
Norm-Agnostic Linear Bandits

Spencer (Brady) Gales
University of Arizona

Sunder Sethuraman
University of Arizona

Kwang-Sung Jun
University of Arizona

Abstract

Linear bandits have a wide variety of applications including recommendation systems yet they make one strong assumption: the algorithms must know an upper bound S on the norm of the unknown parameter θ^* that governs the reward generation. Such an assumption forces the practitioner to guess S involved in the confidence bound, leaving no choice but to wish that $\|\theta^*\| \leq S$ is true to guarantee that the regret will be low. In this paper, we propose novel algorithms that do not require such knowledge for the first time. Specifically, we propose two algorithms and analyze their regret bounds: one for the changing arm set setting and the other for the fixed arm set setting. Our regret bound for the former shows that the price of not knowing S does not affect the leading term in the regret bound and inflates only the lower order term. For the latter, we do not pay any price in the regret for now knowing S . Our numerical experiments show standard algorithms assuming knowledge of S can fail catastrophically when $\|\theta^*\| \leq S$ is not true whereas our algorithms enjoy low regret.

1 INTRODUCTION

Linear bandits (Abe and Long, 1999) have gained popularity since their success in online news recommendation systems (Li et al., 2010) that learn from user feedback interactively. Specifically, at each time step t , the learner chooses an arm $x_t \in \mathbb{R}^d$ from an available pool of arms $\mathcal{X}_t \subset \{x \in \mathbb{R}^d : \|x\| \leq 1\}$ given from the environment, and then receives a reward $y_t \in \mathbb{R}$ that is assumed to follow a linear model $y_t = x_t^\top \theta^* + \eta_t$, where

θ_* is unknown and η is a zero-mean subGaussian noise. This problem is an inductive extension of the classic multi-armed bandit problem (Thompson, 1933) where the learner faces the dilemma between exploration and exploitation. The goal of the algorithm is to minimize the (pseudo-)regret:

$$\text{Reg}_T = \sum_{t=1}^T \max_{x \in \mathcal{X}_t} \langle x, \theta^* \rangle - \langle x_t, \theta^* \rangle. \quad (1)$$

We refer to Lattimore and Szepesvári (2020) for a comprehensive review of bandit algorithms.

Despite the popularity of linear bandits, there is one unrealistic yet commonly-used assumption: the norm of the unknown vector θ^* is bounded by 1 or by an a priori known constant S . Such information is never known in practice. This places the practitioner in an undesirable situation where they must make a guess on the upper bound S of $\|\theta^*\|$ and take a leap of faith that it will be correct. Otherwise, the algorithm does not enjoy any theoretical guarantee. This is not just a theoretical issue as we show the failure of OFUL numerically in our experiments; see Figure 1(c). Such weakness of existing linear bandits raises an interesting question: Can we develop linear bandit algorithms that enjoy low regret without knowledge of S ?

In this paper, we answer this question in the affirmative by proposing two algorithms: NAOFUL (Norm-Agnostic Optimism in the Face of Uncertainty for Linear bandits) for the changing arm set setting, and OLSOFUL (Ordinary Least Squares OFUL) for the fixed arm set setting (i.e., $\mathcal{X}_t = \mathcal{X}$ for some \mathcal{X} , $\forall t$). We show high probability regret bounds up to the time horizon T in Table 1 where NAOFUL also has a free parameter α allowing to trade-off between the leading and the lower-order terms. Taking an absolute constant for α (e.g., 2), both of our algorithms achieve the same order of regret bound, as far as the leading \sqrt{T} term is concerned. For the lower order term (i.e., scaling poly-logarithmically with T), NAOFUL has a slightly larger dependence on $\|\theta^*\|$ in the lower-order term while OLSOFUL achieves the same order of regret bound but can only be used in the fixed arm set setting. This indicates that the price of not knowing anything about

$\|\theta^*\|$ does not affect the order of the leading term in these regret bounds.

