the moment is bounded by $4(1+\upsilon)M_{1+\epsilon}(\lambda;\mathcal{A},\gamma,\beta)$. Moreover, the expected reward estimator for arm a (denoted by $W^{(a)}$) is biased if $\gamma>0$, and we can bound the bias as follows:

$$\left|\langle\theta^*,a\rangle - \mathbb{E}[W^{(a)}]\right| \le \sqrt{\gamma}\beta^{-1}\|\theta^*\|_2 M_{1+\epsilon}(\lambda;\mathcal{A},\gamma,\beta)^{\frac{1}{1+\epsilon}}.$$

Using the triangle inequality and the union bound then gives the desired result. The detailed proof is given in Appendix A. □

The following theorem states our general action set dependent regret bound for *MED-PE*.

Theorem 3. *For any linear bandit problem with finite action set* $\mathcal{A}\subseteq\mathbb{R}^d$*, define*

$$M^*_{1+\epsilon}(\mathcal{A},\gamma,\beta) := \max_{\mathcal{V}\subseteq\mathcal{A}}\min_{\lambda\in\Delta^{\mathcal{V}}} M_{1+\epsilon}(\lambda;\mathcal{V},\gamma,\beta).$$

If $\mathbb{E}[|\eta_t|^{1+\epsilon}]\le\upsilon$*,* $\|\theta^*\|_2\le b$*, and* $\sup_{x\in\mathcal{A}}|a^\top\theta^*|\le 1$*, then MED-PE with the truncated empirical mean estimator (Lemma 1) and* $\gamma = T^{-\frac{2\epsilon}{1+\epsilon}}$ *achieves regret bounded by*

$$R_T \le \left(C_0\beta^{-1}b + C_1(1+\upsilon)^{\frac{1}{1+\epsilon}}\log(|\mathcal{A}|T\log^2 T)^{\frac{\epsilon}{1+\epsilon}}\right) M^*_{1+\epsilon}(\mathcal{A},T^{\frac{-2\epsilon}{1+\epsilon}},\beta)^{\frac{1}{1+\epsilon}}T^{\frac{1}{1+\epsilon}}$$

for some constants C_0 *and* C_1*.*

Proof Sketch. Using Lemma 1, with probability at least $1-(2\ell^2T)^{-1}$, we have

$$\max_{a\in\mathcal{A}_\ell}|a^\top\theta^* - a^\top\widehat{\theta}_\ell| \le \epsilon_\ell + 2\gamma^{1/2}b\beta^{-1}M^*_{1+\epsilon}(\mathcal{A},\gamma,\beta)^{\frac{1}{1+\epsilon}}.$$

Therefore, in the phases where ϵ_ℓ is large compared to $\gamma^{1/2}\beta^{-1}M^*_{1+\epsilon}(\mathcal{A},\gamma,\beta)^{\frac{1}{1+\epsilon}}$, suboptimal arms are eliminated, and no optimal arm is eliminated with high probability. In the phases where ϵ_ℓ is smaller, each arm pull incurs regret $\widetilde{\mathcal{O}}(\gamma^{1/2}\beta^{-1}M^*_{1+\epsilon}(\mathcal{A},\gamma,\beta)^{\frac{1}{1+\epsilon}})$. Setting $\gamma = T^{\frac{-2\epsilon}{1+\epsilon}}$, balances the two regret terms, and leads to the final regret bound. The detailed proof is given in Appendix A. □

Remark 1. *If* $\mathcal{A}$ *is not finite, we can cover the domain with* $T^{O(d)}$ *elements in* $\mathcal{A}$*, such that the expected reward of each arm can be approximated by one of the covered elements with* T^{-1} *error, and therefore the bound of Theorem 3 can be written as*

$$R_T \le \left(C_0\beta^{-1}b + C_1'(1+\upsilon)^{\frac{1}{1+\epsilon}}d^{\frac{\epsilon}{1+\epsilon}}\log(T^2\log^2 T)^{\frac{\epsilon}{1+\epsilon}}\right) M^*_{1+\epsilon}(\mathcal{A},T^{\frac{-2\epsilon}{1+\epsilon}},\beta)^{\frac{1}{1+\epsilon}}T^{\frac{1}{1+\epsilon}}.$$

The quantity $M^*_{1+\epsilon}$ in Theorem 3 may be difficult to characterize precisely in general, but the following lemma gives a universal upper bound.

Lemma 2. *For any action set* $\mathcal{A}$ *and* $\epsilon\in(0,1]$*, setting* $\gamma = T^{\frac{-2\epsilon}{1+\epsilon}}$ *and* $\beta=1$*, we have*

$$M^*_{1+\epsilon}(\mathcal{A},T^{-\frac{2\epsilon}{1+\epsilon}},1)\le d^{\frac{1+\epsilon}{2}}.$$

Moreover, a design λ *with* $M_{1+\epsilon}(\lambda;\mathcal{A},T^{\frac{-2\epsilon}{1+\epsilon}},1) = O(d^{\frac{1+\epsilon}{2}})$ *can be found with* $O(d\log\log d)$ *time.*

Proof. We upper bound the first term in the objective function as follows:

$$\begin{aligned}\mathbb{E}\left[\left|a^\top A^{(\gamma)}(\lambda)^{-1}x\right|^{1+\epsilon}\right] &\le \mathbb{E}\left[\left|a^\top A^{(\gamma)}(\lambda)^{-1}x\right|^2\right]^{\frac{1+\epsilon}{2}} && \text{(Jensen's inequality)}\\ &= \mathbb{E}\left[a^\top A^{(\gamma)}(\lambda)^{-1}xx^\top A^{(\gamma)}(\lambda)^{-1}a\right]^{\frac{1+\epsilon}{2}}\end{aligned}$$

$$\leq \|a\|^{1+\epsilon}_{A^{(\gamma)}(\lambda)^{-1}}. \qquad (\mathbb{E}[xx^\top] = \textstyle\sum_x \lambda(x)xx^\top \preceq A^{(\gamma)}(\lambda))$$

Hence, the minimization of $M_{1+\epsilon}$ is upper bounded in terms of a minimization of $\max_a \|a\|_{A(\lambda)^{-1}}$. This is equivalent to G-optimal design which is well-studied and the following is known (e.g., see [LS20, Chapter 21]): (i) The problem is convex and its optimal value is at most $\sqrt{d}$; (ii) There are efficient algorithms such as Frank–Wolfe that can find a design having $\max_a \|a\|_{A(\lambda)^{-1}} = O(\sqrt{d})$ with $O(d \log\log d)$ iterations. □

Combining Theorem 3 and Lemma 2, we obtain the following.

Corollary 1. *For any action set $\mathcal{A}$, MED-PE achieves regret $\widetilde{\mathcal{O}}(d^{\frac{1+3\epsilon}{2(1+\epsilon)}} T^{\frac{1}{1+\epsilon}})$. Moreover, for a finite action set with $|\mathcal{A}| = n$, the regret bound is lowered to $\widetilde{\mathcal{O}}(\sqrt{d} T^{\frac{1}{1+\epsilon}} (\log n)^{\frac{\epsilon}{1+\epsilon}})$.*

Computational complexity. By Lemma 2, a design over general action sets can be computed efficiently. The truncated sample-mean estimator can also be computed in linear time. Moreover, the minimum-distance estimator for $\widehat{\theta}$ is obtained by solving a linear optimization problem and is therefore computable in polynomial time; in the infinite-dimensional case this is handled via a dual formulation (see Appendix C). The dominant per-round cost is the linear pass over the active arms to update estimates and apply elimination tests, which is standard for finite-arm algorithms.

The bound in Corollary 1 is the worst-case regret over all possible action sets $\mathcal{A}$. However, based on geometry of the action set, we can achieve tighter regret bounds, as we see below.

3.1 Special Cases of the Action Set

Simplex. When $\mathcal{A}$ is the simplex, the problem is essentially one of multi-armed bandits with d arms. Consider λ being uniform over canonical basis; then $A(\lambda) = \frac{1}{d} I$, and for each $a \in \mathcal{A}$, we have

$$\mathbb{E}_{x\sim\lambda}[|a^\top A^{-1} x|^{1+\epsilon}] = \mathbb{E}_{x\sim\lambda}[|da^\top x|^{1+\epsilon}] = d^{1+\epsilon} \sum_{i=1}^{d} d^{-1} |a^\top e_i|^{1+\epsilon} = d^\epsilon \sum_{i=1}^{d} |a_i|^{1+\epsilon} \leq d^\epsilon.$$

Since one of the canonical basis vectors (or its negation) must be optimal when $\mathcal{A}$ is the simplex, we can simply restrict to this subset of $2d$ actions, giving the following corollary, which recovers the well-known scaling for heavy-tailed MAB [BCBL13].

Corollary 2. *For the simplex action set $\mathcal{A} = \Delta^d$, if the assumptions of Theorem 3 hold, then MED-PE, with parameters $\gamma = T^{\frac{-2\epsilon}{1+\epsilon}}, \beta = d^{\frac{\epsilon-1}{2}}$ achieves regret $\widetilde{\mathcal{O}}(d^{\frac{\epsilon}{1+\epsilon}} T^{\frac{1}{1+\epsilon}})$.*

l_p-norm ball with radius r for $p \leq 1+\epsilon$. Similarly to the simplex, if we define λ to be uniform over $\{r\mathbf{e}_i\}_{i=1}^d$, then $A(\lambda) = \frac{r^2}{d} I$ for any $v \in \mathcal{B}(\|\cdot\|_p, r)$, and we have

$$\mathbb{E}_{x\sim\lambda}[|a^\top A^{-1} x|^{1+\epsilon}] = \mathbb{E}_{x\sim\lambda}\Big[\Big|\frac{d}{r^2} a^\top x\Big|^{1+\epsilon}\Big] = d^\epsilon \sum_{i=1}^{d} \Big|\frac{a_i}{r}\Big|^{1+\epsilon} \leq d^\epsilon \sum_{i=1}^{d} \Big|\frac{a_i}{r}\Big|^{p} \leq d^\epsilon,$$

where the last inequality is by the definition of the l_p-norm ball.