Our algorithms are both simple and computationally efficient. Particularly, NAOFUL can be viewed as an *embarrassingly simple* remedy for making OFUL adapt to the unknown norm. Both NAOFUL and OLSOFUL can be implemented in time complexity $\mathcal{O}(d^2|\mathcal{X}_t|+o(1))$ per iteration when the arm set is finite, which is just as efficient as OFUL. Furthermore, OLSOFUL does not use any regularizer, which means that we do not have to regularize the intercept term and OLSOFUL is able to adapt to an arbitrary additive shift in the reward. In contrast, such a shift may result in invalidating OFUL’s regret bound as it has to regularize the bias term as well. We present NAOFUL and OLSOFUL along with their regret bounds in Section 4 and 5 respectively.

We perform numerical evaluations to show the efficacy of our methods compared to existing algorithms in Section 6 and conclude our paper with future directions in Section 7.

2 RELATED WORK

For many online learning problems with bandit feedback, it has been common to assume the knowledge of S such that $\|\theta^*\| \leq S$ (Abe and Long, 1999; Auer, 2002; Abbasi-Yadkori et al., 2011). Some papers further assume that the mean rewards $\{x_t^\top \theta^*\}$ are in a known interval such as $[-1, 1]$ (Abbasi-Yadkori et al., 2011). The assumption on the norm is highly unrealistic, yet researchers mostly left them untouched except for a few studies such as Orabona and Cesa-Bianchi (2011, Section 4), Gentile and Orabona (2014, Theorem 4), and Kocák et al. (2020, Remark 9). These works, however, result in an exponential dependency on $\|\theta^*\|$ in the lower-order term of their performance bound, which is undesirable. For the fixed arm set setting, Dani et al. (2008) does not assume knowledge of S but does assume that the mean rewards are known to be in $[-1, 1]$.

One may attempt to employ model selection to adapt to the unknown norm of θ^* . (Ghosh et al., 2021) consider the K arm mixture bandit algorithm with bias. They do not consider a bound on $\|\theta^*\|$ nor expected rewards. Using a model selection approach, to estimate an upper bound on $\|\theta^*\|$, they incur $\tilde{\mathcal{O}}((1+\|\theta^*\|)(\sqrt{K}+\sqrt{d})\sqrt{T})$. (Wang et al., 2021) estimate the magnitude of a data dependent complexity term (similar to $\|\theta^*\|$) required in their deep active learning algorithm. However, these works consider quite a different problem setup from ours. Furthermore, it is unclear if the model selection approach will enjoy the same bound as ours since many recent results on model selection in bandits indicate that the cost of adaptation is nontrivial and results in

a nontrivial factor to the leading term of the regret bound (Zhu and Nowak, 2020; 2021).

Algorithms with Bayesian regret guarantees do not require the knowledge of the unknown norm of θ^* (Russo and Roy, 2014). However, these algorithms in fact require a stronger assumption than knowing the norm: they typically assume the knowledge of the prior over the unknown θ^* to enjoy a valid regret bound. We remark that our norm-agnostic algorithms still enjoy the claimed regret bound even in the Bayesian environment where θ^* is drawn from a prior distribution, without the knowledge of the prior.

In sum, our work is the first work in the literature to not assume knowledge of S for the changing arm setting nor the range of the mean rewards without introducing an exponential dependence on $\|\theta^*\|$, to our knowledge.

3 BACKGROUNDS

Let us now formally define the linear bandit problem with changing arm set. At each time step t , the learner chooses an arm $x_t \in \mathcal{X}_t \subset \mathbb{R}^d$ from an available pool of arms $\mathcal{X}_t \subset \{x \in \mathbb{R}^d : \|x\| \leq 1\}$ given from the environment and then receives a reward $y_t \in \mathbb{R}$ that we assume to have the following structure $y_t = x_t^\top \theta^* + \eta_t$ where $\theta^* \in \mathbb{R}^d$ is an unknown parameter and η_t is a zero-mean R^2 -sub-Gaussian noise conditioned on all the previous observations; i.e., $\forall v, \mathbb{E}[\exp(v\eta_t) \mid x_{1:t}, \eta_{1:t-1}] \leq \exp(v^2 R^2/2)$ where we denote z_1, \dots, z_t by $z_{1:t}$.