Corollary 3. *For the action set $\mathcal{A} = \{x : \|x\|_p \leq r\}$ with $p \leq 1+\epsilon$, if the assumptions of Theorem 3 hold, then MED-PE, with parameters $\gamma = T^{\frac{-2\epsilon}{1+\epsilon}}, \beta = d^{\frac{\epsilon-1}{2}}$, has regret of $\widetilde{\mathcal{O}}(d^{\frac{2\epsilon}{1+\epsilon}} T^{\frac{1}{1+\epsilon}})$.*

Matérn Kernels. Our algorithm does not require the action features to lie in a finite-dimensional space, as long as the design and the estimator $a^\top A^{(\gamma)}(\lambda)^{-1} x$ can be computed efficiently. In particular, following the approach of [CJKS21], our method extends naturally to kernel bandits, where the reward function belongs to a reproducing kernel Hilbert space (RKHS) associated with a kernel K satisfying $K(x, y) = \phi(x)^\top \phi(y)$ for some (possibly infinite-dimensional) feature map ϕ. Since our focus is on linear bandits, we defer a full description of the kernel setting to Appendix C, where we also establish the following corollary (stated informally here, with the formal version deferred to Appendix C).

Corollary 4. (Informal) *For the kernel bandit problem with domain $[0, 1]^d$ for a constant value of d, under the Matérn kernel with smoothness parameter $\nu > 0$, the kernelized version of MED-PE (with suitably-chosen parameters) achieves regret $\widetilde{\mathcal{O}}(T^{1-\frac{\epsilon}{1+\epsilon}\cdot\frac{2\nu}{2\nu+d}})$.*

While this does not match the known lower bound (except when $\epsilon = 1$ or in the limit as $\epsilon \to 0$), it significantly improves over the best existing upper bound [CG19], which is only sublinear in T for a relatively narrow range of (ϵ, d, ν). In contrast, our bound is sublinear in T for all such choices.

4 Conclusion

In this paper, we revisited stochastic linear bandits with heavy-tailed rewards and substantially narrowed the gap between known minimax lower and upper regret bounds in both the infinite- and finite-action settings. Our new regression estimator, guided by geometry-aware experimental design, yields improved instance-dependent guarantees that leverage the structure of the action set. Since our geometry-dependent bounds recover the $d^{\frac{2\epsilon}{1+\epsilon}}$ dimension dependence that also appears in our minimax lower bound, it is plausible that this gives the correct minimax rate for general action sets. Closing the remaining gap to establish true minimax-optimal rates for all moment parameters, and precisely characterizing the action-set-dependent complexity term under different geometries, remain promising directions for future work.

Acknowledgement

This work was supported by the Singapore National Research Foundation (NRF) under its AI Visiting Professorship programme, and by NSF Award TRIPODS 202323.

References

[AC11] Jean-Yves Audibert and Olivier Catoni. Robust linear least squares regression. *The Annals of Statistics*, 39(5), October 2011.

[BCBL13] Sébastien Bubeck, Nicolo Cesa-Bianchi, and Gábor Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013.

[BJL15] Christian Brownlees, Emilien Joly, and Gábor Lugosi. Empirical risk minimization for heavy-tailed losses. *The Annals of Statistics*, 43(6), December 2015.

[BTA02] François Baccelli, Gérard H. Taché, and Etienne Altman. Flow complexity and heavy-tailed delays in packet networks. *Performance Evaluation*, 49(1–4):427–449, 2002.

[Cat12] Olivier Catoni. *Challenging the Empirical Mean and Empirical Variance: A Deviation Study*, volume 1906 of *Lecture Notes in Mathematics*. Springer, 2012.

[CB00] Rama Cont and Jean-Philippe Bouchaud. Herd behavior and aggregate fluctuations in financial markets. *Macroeconomic Dynamics*, 4(2):170–196, 2000.

[CG19] Sayak Ray Chowdhury and Aditya Gopalan. Bayesian optimization under heavy-tailed payoffs. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[CHDH25] Yu Chen, Jiatai Huang, Yan Dai, and Longbo Huang. uniINF: Best-of-both-worlds algorithm for parameter-free heavy-tailed MABs. In *International Conference on Learning Representations (ICLR)*, 2025.

[CJKS21] Romain Camilleri, Kevin Jamieson, and Julian Katz-Samuels. High-dimensional experimental design and kernel bandits. In *International Conference on Machine Learning (ICML)*, pages 1227–1237. PMLR, 2021.

[Con01] R. Cont. Empirical properties of asset returns: Stylized facts and statistical issues. *Quantitative Finance*, 1(2):223–236, 2001.

[CvdLG20] Yong-Kyu Choi, Erko van der Laan, and Omar Ghattas. Modeling heavy-tailed conversion values in real-time bidding. In *ACM International Conference on Web Search and Data Mining (WSDM)*, pages 870–878, 2020.

[DLLO16] Luc Devroye, Matthieu Lerasle, Gabor Lugosi, and Roberto I. Oliveira. Sub-Gaussian mean estimators. *The Annals of Statistics*, 44(6):2695 – 2725, 2016.

[HDH22] Jiatai Huang, Yan Dai, and Longbo Huang. Adaptive best-of-both-worlds algorithm for heavy-tailed multi-armed bandits. In *International Conference on Machine Learning (ICML)*, volume 162, pages 9173–9200. PMLR, 17–23 Jul 2022.

[HS14] Daniel Hsu and Sivan Sabato. Heavy-tailed regression with a generalized median-of-means. In *International Conference on Machine Learning (ICML)*, volume 32, pages 37–45. PMLR, 22–24 Jun 2014.

[HW19] Qiyang Han and Jon A. Wellner. Convergence rates of least squares regression estimators with heavy-tailed errors. *The Annals of Statistics*, 47(4):2286 – 2319, 2019.

[HZWY23] Jiayi Huang, Han Zhong, Liwei Wang, and Lin Yang. Tackling heavy-tailed rewards in reinforcement learning with function approximation: Minimax optimal and instance-dependent regret bounds. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[JSP21] Saravanan Jebarajakirthy, Paurav Shukla, and Prashant Palvia. Heavy-tailed distributions in online ad response: A marketing analytics perspective. *Journal of Business Research*, 124:818–830, 2021.

[KK23] Minhyun Kang and Gi-Soo Kim. Heavy-tailed linear bandit with Huber regression. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 216, pages 1027–1036. PMLR, 31 Jul–04 Aug 2023.

[LM19a] Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019.

[LM19b] Gábor Lugosi and Shahar Mendelson. Sub-Gaussian estimators of the mean of a random vector. *The Annals of Statistics*, 47(2):783 – 794, 2019.

[LS20] Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

[LS24] Xiang Li and Qiang Sun. Variance-aware decision making with linear function approximation under heavy-tailed rewards. *Transactions on Machine Learning Research*, 2024.

[LWHZ19] Shiyin Lu, Guanghui Wang, Yao Hu, and Lijun Zhang. Optimal algorithms for Lipschitz bandits with heavy-tailed rewards. In *International Conference on Machine Learning (ICML)*, volume 97, pages 4154–4163. PMLR, 09–15 Jun 2019.

[LYLO20] Kyungjae Lee, Hongjun Yang, Sungbin Lim, and Songhwai Oh. Optimal algorithms for stochastic multi-armed bandits with heavy tailed rewards. In *Conference on Neural Information Processing Systems (NeurIPS)*, volume 33, pages 8452–8462, 2020.

[MY16] Andres Muñoz Medina and Scott Yang. No-regret algorithms for heavy-tailed linear bandits. In *International Conference on Machine Learning (ICML)*, pages 1642–1650, 2016.

[RVHH15] James A. Roberts, László A. E. Varnai, Brenton H. Houghton, and David Hughes. Heavy-tailed distributions in the amplitude of neural oscillations. *Journal of Neuroscience*, 35(19):7313–7323, 2015.

[Sas15] Igal Sason. An improved reverse pinsker inequality for probability distributions on a finite set. *CoRR*, abs/1503.03417, 2015.

[SBC17] Jonathan Scarlett, Ilija Bogunovic, and Volkan Cevher. Lower bounds on regret for noisy Gaussian process bandit optimization. In *Conference on Learning Theory (COLT)*. 2017.

[SYKL18] Han Shao, Xiaotian Yu, Irwin King, and Michael R Lyu. Almost optimal algorithms for linear stochastic bandits with heavy-tailed payoffs. In *Conference on Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

[SZF20] Qiang Sun, Wen-Xin Zhou, and Jianqing Fan. Adaptive Huber regression. *Journal of the American Statistical Association*, 115(529):254–265, 2020.

[VBJ+21] Sattar Vakili, Nacime Bouziani, Sepehr Jalali, Alberto Bernacchia, and Da-shan Shiu. Optimal order simple regret for Gaussian process bandits. *Conference on Neural Information Processing Systems (NeurIPS)*, 34:21202–21215, 2021.

[VKP21] Sattar Vakili, Kia Khezeli, and Victor Picheny. On information gain and regret bounds in Gaussian process bandits. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.

[WS21] Lai Wei and Vaibhav Srivastava. Minimax policy for heavy-tailed bandits. *IEEE Control Systems Letters*, 5(4):1423–1428, 2021.

[WZZZ25] Jing Wang, Yu-Jie Zhang, Peng Zhao, and Zhi-Hua Zhou. Heavy-tailed linear bandits: Huber regression with one-pass update. *arXiv preprint arXiv:2503.00419*, 2025.

[XWW+23] Bo Xue, Yimu Wang, Yuanyu Wan, Jinfeng Yi, and Lijun Zhang. Efficient algorithms for generalized linear bandits with heavy-tailed rewards. In *Conference on Neural Information Processing Systems (NeurIPS)*, volume 36, pages 70880–70891, 2023.

[XWWZ20] Bo Xue, Guanghui Wang, Yimu Wang, and Lijun Zhang. Nearly optimal regret for stochastic linear bandits with heavy-tailed payoffs. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2936–2942, 7 2020.

[YNKL18] Xiaotian Yu, Yichuan Nevmyvaka, Irwin King, and Michael R. Lyu. Pure exploration of multi-armed bandits with heavy-tailed payoffs. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2018.

[ZHYW21] Han Zhong, Jiayi Huang, Lin Yang, and Liwei Wang. Breaking the moments condition barrier: No-regret algorithm for bandits with super heavy-tailed payoffs. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

A Upper Bound Proofs

A.1 Proof of Lemma 1 (Confidence Interval)

We first state a well known guarantee of the truncated mean estimator.

Lemma 3. ***(Lemma 1 of [BCBL13])*** *Let $X_1, \ldots, X_n$ be i.i.d. random variables such that $\mathbb{E}[|X_i|^{1+\epsilon}] \le u$ for some $\epsilon \in (0, 1]$. Then the truncated empirical mean estimator $\widehat{\mu}(\{X_i\}_{i=1}^n, \delta) := \frac{1}{n}\sum_{i=1}^n X_i \mathbb{I}\{|X_i| \le \left(\frac{ut}{\log(\delta^{-1})}\right)^{\frac{1}{1+\epsilon}}\}$ satisfies with probability at least $1-\delta$ that*

$$|\widehat{\mu}(\{X_i\}_{i=1}^n, \delta) - \mu| \le 4u^{\frac{1}{1+\epsilon}} \left(\frac{\log(\delta^{-1})}{n}\right)^{\frac{\epsilon}{1+\epsilon}}.$$

Let $W^{(a)} := \widehat{\mu}\left(\{a^\top A^{(\gamma)}(\lambda)^{-1} x_i\, y_i\}_{i=1}^n, \frac{\delta}{|\mathcal{A}|}\right)$. We first observe that

$$\begin{aligned}
\max_{a\in\mathcal{A}} |a^\top \widehat{\theta}(\gamma) - a^\top \theta^*| &= \max_{a\in\mathcal{A}} |a^\top \widehat{\theta}(\gamma) - W^{(a)} + W^{(a)} - a^\top \theta^*| \\
&\le \max_{a\in\mathcal{A}} |a^\top \widehat{\theta}(\gamma) - W^{(a)}| + \max_{a\in\mathcal{A}} |W^{(a)} - a^\top \theta^*| \\
&= \min_{\theta} \max_{a\in\mathcal{A}} |a^T\theta - W^{(a)}| + \max_{a\in\mathcal{A}} |W^{(a)} - a^\top \theta^*| && \text{(def. } \widehat{\theta}(\gamma)) \\
&\le 2\max_{a\in\mathcal{A}} |W^{(a)} - a^\top \theta^*|.
\end{aligned}$$

For fixed a, we bound the $(1+\epsilon)$-moment of $a^\top A^{(\gamma)}(\lambda)^{-1} xy$, where $x \sim \lambda$ and $y = x^\top \theta^* + \eta$, as follows:

$$\begin{aligned}
\mathbb{E}\left[\left|a^\top A^{(\gamma)}(\lambda)^{-1} xy\right|^{1+\epsilon}\right] &= \mathbb{E}\left[\left|a^\top A^{(\gamma)}(\lambda)^{-1} x(x^\top\theta^* + \eta)\right|^{1+\epsilon}\right] \\
&\le 2^{1+\epsilon}\mathbb{E}\left[\left|a^\top A^{(\gamma)}(\lambda)^{-1} x(x^\top\theta^*)\right|^{1+\epsilon}\right] + 2^{1+\epsilon}\mathbb{E}\left[\left|a^\top A^{(\gamma)}(\lambda)^{-1} x\right|^{1+\epsilon}|\eta|^{1+\epsilon}\right] \\
& \qquad\qquad (|a+b| \le 2\max\{|a|, |b|\}) \\
&\le 4\mathbb{E}\left[\left|a^\top A^{(\gamma)}(\lambda)^{-1} x\right|^{1+\epsilon}\right] + 4\upsilon\mathbb{E}\left[\left|a^\top A^{(\gamma)}(\lambda)^{-1} x\right|^{1+\epsilon}\right] \\
& \qquad\qquad (\,|x^\top\theta^*| \le 1 \text{ and } \mathbb{E}[|\eta|^{1+\epsilon}] \le \upsilon) \\
&\le 4(1+\upsilon) M_{1+\epsilon}(\lambda; \mathcal{A}, \gamma, \beta). && \text{(def. } M_{1+\epsilon})
\end{aligned}$$

Using this moment bound and Lemma 3, for any a, we have with probability at least $1 - \frac{\delta}{|\mathcal{A}|}$ that

$$|W^{(a)} - \mathbb{E}[W^{(a)}]| \le 16(1+\upsilon)^{\frac{1}{1+\epsilon}} M_{1+\epsilon}(\mathcal{A}, \gamma, \beta)^{\frac{1}{1+\epsilon}} \left(\frac{\log(\delta^{-1}|\mathcal{A}|)}{n}\right)^{\frac{\epsilon}{1+\epsilon}}.$$

Moreover, we have

$$\begin{aligned}
|a^\top\theta^* - \mathbb{E}[W^{(a)}]| &= |\langle\theta^*, a\rangle - \mathbb{E}[a^\top A^{(\gamma)}(\lambda)^{-1} xx^\top\theta^*]| && \text{(def. } W^{(a)}) \\
&= |\langle\theta^*, a\rangle - a^\top A^{(\gamma)}(\lambda)^{-1} A(\lambda)\theta^*| && \text{(where } A(\lambda) = \mathbb{E}[xx^T]) \\
&= |\langle\theta^*, a\rangle - a^\top A^{(\gamma)}(\lambda)^{-1}\big(A^{(\gamma)}(\lambda) - \gamma I\big)\theta^*| && (A(\lambda) = A^{(\gamma)}(\lambda) - \gamma I) \\
&= \gamma|a^\top A^{(\gamma)}(\lambda)^{-1}\theta^*| \\
&= \gamma|a^\top (A(\lambda) + \gamma I)^{-1/2}(A(\lambda) + \gamma I)^{-1/2}\theta^*| \\
&\le \gamma\|a\|_{A^{(\gamma)}(\lambda)^{-1}}\gamma^{-1/2}\|\theta^*\|_{(I+\gamma^{-1}A(\lambda))^{-1}} && \text{(Cauchy–Schwarz)} \\
&\le \gamma^{1/2}\|a\|_{A^{(\gamma)}(\lambda)^{-1}}\|\theta^*\|_2 && (I + \gamma^{-1}A(\lambda) \succeq I) \\
&\le \gamma^{1/2}\beta^{-1}\|\theta^*\|_2 M_{1+\epsilon}(\lambda; \mathcal{A}, \gamma, \beta)^{\frac{1}{1+\epsilon}}. && \text{(def. } M_{1+\epsilon})
\end{aligned}$$

Putting the two inequalities together, and using the union bound completes the proof.

A.2 Proof of Theorem 3 (Regret Bound for *MED-PE*)

Using Lemma 1 for action set $\mathcal{A}_\ell$, we have with probability of at least $1 - \frac{1}{2\ell^2 T}$,

$$\begin{aligned}
\max_{a\in\mathcal{A}_\ell} |a^\top\theta^* - a^\top\widehat{\theta}_\ell| &\le \left(2\gamma^{1/2}\|\theta^*\|_2\beta^{-1} + 32(1+\upsilon)^{\frac{1}{1+\epsilon}}\left(\frac{\log(2l^2T|\mathcal{A}_\ell|)}{\tau_\ell}\right)^{\frac{\epsilon}{1+\epsilon}}\right) M_{1+\epsilon}(\lambda_\ell^*;\mathcal{A}_\ell,\gamma,\beta)^{\frac{1}{1+\epsilon}} \\
&\le 2\gamma^{1/2}b\beta^{-1}M_{1+\epsilon}(\lambda_\ell^*;\mathcal{A}_\ell,\gamma,\beta)^{\frac{1}{1+\epsilon}} + \epsilon_\ell && \text{(choice of } \tau_\ell \text{ in Algorithm 1)} \\
&\le 2\gamma^{1/2}b\beta^{-1}M^*_{1+\epsilon}(\mathcal{A},\gamma,\beta)^{\frac{1}{1+\epsilon}} + \epsilon_\ell && \text{(def. } M^*_{1+\epsilon})
\end{aligned}$$

Now we define the event $\mathcal{E} := \bigcap_{\ell=1}^\infty \bigcap_{x\in\mathcal{A}_\ell} \mathcal{E}_{x,l}(\mathcal{A}_\ell)$, where

$$\mathcal{E}_{x,l}(\mathcal{V}) := \left\{|x^\top\widehat{\theta}_\ell(\mathcal{V}) - x^\top\theta^*| \le \epsilon_\ell + 2b\gamma^{1/2}\beta^{-1}M^*_{1+\epsilon}(\mathcal{A},\gamma,\beta)^{\frac{1}{1+\epsilon}}\right\},$$

with $\widehat{\theta}_\ell(\cdot)$ corresponding to $\widehat{\theta}_\ell$ in Algorithm 1 with an explicit dependence on the action subset. Then, we have