Assumptions on the arm sets. Our theoretical bounds will still be valid when \mathcal{X}_t depends on anything that happens up to the end of the time step $t-1$. However, such a generic setting makes an interpretability issue on the regret (Eq. (1)) since the comparator $\max_{x \in \mathcal{X}_t} x^\top \theta^*$ in the definition of regret is now dependent on the algorithm’s behavior. To avoid any confusion, we assume that, for each $t \geq 1$, the arm set \mathcal{X}_t is determined by the environment before the game begins and revealed to the learner at the beginning of the time step t . Note that the fixed arm set setting is a simpler setup where $\mathcal{X}_t = \mathcal{X}, \forall t$ and \mathcal{X} is known to the learner before the game starts.

Notations. Define $S_* = \|\theta^*\|$. The order symbol $\tilde{\mathcal{O}}$ ignores polylogarithmic factors. For the matrix $A \in \mathbb{R}^{d \times d}$, define $|A|$ to be the determinant of A . For the set $\mathcal{X} \subset \mathbb{R}^d$, define $|\mathcal{X}|$ to be the cardinality of \mathcal{X} .

OFUL. One of the most popular and performant linear bandit algorithm is OFUL (Abbasi-Yadkori et al., 2011). We present OFUL in Algorithm 1 where

$$V_t(\lambda) = \sum_{s=1}^t x_s x_s^\top + \lambda I. \quad (2)$$

| Algorithms | Norm-agnostic | Regret bound | Arm set |
|--------------------|---------------|--|----------|
| OFUL | ✗ | $dR\sqrt{T} + S_*d$ | Changing |
| NAOFUL(α) | ✓ | $\alpha dR\sqrt{T} + \alpha dS_* + S_*^{2/\alpha+1}$ | Changing |
| OLSOFUL | ✓ | $dR\sqrt{T} + S_*d$ | Fixed |

Table 1: Comparison of OFUL and our algorithms. The regret bounds are expressed order-wise only, ignoring polylogarithmic factors. The regret bound of OFUL here is based on an improved analysis that was not reported in the literature explicitly, which we report in our appendix. For NAOFUL(α), we choose a polynomially decaying regularizer (see Corollary 4(ii)).

OFUL maintains a regularized MLE

$$\hat{\theta}_t = \arg \min_{\theta} \sum_{s=1}^t \frac{1}{2} (y_s - x_s^\top \theta)^2 + \frac{\lambda}{2} \|\theta\|^2 \quad (3)$$

and a confidence set around it. OFUL then chooses an arm by the celebrated optimism principle (Eq. (5)) where we choose an arm that maximizes the upper confidence bound on its mean reward. The original guarantee of OFUL along with proper tuning of λ is as follows:

Theorem 1 (Theorem 3 of (Abbasi-Yadkori et al., 2011)). *Assume that the hyperparameter S of OFUL (Algorithm 1) satisfies $S_* \leq S$. If $\langle x, \theta^* \rangle \in [-1, 1]$, $\forall x \in \mathcal{X}_t, t \in [T]$, then with $\lambda = \frac{R}{S^2}$, OFUL satisfies, w.p. at least $1 - \delta$,*

$$\text{Reg}_T = \tilde{O}(dR\sqrt{T}). \quad (4)$$

Throughout the paper, all the proofs are deferred to the appendix unless noted otherwise. We remark that we hide the dependency on δ to reduce clutter but all the regret bounds in our paper have a multiplicative factor of $\sqrt{\log(1/\delta)}$ in the leading term.

However, Theorem 1 is in fact missing an additive lower order term $\Omega(dS^*)$, which we elaborate in Appendix A along with a minor error in Abbasi-Yadkori et al. (2011). The key idea is that even if there is no noise (i.e., $R = 0$), there exists a problem where OFUL is forced to pulling $d - 1$ suboptimal arms for the first $d - 1$ iterations, incurring S^* instantaneous regret at each time step.

Furthermore, the assumption in Theorem 1 that $\langle x, \theta^* \rangle \in [-1, 1]$ is restrictive and hard to verify in practice. One can drop such an assumption and prove a regret bound using the same proof structure as for OFUL, giving a regret bound of $\tilde{O}((1 + S_*)dR\sqrt{T})$ in terms of an extra factor of $1 + S_*$ which seems restrictive. In fact, there exists a novel technique sketched in Lattimore and Szepesvári (2020, Exercise 19.3) that removes such a factor.