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{\ell=1}^\infty \bigcup_{x\in\mathcal{A}_\ell} \{\mathcal{E}^c_{x,\ell}(\mathcal{A}_\ell)\}\right) &\le \sum_{\ell=1}^\infty \mathbb{P}\left(\bigcup_{x\in\mathcal{A}_\ell}\{\mathcal{E}^c_{x,\ell}(\mathcal{A}_\ell)\}\right) \\
&= \sum_{\ell=1}^\infty \sum_{\mathcal{V}\subseteq\mathcal{A}} \mathbb{P}\left(\bigcup_{x\in\mathcal{V}}\{\mathcal{E}^c_{x,\ell}(\mathcal{V})\}\,\middle|\,\mathcal{A}_\ell = \mathcal{V}\right)\mathbb{P}(\mathcal{A}_\ell = \mathcal{V}) \\
&\le \sum_{\ell=1}^\infty \sum_{\mathcal{V}\subseteq\mathcal{A}} \tfrac{1}{2\ell^2T}\mathbb{P}(\mathcal{A}_\ell = \mathcal{V}) \le \frac{1}{T}, \quad \text{(union bound and } \textstyle\sum_{\ell=1}^\infty \frac{1}{\ell^2} < 2)
\end{aligned}$$

As $\mathbb{E}[R_T\mathbf{1}_{\mathcal{E}^c}] = \mathbb{E}[R_T|\mathcal{E}^c]\mathbb{P}[\mathcal{E}^c] \le (\sup_{x,x'} x'^\top\theta^* - x^\top\theta^*)T\frac{1}{T} \le 2$, for the rest of the proof we assume event $\mathcal{E}$.

Let $x^* = \operatorname{argmax}_{x\in\mathcal{A}} x^\top\theta^*$; then, for every ℓ such that $2\epsilon_\ell \ge 4b\gamma^{1/2}\beta^{-1}M_{1+\epsilon}(\mathcal{A},\gamma,\beta)^{\frac{1}{1+\epsilon}}$ and any $x \in \mathcal{A}_\ell$, we have

$$\begin{aligned}
x^\top\widehat{\theta}_\ell - x^{*\top}\widehat{\theta}_\ell &= (x^\top\widehat{\theta}_\ell - x^\top\theta^*) + (x^\top\theta^* - x^{*\top}\theta^*) + (x^{*\top}\theta^* - x^{*\top}\widehat{\theta}_\ell) \\
&\le 2\epsilon_\ell + 4b\gamma^{1/2}\beta^{-1}M^*_{1+\epsilon}(\mathcal{A},\gamma,\beta)^{\frac{1}{1+\epsilon}} && \text{(def. } \mathcal{E} \text{ and def. } x^*) \\
&\le 4\epsilon_\ell. && \text{(assumption on } \epsilon_\ell)
\end{aligned}$$

Therefore, recalling the elimination rule in Algorithm 1, we have by induction that $x^* \in \mathcal{A}_{\ell+1}$. We also claim that all suboptimal actions of gap more than $8\epsilon_\ell = 16\epsilon_{\ell+1}$ are eliminated at the end of epoch ℓ. To see this, let $x' \in \mathcal{A}_\ell$ be such an action, and observe that

$$\begin{aligned}
\max_{x\in\mathcal{A}_\ell}\left(x'^\top\widehat{\theta}_\ell - x^\top\widehat{\theta}_\ell\right) &\ge x^{*\top}\widehat{\theta}_\ell - x^\top\widehat{\theta}_\ell && (x^* \in \mathcal{A}_\ell) \\
&\ge x^{*\top}\theta^* - x^\top\theta^* - 2\epsilon_\ell - 4b\gamma^{1/2}\beta^{-1}M^*_{1+\epsilon}(\mathcal{A},\gamma,\beta)^{\frac{1}{1+\epsilon}} && \text{(shown above)} \\
&\ge x^{*\top}\theta^* - x^\top\theta^* - 4\epsilon_\ell && \text{(assumption on } \epsilon_\ell) \\
&> 4\epsilon_\ell. && \text{(gap exceeds } 8\epsilon_\ell)
\end{aligned}$$

In summary, the above arguments show that when $2\epsilon_\ell \ge 4b\gamma^{1/2}\beta^{-1}M^*_{1+\epsilon}(\mathcal{A},\gamma,\beta)^{\frac{1}{1+\epsilon}}$, the regret incurred in epoch $\ell+1$ is at most $16\epsilon_{\ell+1}$. Since $\mathcal{A}_{\ell+1} \subseteq \mathcal{A}_\ell$, this also implies that even when ℓ increases beyond such a point, we still incur regret at most $32b\gamma^{1/2}\beta^{-1}M^*_{1+\epsilon}(\mathcal{A},\gamma,\beta)^{\frac{1}{1+\epsilon}}$.

Finally, we can upper bound the regret as follows:

$$\begin{aligned}
\mathbb{E}[R_T] &\le \sum_\ell \tau_\ell\Big(\sup_{x\in\mathcal{A}_\ell} x^{*T}\theta^* - x^T\theta^*\Big) \\
&\le \sum_\ell \tau_\ell \max\{16\epsilon_\ell, 32b\gamma^{1/2}\beta^{-1}M^*_{1+\epsilon}(\mathcal{A},\gamma,\beta)^{\frac{1}{1+\epsilon}}\} && \text{(shown above)} \\
&\le \sum_\ell 16\tau_\ell\epsilon_\ell + T\zeta && (\zeta := 32b\gamma^{1/2}\beta^{-1}M_{1+\epsilon}(\mathcal{A},\gamma,\beta)^{\frac{1}{1+\epsilon}})
\end{aligned}$$

$$\le \sum_{\ell\,:\,16\epsilon_\ell \ge \omega} 16\epsilon_\ell \tau_\ell + T\omega + T\zeta \qquad \text{(for any } \omega \ge 0)$$

$$\le \sum_{\ell\,:\,16\epsilon_\ell \ge \omega} 16\epsilon_\ell 32^{\frac{1+\epsilon}{\epsilon}} (1+\upsilon)^{\frac{1}{\epsilon}} \varepsilon_\ell^{-\frac{1+\epsilon}{\epsilon}} M^*_{1+\epsilon}(\mathcal{A},\gamma,\beta)^{\frac{1}{\epsilon}} \log(2l^2|\mathcal{A}|T) + T(\omega+\zeta)$$

(def. τ_ℓ in Alg. 1)

$$\le \sum_{\ell\,:\,16\epsilon_\ell \ge \omega} C_1'(1+\upsilon)^{\frac{1}{\epsilon}} \varepsilon_\ell^{-\frac{1}{\epsilon}} M^*_{1+\epsilon}(\mathcal{A},\gamma,\beta)^{\frac{1}{\epsilon}} \log(2l^2|\mathcal{A}|T) + T(\omega+\zeta)$$

(for some constant C_1')

$$\le C_1(1+\upsilon)^{\frac{1}{1+\epsilon}} M^*_{1+\epsilon}(\mathcal{A},\gamma,\beta)^{\frac{1}{1+\epsilon}} \log(2|\mathcal{A}|T\log_2^2 T)^{\frac{\epsilon}{1+\epsilon}} T^{\frac{1}{1+\epsilon}} + T\zeta$$

($\omega := M^*_{1+\epsilon}(\cdot)^{\frac{1}{1+\epsilon}} \log(2|\mathcal{A}|T\log_2^2 T)^{\frac{\epsilon}{1+\epsilon}} T^{\frac{-\epsilon}{1+\epsilon}}$ and $\ell \le \log_2 T$; see below)

$$\le \left(C_0\beta^{-1}b + C_1(1+\upsilon)^{\frac{1}{1+\epsilon}} \log(|\mathcal{A}|T\log_2^2 T)^{\frac{\epsilon}{1+\epsilon}}\right) M^*_{1+\epsilon}(\mathcal{A}, T^{\frac{-2\epsilon}{1+\epsilon}}, \beta)^{\frac{1}{1+\epsilon}} T^{\frac{1}{1+\epsilon}}.$$

(def. ζ and $\gamma = T^{\frac{-2\epsilon}{1+\epsilon}}$)

In more detail, the second-last step upper bounds $\sum_{\ell\,:\,16\epsilon_\ell \ge \omega} \epsilon_\ell^{-\frac{1}{\epsilon}}$ by a constant times its largest possible term $\omega^{-\frac{1}{\epsilon}}$, since $\{\epsilon_\ell\}_{\ell \ge 1}$ is exponentially decreasing. Since the choice of ω contains $(M^*_{1+\epsilon})^{\frac{1}{1+\epsilon}}$, the overall $M^*_{1+\epsilon}$ dependence simplifies as $\left(\frac{M^*_{1+\epsilon}}{(M^*_{1+\epsilon})^{\frac{1}{1+\epsilon}}}\right)^{\frac{1}{\epsilon}} = \left((M^*_{1+\epsilon})^{\frac{\epsilon}{1+\epsilon}}\right)^{\frac{1}{\epsilon}} = (M^*_{1+\epsilon})^{\frac{1}{1+\epsilon}}$.

B Unit Ball Lower Bound

In this appendix, we prove the following lower bound for the case that the action set is the unit ball.