Theorem 2. *Assume that the hyperparameter S of OFUL (Algorithm 1) satisfies $S_* \leq S$. Let $\lambda = \frac{R}{S^2}$. Then, OFUL satisfies, w.p. at least $1 - \delta$,*

$$\text{Reg}_T = \tilde{O}(dR\sqrt{T} + S_*d)$$

Algorithm 1 OFUL (Abbasi-Yadkori et al., 2011)

- 1: **Input:** $S > 0, \lambda > 0$
- 2: **for** $t = 1, 2, \dots$ **do**
- 3: Compute $\hat{\theta}_{t-1}$ by Eq. (3).
- 4: $\sqrt{\beta_{t-1}(\lambda)} := R\sqrt{\log\left(\frac{|V_{t-1}(\lambda)|}{|\lambda I| \delta^2}\right)} + \sqrt{\lambda}S$.
- 5: Pull

$$x_t = \arg \max_{x \in \mathcal{X}_t} \max_{\theta \in C_{t-1}} \langle x, \theta \rangle \quad (5)$$

where

$$C_{t-1} = \left\{ \theta \in \mathbb{R}^d : \|\hat{\theta}_{t-1} - \theta\|_{V_{t-1}(\lambda)} \leq \sqrt{\beta_{t-1}(\lambda)} \right\}.$$

- 6: Receive reward $y_t \in \mathbb{R}^d$.
 - 7: **end for**
-

where the dependence on S is logarithmic and thus omitted.

This matches the lower bound (Dani et al., 2008), as far as the leading term is concerned. The lower order term of S_*d is not restrictive – one can extend the argument described in Appendix A to any algorithm. That is, one can show that there exists a set of problems where any algorithm needs to make at least $\Theta(d)$ suboptimal arm pulls before identifying the best arm in one of those problems where each suboptimal arm pulls incur an instantaneous regret of S_* .

4 THE CHANGING ARM SET

To lift the common assumption of knowing a bound S on $\|\theta^*\|$, we propose a new algorithm called NAOFUL (Norm Agnostic OFUL) described in Algorithm 2. Our algorithm makes two simple adjustments from OFUL: introduction of (i) time dependent regularizer λ_t and (ii) a novel confidence width $\sqrt{\beta_t}$. We emphasize that $\sqrt{\beta_t}$ does not involve S or S_* , which means that the algorithm does not require any knowledge on θ^* . One can show that to obtain a valid version of the confidence ellipsoid for θ^* at any time step t , one needs to use

Algorithm 2 NAOFUL (Norm-Agnostic OFUL)

- 1: **Input:** a nonincreasing sequence $\{\lambda_t > 0\}_{t \geq 1}$
- 2: Let $\delta_t = \frac{1}{t^2} \frac{6}{\pi^2} > 0, \forall t \geq 1$.
- 3: **for** $t = 1, 2, \dots$ **do**
- 4: Compute $\hat{\theta}_{t-1}$ by Eq. (3).
- 5: $\sqrt{\beta_{t-1}} := R \sqrt{\log \left(\frac{|V_{t-1}(\lambda_t)|}{|\lambda_t| \delta_t^2} \right)} + R\sqrt{d}$
- 6: Pull $x_t = \arg \max_{x \in \mathcal{X}_t} \max_{\theta \in \bar{C}_{t-1}} \langle x, \theta \rangle$ where

$$\bar{C}_{t-1} = \left\{ \theta \in \mathbb{R}^d : \|\hat{\theta}_{t-1} - \theta\|_{V_{t-1}(\lambda_t)} \leq \sqrt{\beta_{t-1}} \right\}.$$
- 7: Receive reward $y_t \in \mathbb{R}^d$.
- 8: **end for**

$\sqrt{\lambda_t} S_*$ in place of $R\sqrt{d}$ in the definition of $\sqrt{\beta_{t-1}}$. The key idea is that there exists a critical time step t_0 after which $\sqrt{\lambda_t} S_* \leq R\sqrt{d}$ becomes true so that the confidence ellipsoid is true. This leaves us to analyze the regret for a time-varying regularizer and the regret before the time step t_0 . NAOFUL enjoys the following regret bound.