Theorem 4. *Let the action set be $\mathcal{A} = \{x \in \mathbb{R}^d : \|x\|_2 \le 1\}$, and the $(1+\epsilon)$-absolute moment of the error distribution be bounded by 1. Then, for any algorithm, there exists $\theta^* \in \mathbb{R}^d$ such that $\sup_{x\in\mathcal{A}} |x^\top \theta^*| \le 1$, and such that for $T \ge d^2$, the regret incurred is $\Omega(d^{\frac{2\epsilon}{1+\epsilon}} T^{\frac{1}{1+\epsilon}})$.*

Since the KL divergence between Bernoulli random variables Ber(p) and Ber(q) goes to infinity as $p \to 0$, and $\theta^\top x$ can be zero for unit ball, we cannot use the same reward distribution as before. However, we can overcome this by shifting all probabilities and adding -1 to the support of the reward random variable. Specifically, we set the error distribution to be:

$$y(x) = \begin{cases} (\frac{1}{\gamma})^{\frac{1}{\epsilon}} & w.p.\ \gamma^{\frac{1}{\epsilon}}(\theta^\top x + 2\sqrt{d}\Delta) \\ 0 & w.p.\ 1-\gamma^{\frac{1}{\epsilon}}(\theta^\top x + 2\sqrt{d}\Delta) - 2\sqrt{d}\Delta \\ -1 & w.p.\ 2\sqrt{d}\Delta \end{cases}$$

with $\gamma := 24\sqrt{d}\Delta$ and Δ to be specified later. For any $\theta \in \{\pm\Delta\}^d$, the absolute value of rewards are bounded by $\sum_{i=1}^d \frac{1}{\sqrt{d}}\Delta = \sqrt{d}\Delta$. Then, assuming $\Delta \le \frac{1}{24\sqrt{d}}$, we have $|\theta^\top x| \le \sqrt{d}\Delta \le \frac{1}{8}$ and $\|\theta\|_2 \le 1$ as well as $\gamma \le 1$, and the $(1+\epsilon)$-central absolute moment is bounded by:

$$\mathbb{E}[|y(x) - \theta^\top x|^{1+\epsilon} \mid x]$$

$$\le |\gamma^{-\frac{1}{\epsilon}} - \theta^\top x|^{1+\epsilon}(\theta^\top x + 2\sqrt{d}\Delta) + |\theta^\top x|^{1+\epsilon} + |-1-\theta^\top x|^{1+\epsilon} 2\sqrt{d}\Delta \qquad (\gamma \le 1)$$

$$\le 2^{1+\epsilon}\gamma^{-1}3\sqrt{d}\Delta + (\sqrt{d}\Delta)^{1+\epsilon} + 2\sqrt{d}\Delta(\sqrt{d}\Delta+1)^{1+\epsilon} \qquad (|\theta^\top x| \le \sqrt{d}\Delta \le 1 \text{ and } \gamma^{-\frac{1}{\epsilon}} \ge 1)$$

$$\le \frac{2^{1+\epsilon}}{8} + \left(\frac{1}{24}\right)^{1+\epsilon} + \frac{1}{12}\left(\frac{9}{24}\right)^{1+\epsilon} < 1. \qquad (\text{def. } \gamma,\ \Delta \le \tfrac{1}{24\sqrt{d}}, \text{ and } \epsilon \in (0,1])$$

Defining $T_i := T \wedge \min(s : \sum_{t=1}^s x_{t,i}^2 \ge \frac{T}{d})$, we have

$$R_T(\mathcal{A},\theta) = \Delta\mathbb{E}_\theta\left[\sum_{t=1}^T\sum_{i=1}^d\left(\frac{1}{\sqrt{d}} - x_{t,i}\mathtt{sign}(\theta_i)\right)\right]$$

$$\ge \frac{\Delta\sqrt{d}}{2}\mathbb{E}_\theta\left[\sum_{t=1}^T\sum_{i=1}^d\left(\frac{1}{\sqrt{d}} - x_{t,i}\mathtt{sign}(\theta_i)\right)^2\right]$$

(by expanding the square and applying $\|x_t\|_2^2 \le 1$)

$$\geq \frac{\Delta\sqrt{d}}{2}\sum_{i=1}^{d}\mathbb{E}_\theta\left[\sum_{t=1}^{T_i}\left(\frac{1}{\sqrt{d}} - x_{t,i}\mathtt{sign}(\theta_i)\right)^2\right].$$

Now we define $U_i(b) := \sum_{t=1}^{T_i}\left(\frac{1}{\sqrt{d}} - x_{t,i}b\right)^2$, which gives

$$U_i(1) \leq 2\sum_{t=1}^{T_i}\frac{1}{d} + 2\sum_{t=1}^{T_i}x_{t,i}^2 \leq \frac{4T}{d} + 2.$$

Then, for any $\theta, \theta' \in \{\pm\Delta\}^d$ that only differ in i-th element, we have

$$\begin{aligned}
\mathbb{E}_\theta[U_i(1)] &\geq \mathbb{E}_{\theta'}[U_i(1)] - \left(\frac{4T}{d}+2\right)\sqrt{\frac{1}{2}\mathrm{KL}(\mathbb{P}_\theta\|\mathbb{P}_{\theta'})} && \text{(Pinsker's inequality)}\\
&\geq \mathbb{E}_{\theta'}[U_i(1)] - \left(\frac{4T}{d}+2\right)\sqrt{\frac{1}{2}\mathbb{E}_\theta\left[\sum_{t=1}^{T_i}\mathrm{KL}(y_\theta(x_t)\|y_{\theta'}(x_t))\right]} && \text{(Chain rule)}\\
&\geq \mathbb{E}_{\theta'}[U_i(1)] - \left(\frac{4T}{d}+2\right)\sqrt{\frac{1}{2}\mathbb{E}_\theta\left[\sum_{t=1}^{T_i}24^{\frac{1}{\epsilon}}8\sqrt{d}^{\frac{1-\epsilon}{\epsilon}}\Delta^{\frac{1+\epsilon}{\epsilon}}x_{t,i}^2\right]}\\
& && \text{(Inverse Pinsker's inequality; see below)}\\
&\geq \mathbb{E}_{\theta'}[U_i(1)] - 24^{\frac{1}{2\epsilon}}2\Delta^{\frac{1+\epsilon}{2\epsilon}}\sqrt{d}^{\frac{1-\epsilon}{2\epsilon}}\left(\frac{4T}{d}+2\right)\sqrt{\mathbb{E}_\theta\left[\sum_{t=1}^{T_i}x_{t,i}^2\right]}\\
&\geq \mathbb{E}_{\theta'}[U_i(1)] - 24^{\frac{1}{2\epsilon}}12\sqrt{2}\Delta^{\frac{1+\epsilon}{2\epsilon}}\sqrt{d}^{\frac{1-\epsilon}{2\epsilon}}\frac{T}{d}\sqrt{\frac{T}{d}}. && (d \leq T, \textstyle\sum_{t=1}^{T_i}x_{t,i}^2 \leq \frac{T}{d}+1)
\end{aligned}$$

Note that the version of the chain rule with a random stopping time can be found in [LS20, Exercise 15.7]. We detail the step using inverse Pinsker's inequality ([Sas15]) as follows:

$$\begin{aligned}
\mathrm{KL}(y_\theta(x_t)\|y_{\theta'}(x_t)) &\leq \frac{2}{\min_{a\in\{\gamma^{-\frac{1}{\epsilon}},0,-1\}}\mathbb{P}[y_{\theta'}(x_t)=a]}\sup_a|\mathbb{P}[y_\theta(x_t)=a]-\mathbb{P}[y_{\theta'}(x_t)=a]|^2\\
&\leq \frac{2}{\gamma^{\frac{1}{\epsilon}}\sqrt{d}\Delta}(\gamma^{\frac{1}{\epsilon}}2\Delta x_{t,i})^2\\
&\leq 24^{\frac{1}{\epsilon}}8\sqrt{d}^{\frac{1}{\epsilon}-1}\Delta^{\frac{1}{\epsilon}+1}x_{t,i}^2. && (\gamma = 24\sqrt{d}\Delta)
\end{aligned}$$

Using the above lower bound on $\mathbb{E}_\theta[U_i(1)]$, and setting $\Delta := 24^{\frac{-1}{1+\epsilon}}d^{\frac{3\epsilon-1}{2(1+\epsilon)}}(288T)^{\frac{-\epsilon}{1+\epsilon}}$ (noting $288 = (12\sqrt{2})^2$), we have the following:

$$\begin{aligned}
\mathbb{E}_\theta[U_i(1)] + \mathbb{E}_{\theta'}[U_i(-1)] &\geq \mathbb{E}_{\theta'}[U_i(1)+U_i(-1)] - 24^{\frac{1}{2\epsilon}}12\sqrt{2}\Delta^{\frac{1+\epsilon}{2\epsilon}}\sqrt{d}^{\frac{1-\epsilon}{2\epsilon}}\frac{T}{d}\sqrt{\frac{T}{d}}\\
&= 2\mathbb{E}_{\theta'}\left[\frac{T_i}{d} + \sum_{t=1}^{T_i}x_{t,i}^2\right] - 24^{\frac{1}{2\epsilon}}12\sqrt{2}\Delta^{\frac{1+\epsilon}{2\epsilon}}\sqrt{d}^{\frac{1-\epsilon}{2\epsilon}}\frac{T}{d}\sqrt{\frac{T}{d}}\\
&\geq \frac{2T}{d} - \frac{T}{d} = \frac{T}{d}. && (T_i \geq 0, \text{ def. } T_i, \text{ choice of } \Delta)
\end{aligned}$$

Note also that $\Delta \leq \frac{1}{24\sqrt{d}}$ (as required earlier) since $T \geq d^2$. We now combine the preceding equation with our earlier lower bound on R_T. By averaging overall $\theta \in \{\pm\Delta\}^d$, we conclude that there exists some θ^* such that

$$\begin{aligned}
R_T(\mathcal{A},\theta^*) &\geq \frac{\Delta\sqrt{d}}{2}\frac{1}{2^d}\sum_{\theta\in\{-\Delta,\Delta\}^d}R_T(\mathcal{A},\theta)\\
&\geq \frac{\Delta\sqrt{d}}{4}\sum_{i=1}^{d}\sum_{\theta_i\in\{-\Delta,\Delta\}}\mathbb{E}_\theta[U_i(\mathtt{sign}(\theta_i))]. && (R_T \text{ bound and } \textstyle\sum_{\{\theta_j\}_{j\neq i}} 1 = 2^{d-1})
\end{aligned}$$

$$\geq \frac{1}{4}T\sqrt{d}\Delta \qquad (\mathbb{E}_\theta[U_i(1)] + \mathbb{E}_{\theta'}[U_i(-1)] \geq \tfrac{T}{d})$$

$$\geq \frac{1}{4 \cdot 24 \cdot 12\sqrt{2}} d^{\frac{2\epsilon}{1+\epsilon}} T^{\frac{1}{1+\epsilon}}. \qquad (\text{choice of } \Delta,\ \epsilon \in [0,1])$$

C Extension to Kernel Bandits

C.1 Problem Setup

We consider an unknown reward function $f : \mathcal{A} \to \mathbb{R}$ lying in the reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ associated with a given kernel K, i.e., $f(x) = \langle f, K(x, \cdot)\rangle_K$. Similar to the linear bandit setting, we assume that $\max_{x\in\mathcal{A}} |f(x)| \leq 1$ and $\|f\|_K \leq b$ for some $b > 0$.

At each round $t = 1, 2, \ldots, T$, the learner chooses an action $x_t \in \mathcal{A} \subseteq [0, 1]^d$ and observes the reward

$$y_t \;=\; f(x_t) \;+\; \eta_t,$$

where η_t are independent noise terms that satisfy $\mathbb{E}[\eta_t] = 0$ and $\mathbb{E}\big[|\eta_t|^{1+\epsilon}\big] \leq \upsilon$ for some $\epsilon \in (0, 1]$ and finite $\upsilon > 0$. Letting $x^\star \in \arg\max_{x\in[0,1]^d} f(x)$ be an optimal action, the cumulative expected regret after T rounds is

$$R_T \;=\; \sum_{t=1}^{T} \big(f(x^*) - f(x_t)\big).$$

Given $(\mathcal{A}, \epsilon, \upsilon)$, the objective is to design a policy for sequentially selecting the points (i.e., x_t for $t = 1, \ldots, T$) in order to minimize R_T. We focus on the Matérn kernel, defined as follows:

$$K_{\nu,l}(x, x') := \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\|x - x'\|_2 \sqrt{2\nu}}{l} \right)^{\nu} B_\nu \left(\frac{\|x - x'\|_2 \sqrt{2\nu}}{l} \right),$$

where Γ is the Gamma function, B_ν is the modified Bessel function, and (ν, l) are parameters corresponding to smoothness and lengthscale.

We focus on the case that $\mathcal{A}$ is a *finite* subset of $[0, 1]^d$, but it is well known (e.g., see [VBJ+21, Assumption 4]) that the resulting regret bounds extend to the continuous domain $[0, 1]^d$ via a discretization argument with with $\log |\mathcal{A}| = O(\log T)$.

C.2 Proof of Corollary 4

We state a more precise version of Corollary 4 as follows.

Theorem 5. *For any unknown reward function $f : \mathcal{A} \to \mathbb{R}$ lying in the RKHS of the Matérn kernel with parameters (ν, l), for some finite set $\mathcal{A} \subseteq [0, 1]^d$, assuming that $\max_{x\in\mathcal{A}} |f(x)| \leq 1$ and $\|f\|_K \leq b$ for some $b > 0$, we have*

$$M^*(\mathcal{A}, T^{\frac{-2\epsilon}{1+\epsilon}}, 1) \leq C T^{\epsilon \cdot \frac{d}{2\nu+d}},$$

for some constant C, and Algorithm 1 achieves regret of

$$R_T(f, \mathcal{A}) \leq \left(C_0' b + C_1' (1 + \upsilon)^{\frac{1}{1+\epsilon}} \log(|\mathcal{A}| T \log^2 T)^{\frac{\epsilon}{1+\epsilon}} \right) T^{1 - \frac{\epsilon}{1+\epsilon}\frac{2\nu}{2\nu+d}},$$

for some constants C_0', C_1'. Note that the constants may depend on the kernel parameters (ν, l) and the dimension d.

We now proceed with the proof. We first argue that Algorithm 1 and Theorem 3 can still be applied (with x replacing a and $f(x)$ replacing $a^\top \theta^*$) in the kernel setting. The reasoning is the same as the case $\epsilon = 1$ handled in [CJKS21], so we keep the details brief.

Recall that for any kernel K, there exists a (possibly infinite dimensional) feature map $\phi : \mathcal{A} \to \mathcal{H}$ such that $K(x, x') = \phi(x)^\top \phi(x')$. For any $\lambda \in \Delta_{\mathcal{A}}$, we define $k_\lambda(\cdot) \in \mathbb{R}^{|\mathcal{A}|}$ such that for $\psi \in \mathcal{H}$, $k_\lambda(\psi)_i := \sqrt{\lambda_i}\phi(x_i)^\top \psi$, and $K_\lambda \in \mathbb{R}^{|\mathcal{A}|\times|\mathcal{A}|}$ such that $(K_\lambda)_{i,j} := \sqrt{\lambda_i}\sqrt{\lambda_j} K(x_i, x_j)$. Then similar to [CJKS21, Lemma 2], we have for any $\psi, \rho \in \mathcal{H}$ that

$$\psi^\top A^{(\gamma)}(\lambda)^{-1} \rho = \gamma^{-1} \psi^\top \rho - \gamma^{-1} k_\lambda(\psi)(K_\lambda + I_{|\mathcal{A}|})^{-1} k_\lambda(\rho).$$

Then the gradient for the experimental design problem $\inf_{\lambda\in\Delta_{\mathcal{V}}} \max_{v\in\mathcal{V}} \|\phi(v)\|_{A^{(\gamma)}(\lambda)^{-1}}$ (which is an upper bound for our experimental design objective $M_{1+\epsilon}(\lambda;\mathcal{V},\gamma,1)$ by the proof of Lemma 2) can be computed efficiently. Moreover, Theorem 3 still holds because the the kernel setup can be viewed as a linear setup in an infinite-dimensional feature space (after applying the feature map ϕ to the action set), and our analysis does not use the finiteness of the dimension.

Given Theorem 3, the main remaining step is to upper bound $M^*_{1+\epsilon}$. To do so, we use the well-known polynomial eigenvalue decay of the Matérn kernel. Specifically, the j-th eigenvalue φ_j satisfies $\varphi_j \le \mathcal{O}(j^{-\kappa})$ with $\kappa = \frac{2\nu+d}{d}$ (e.g., see [VBJ+21]). We let $\lambda^*_D \in \arg\max_{\lambda\in\Delta_{\mathcal{A}}} \log\det\big(A^{(\gamma)}(\lambda)\big)$, and proceed as follows:

$$
\begin{aligned}
M^*_{1+\epsilon}(\mathcal{A},\gamma,1)^{\frac{2}{1+\epsilon}} &\le \max_{\mathcal{V}\in\mathcal{A}} \inf_{\lambda\in\Delta_{\mathcal{V}}} 2^{\frac{2}{1+\epsilon}} \max_{v\in\mathcal{V}} \|\phi(v)\|^2_{A^{(\gamma)}(\lambda)^{-1}} && \text{(shown in the proof of Lemma 2)}\\
&\le 4\mathrm{Tr}\left(A(\lambda^*_D)(A(\lambda^*_D)+\gamma I)^{-1}\right) && \text{[CJKS21, Lemma 3]}\\
&= 4\mathrm{Tr}\left(K_{\lambda^*_D}(K_{\lambda^*_D}+\gamma I)^{-1}\right)\\
&= 4\sum_{j=1}^{|\mathcal{A}|} \frac{\varphi_j}{\varphi_j+\gamma}\\
&\le 4\sum_{j=1}^{|\mathcal{A}|} \frac{cj^{-\kappa}}{cj^{-\kappa}+\gamma} && \text{(for some constant } c\ge 1 \text{ dependent on } l,\nu,d)\\
&\le 4c\sum_{j\le\gamma^{-\frac{1}{\kappa}}} \frac{j^{-\kappa}}{j^{-\kappa}+\gamma} + 4c\sum_{j>\gamma^{-\frac{1}{\kappa}}} \frac{j^{-\kappa}}{j^{-\kappa}+\gamma} && (c\ge 1)\\
&\le 4c\gamma^{-1/\kappa} + 4c\sum_{j>\gamma^{-\frac{1}{\kappa}}} \frac{j^{-\kappa}}{\gamma} && \text{(dropping terms in denominators)}\\
&\le 4c\gamma^{-\frac{1}{\kappa}} + 4c(\gamma^{-\frac{1}{\kappa}})^{1-\kappa}\frac{1}{(\kappa-1)\gamma} && \text{(bounding sum by integral; } \kappa>1)\\
&= 4c\gamma^{-\frac{1}{\kappa}}\left(1+\frac{1}{\kappa-1}\right)\\
&= 4c\frac{2\nu+d}{2\nu}T^{\frac{2\epsilon}{1+\epsilon}\frac{d}{2\nu+d}}. && (\gamma = T^{\frac{-2\epsilon}{1+\epsilon}} \text{ and } \kappa = \tfrac{2\nu+d}{d})
\end{aligned}
$$

Taking the square root on both sides gives $M^*_{1+\epsilon}(\mathcal{A},\gamma,1)^{\frac{1}{1+\epsilon}} = \widetilde{\mathcal{O}}\big(T^{\frac{\epsilon}{1+\epsilon}\frac{d}{2\nu+d}}\big)$, and multiplying by $\widetilde{\mathcal{O}}(T^{\frac{1}{1+\epsilon}}) = \widetilde{\mathcal{O}}(T^{1-\frac{\epsilon}{1+\epsilon}})$ from the regret bound in Theorem 3 gives $\widetilde{\mathcal{O}}(T^{1-\frac{\epsilon}{1+\epsilon}\cdot\frac{2\nu}{2\nu+d}})$ regret as claimed in Corollary 4. By the same reasoning but keeping track of the logarithmic terms, we obtain the regret bound stated in Theorem 5.

C.3 Comparisons of Bounds

Comparison to existing lower bound. In Figure 2, we compare our regret upper bound to the lower bound of $\Omega\big(T^{\frac{\nu+d\epsilon}{\nu(1+\epsilon)+d\epsilon}}\big)$ proved in [CG19]. We see that the upper and lower bounds coincide in certain limits and extreme cases:

- As $\nu/d \to \infty$, the regret approaches $T^{\frac{1}{1+\epsilon}}$ scaling, which matches the regret of linear heavy-tailed bandits in constant dimension.
- As $\nu/d \to 0$ and/or $\epsilon \to 0$, the regret approaches trivial linear scaling in T.
- When $\epsilon = 1$, the regret scales as $\widetilde{\Theta}\big(T^{\frac{\nu+d}{2\nu+d}}\big)$, which matches the optimal scaling for the sub-Gaussian noise setting [SBC17]. As we discussed earlier, this finite-variance setting was already handled in [CJKS21].

For finite ν/d and fixed $\epsilon \in (0,1)$, we observe from Figure 2 that gaps still remain between the upper and lower bounds, but they are typically small, especially when ν/d is not too small.

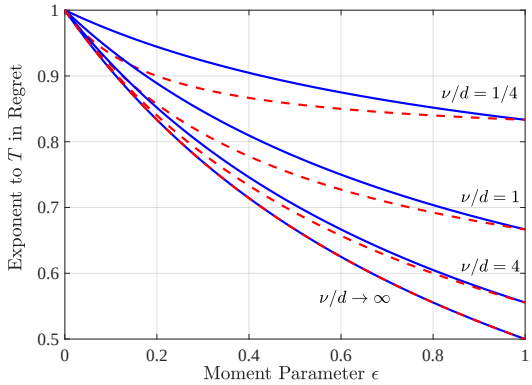


Figure 2: Comparison of our regret upper bound (solid) and the lower bound of [CG19] (dashed). We plot the exponent c such that the regret bound has dependence T^c, with the 4 pairs of curves corresponding to $\nu/d \in \{0.25, 1, 4\}$ and $\nu/d \to \infty$.

Comparison to existing upper bound. In [CG19], a regret upper bound of $\widetilde{\mathcal{O}}(\gamma_T T^{\frac{2+\epsilon}{2(1+\epsilon)}})$ was established, where γ_T is an *information gain* term that satisfies $\gamma_T = \widetilde{\mathcal{O}}(T^{\frac{d}{2\nu+d}})$ for the Matérn kernel [VKP21]. We did not plot this upper bound in Figure 2, because its high degree of suboptimality is easier to describe textually:

- For $\nu/d = 1/4$ and $\nu/d = 1$, their bound exceeds the trivial $\mathcal{O}(T)$ bound for all $\epsilon \in (0, 1]$.
- For $\nu/d = 4$, their bound still exceeds $\mathcal{O}(T)$ for $\epsilon \lesssim 0.28$, and is highly suboptimal for larger ϵ.
- As $\nu/d \to \infty$, the γ_T term becomes insignificant and their bound simplifies to $\widetilde{\mathcal{O}}(T^{\frac{2+\epsilon}{2(1+\epsilon)}})$, which is never better than $\widetilde{\mathcal{O}}(T^{3/4})$ (achieved when $\epsilon = 1$).
- A further weakness when $\epsilon = 1$ is that the optimal γ_T dependence should be $\sqrt{\gamma_T}$ rather than linear in γ_T [SBC17, CJKS21].

For the *squared exponential kernel*, which has exponentially decaying eigenvalues rather than polynomial, these weaknesses were overcome in [CG19] using kernel approximation techniques, to obtain an optimal $\widetilde{\mathcal{O}}(T^{\frac{1}{1+\epsilon}})$ regret bound. Our main contribution above is to establish a new state of the art for the Matérn kernel, which is significantly more versatile in being able to model both highly smooth (high ν) and less smooth (small ν) functions.

D Numerical Experiments

In this section, we perform a simple proof-of-concept experiment to demonstrate that our algorithm can outperform existing methods as the ambient dimension increases. However, we emphasize that our main contributions are theoretical, and we leave detailed experimental studies for future work.

We conduct experiments with horizon $T = 100{,}000$ and action set size $N = 2d$. The true parameter is $\theta^\star = \frac{1}{\sqrt{d}}\mathbf{1}$ (so $\|\theta^\star\|_2 = 1$), and the action set is the subset of the normalized hypercube given by the signed coordinate directions $\mathcal{A} = \{\pm e_i\}_{i=1}^d$. Rewards follow $r_t = x_t^\top \theta^\star + \eta_t$ with heavy-tailed noise $\eta_t \sim \text{ParetoII}(\alpha = 2, \sigma = 1) - \mathbb{E}[\text{ParetoII}(\alpha = 2, \sigma = 1)]$ (centered to zero mean). The proposed algorithm is instantiated with input $\epsilon = 0.5$ and evaluated against the *Confidence Region with Truncated Mean* (CRTM) Algorithm in [XWW+23]. Performance is measured via cumulative pseudo-regret $\sum_{t=1}^T \left(x_t^{\star\top}\theta^\star - x_t^\top \theta^\star\right)$, aggregated over 10 independent repetitions (with identical arm sets and independent noise).

Results. As shown in Figure 3, for all $d \geq 40$ Algorithm 1 achieves comparable or lower mean regret than CRTM, and notably, the gap widens as d increases. Both procedures remain sublinear in T in this controlled setting; however, the regret of Algorithm 1 grows more slowly with d, consistent with the guarantees in Theorem 3.

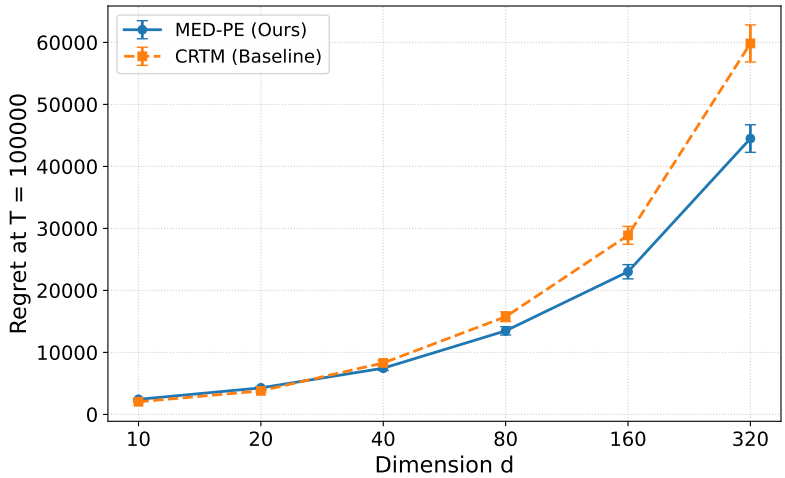


Figure 3: Regret vs dimension d with time horizon $T = 100,000$, and $N = 2d$ arms

E Proof of Theorem 2 (Finite-Arm Lower Bound)

Consider $\log(\cdot)$ with base 2, and define m to be the smallest integer such that $\frac{m}{\log m} \geq \frac{d}{\log n}$. From the assumption $n \in [d, 2^{\lfloor \frac{d}{4} \rfloor}]$ we can readily verify that $d > 4$ and $m \in [4, d]$. For convenience, we assume that d is a multiple of m, since otherwise we can form the construction of the lower bound with $d' = d - (d \bmod m)$ and pad the action vectors with zeros. Letting $d_i := (i-1)m$, we define the action set and the parameter set as follows for some Δ to be specified later:

$$\mathcal{A} := \left\{ a \in \{0,1\}^d : \sum_{j=d_i+1}^{d_{i+1}} a_j = 1, \ \ \forall i \in [d/m] \right\}$$

$$\theta^* \in \Theta := \left\{ \theta \in \{\Delta, 2\Delta\}^d : \sum_{j=d_i+1}^{d_{i+1}} \theta_j = (m+1)\Delta, \ \ \forall i \in [d/m] \right\}.$$

In simple terms, the d-dimensional vectors are arranged in d/m groups of size m; each block in $a \in \mathcal{A}$ has a single entry of 1 (with 0 elsewhere), and each block in θ^* has a single entry of 2Δ (with Δ elsewhere). Observe that if $\Delta \leq \min(\frac{m}{4d}, \frac{1}{4\sqrt{d}})$, then $\|\theta^*\|_2 \leq 1$ and $x^\top \theta^* \leq 1$ as required. Moreover, we have $|\mathcal{A}| = m^{\frac{d}{m}}$, and thus $\log |\mathcal{A}| = \frac{d}{m} \log m \leq \log n$ by the definition of m.

Similar to Theorem 1, we let the reward distribution be

$$y(x) = \begin{cases} (\frac{1}{\gamma})^{\frac{1}{\epsilon}} & \text{w.p. } \gamma^{\frac{1}{\epsilon}} \theta^\top x \\ 0 & \text{w.p. } 1 - \gamma^{\frac{1}{\epsilon}} \theta^\top x \end{cases}$$

with $\gamma := 2\Delta \frac{d}{m}$. The choices of $\mathcal{A}$ and Θ give $\theta^\top x \leq 2\Delta \frac{d}{m}$, so by the same reasoning as in Theorem 1, the $(1+\epsilon)$-moment of the reward distribution is bounded by 1.

Let $\text{ind}_i(x) := \arg\max_{b \in [m]} (x_{d_i+b})$ for fixed $x \in \mathcal{A} \cup \Theta$, and define $T_{i,b} := |\{t : x_{t,d_i+b} = 1\}|$. Moreover, define t_{U} to be a random integer drawn uniformly from $[T]$, which immediately implies that $\mathbb{P}_\theta[x_{t_{\text{U}},d_i+b} = 1] = \frac{\mathbb{E}_\theta[T_{i,b}]}{T}$. Then,

$$\begin{aligned} R_T(\mathcal{A}, \theta) &= \sum_{t=1}^{T} \sum_{i=1}^{d/m} \left(\Delta - \Delta \mathbb{I}\{\text{ind}_i(\theta) = \text{ind}_i(x_t)\} \right) \\ &= \Delta \sum_{i=1}^{d/m} \left(T - \mathbb{E}_\theta \left[T_{i,\text{ind}_i(\theta)} \right] \right) \end{aligned}$$

$$= \Delta T \sum_{i=1}^{d/m} \left(1 - \mathbb{P}_\theta[x_{t_{\mathrm{U}}, d_i + \mathrm{ind}_i(\theta)} = 1]\right).$$

For fixed $\theta \in \Theta$ and $i \in [\frac{d}{m}]$, and any $b \in [m]$, we define $\theta^{(b)} \in \Theta$ to have entries given by $\theta_j^{(b)} = \begin{cases} \Delta + \Delta \mathbb{I}\{j = d_i + b\} & j \in [d_i + 1, d_{i+1}] \\ \theta_j & \text{otherwise} \end{cases}$; and define the base parameter $\theta^{(0)}$ with entries $\theta_j^{(0)} = \begin{cases} \Delta & j \in [d_i + 1, d_{i+1}] \\ \theta_j & \text{otherwise} \end{cases}$. Note that $\theta^{(\mathrm{ind}_i(\theta))} = \theta$, and that the dependence of $\theta^{(b)}$ on i is left implicit.

Then, for $b \in [m]$, we have

$$\mathbb{P}_{\theta^{(b)}}[x_{t, d_i + b} = 1] \le \mathbb{P}_{\theta^{(0)}}[x_{t, d_i + b} = 1] + \sqrt{\frac{1}{2} \mathrm{KL}(\mathbb{P}_{\theta^{(0)}} \| \mathbb{P}_{\theta^{(b)}})} \qquad \text{(Pinsker's Inequality)}$$

$$= \mathbb{P}_{\theta^{(0)}}[x_{t, d_i + b} = 1] + \sqrt{\frac{1}{2} \mathbb{E}_{\theta^{(0)}} \left[\sum_{t=1}^{T} \mathrm{KL}\left(\mathrm{Ber}(\gamma^{\frac{1}{\epsilon}} {\theta^{(0)}}^\top x_t) \| \mathrm{Ber}(\gamma^{\frac{1}{\epsilon}} {\theta^{(b)}}^\top x_t) \right) \right]}. \qquad \text{(Chain rule)}$$

Similarly to the proof of Theorem 1, applying $\mathrm{KL}(\mathrm{Ber}(p) \| \mathrm{Ber}(q)) \le \frac{(p-q)^2}{q(1-q)}$ along with $\Delta d/m \le \theta^\top x \le 2\Delta d/m$ and $|(\theta^{(0)} - \theta^{(b)})^\top x| \le \Delta$ gives

$$\mathrm{KL}\left(\mathrm{Ber}(\gamma^{\frac{1}{\epsilon}} {\theta^{(0)}}^\top x_t) \| \mathrm{Ber}(\gamma^{\frac{1}{\epsilon}} {\theta^{(b)}}^\top x_t) \right) \le \frac{2(\gamma^{\frac{1}{\epsilon}} (\theta^{(0)} - \theta^{(b)})^\top x_t)^2}{\gamma^{\frac{1}{\epsilon}} {\theta^{(b)}}^\top x_t}$$

$$\le \frac{2^{\frac{2+\epsilon}{\epsilon}} \Delta^{\frac{2}{\epsilon}} (\frac{d}{m})^{\frac{2}{\epsilon}} \Delta^2 \mathbb{I}\{x_{t, d_i + b} = 1\}}{2^{\frac{1}{\epsilon}} \Delta^{\frac{1+\epsilon}{\epsilon}} (\frac{d}{m})^{\frac{1+\epsilon}{\epsilon}}} = 2^{\frac{1+\epsilon}{\epsilon}} \Delta^{\frac{1+\epsilon}{\epsilon}} \left(\frac{d}{m}\right)^{\frac{1-\epsilon}{\epsilon}} \mathbb{I}\{x_{t, d_i + b} = 1\}.$$

We set $\Delta := \frac{1}{8} \left(\frac{d}{m}\right)^{\frac{\epsilon-1}{1+\epsilon}} \left(\frac{T}{m}\right)^{\frac{-\epsilon}{1+\epsilon}}$. We claim that under this choice, the condition $T \ge 4^{\frac{1+\epsilon}{\epsilon}} d^{\frac{1+\epsilon}{\epsilon}}$ implies $\Delta \le \min(\frac{m}{4d}, \frac{1}{4\sqrt{d}})$, as we required earlier. To see this, we rewrite $\Delta = \frac{1}{8} d^{\frac{\epsilon-1}{1+\epsilon}} m^{\frac{1}{1+\epsilon}} T^{-\frac{\epsilon}{1+\epsilon}}$ and substitute the bound on T to obtain $\Delta \le \frac{1}{32} d^{\frac{\epsilon-1}{1+\epsilon}} m^{\frac{1}{1+\epsilon}} d^{-1}$. Dividing both sides by m gives $\frac{\Delta}{m} \le \frac{1}{32d}$, whereas applying $m \le d$ gives $\Delta \le \frac{1}{32} d^{-\frac{1}{1+\epsilon}} \le \frac{1}{32\sqrt{d}}$.

Combining the preceding two display equations and averaging over all $b \in m$, we have

$$\frac{1}{m} \sum_b \mathbb{P}_{\theta^{(b)}}[x_{t, d_i + b} = 1] \le \frac{1}{m} + \frac{1}{m} \sum_b \sqrt{2^{\frac{1+\epsilon}{\epsilon}} \Delta^{\frac{1+\epsilon}{\epsilon}} \left(\frac{d}{m}\right)^{\frac{1-\epsilon}{\epsilon}} \mathbb{E}_{\theta^{(0)}}[T_{i,b}]}$$

$$\le \frac{1}{m} + \sqrt{2^{\frac{1+\epsilon}{\epsilon}} \frac{1}{m} \Delta^{\frac{1+\epsilon}{\epsilon}} \left(\frac{d}{m}\right)^{\frac{1-\epsilon}{\epsilon}} \sum_b \mathbb{E}_{\theta^{(0)}}[T_{i,b}]} \le \frac{1}{m} + \frac{1}{2}.$$

(Jensen, $\sum_b T_{i,b} = T$ & choice of Δ)

Averaging over all $\theta \in \Theta$, summing over $i \in [d/m]$, and recalling that $m \ge 4$, we obtain

$$\frac{1}{|\Theta|} \sum_{\theta \in \Theta} \sum_{i=1}^{d/m} \left(1 - \mathbb{P}_\theta[x_{t, d_i + \mathrm{ind}_i(\theta)} = 1]\right) \ge \frac{d}{m}\left(1 - \frac{1}{m} - \frac{1}{2}\right) \ge \frac{d}{4m}.$$

Hence, there exists $\theta^* \in \Theta$ such that $\sum_{i=1}^{d/m} \left(1 - \mathbb{P}_{\theta^*}[x_{t, d_i + \mathrm{ind}_i(\theta^*)} = 1]\right) \ge \frac{d}{4m}$. Substituting into our earlier lower bound on R_T and again using our choice of Δ, we obtain

$$R_T(\mathcal{A}, \theta^*) \ge \frac{d}{4m} \Delta T = \frac{1}{32} d^{\frac{\epsilon}{1+\epsilon}} \left(\frac{d}{m}\right)^{\frac{\epsilon}{1+\epsilon}} T^{\frac{1}{1+\epsilon}}.$$

Since $f(x) = \frac{x}{\log x}$ is increasing for $x \geq e$, and $m \in [4, d]$, the definition of m gives the following:

$$\frac{d}{\log n} > \frac{m-1}{\log(m-1)} > \frac{m-1}{\log m} \geq \frac{m-1}{\log d}.$$

Rearranging the above, we obtain $\frac{d}{m} > \frac{\log n}{\log d}\left(1 - \frac{1}{m}\right) \geq \frac{\log n}{2 \log d}$, completing the proof.