Theorem 3. *Suppose we run NAOFUL (Algorithm 2) with a nonincreasing sequence $\{\lambda_t > 0\}_{t \geq 1}$. Let $t_0 := \min\{t \geq 1 : \lambda_t^{1/2} S_* \leq R\sqrt{d}\}$. Then, for any $T \geq 1$, with probability at least $1 - \delta$,*

$$\text{Reg}_T = \mathcal{O} \left(t_0 S_* + R\sqrt{T} \cdot d \log \left(\frac{T}{\lambda_T} \right) + S_* d \log \left(\frac{1}{\lambda_T} \right) \right)$$

Theorem 3 is stated for a general sequence $\{\lambda_t\}$. In the corollary below, we provide a few choices.

Corollary 4. *Under the same assumption as Theorem 3,*

(i) *If $\lambda_t = \frac{R^2 d}{\log^\gamma(1+t)}$ for $\gamma > 0$, w.p. at least $1 - \delta$,*

$$\text{Reg}_T = \tilde{\mathcal{O}}(Rd\sqrt{T} + S_*(d + \exp(S_*^{2/\gamma})))$$

(ii) *If $\lambda_t = \frac{R^2 d}{t^\alpha}$ for $\alpha \geq 0$, w.p. at least $1 - \delta$,*

$$\text{Reg}_T = \tilde{\mathcal{O}}(Rd\sqrt{(\alpha^2 + 1)T} + S_*(d\alpha + S_*^{2/\alpha}))$$

(iii) *If $\lambda_t = \frac{R^2 d}{\exp(t^q)}$ for $q \in (0, 1)$, w.p. at least $1 - \delta$,*

$$\text{Reg}_T = \tilde{\mathcal{O}}(Rd\sqrt{T^{1+q}} + S_*(dT^q + \log^{1/q}(S_*^2)))$$

Various choices of $\{\lambda_t\}$ exhibit different tradeoffs between the leading term (involving \sqrt{T}) and the lower order term (involving S_*) in the regret bound. If the practitioner knows a rough range of S_* and the time

horizon T , one can make an informed choice accordingly. Specifically, if $R\sqrt{T}$ is much larger than S_* then the rate $\lambda_t = \mathcal{O}(1/\log^\gamma(t))$ is preferred whereas if $S_* \gg RT^{1-(q/2)}$ then the rate $\lambda_t = \exp(-t^q)$ is preferred. If there is absolutely no information on S_* , then the polynomial decay $R^2 d/t^\alpha$ seems most reasonable because it does not introduce an exponential dependence on S_* and does not change the order w.r.t. T .

Computational complexity. Note that when the arm set is finite, OFUL can be implemented with $\mathcal{O}(d^2|\mathcal{X}_t|)$ time complexity per time step by updating $V_t^{-1}(\lambda)$ using the matrix inversion lemma and $\log(|V_t(\lambda)|) = \log(|V_{t-1}(\lambda)|) + \log(1 + \|x_t\|_{V_{t-1}(\lambda)}^2)$. Such updates are nontrivial for NAOFUL since it has a time-varying λ_t . However, for the polynomial decay of $\lambda_t = \frac{R^2 d}{t^\alpha}$, one can achieve a similar decay by a piecewise-linear schedule of $\lambda'_t = R^2 d \exp(-[\alpha \log(t)])$ without changing the regret guarantee. This way, we can make the same efficient updates as OFUL while λ'_t does not change. Up to time T , the regularizer λ'_t changes only $\mathcal{O}(\log(T))$ times, and when such a change happens we pay $\mathcal{O}(d^3)$ for recomputing the inversion. Therefore, the average per-time-step time complexity is $\mathcal{O}(d^2|\mathcal{X}_t| + d^3 \log(T)/T)$ where the latter term disappears quickly as T gets large, making NAOFUL as efficient as OFUL.

Remark 1. We remark that Theorem 3 can be further tightened by replacing the definition of t_0 to $t'_0 := \min\{t \geq 1 : \lambda\|\theta^*\|_{V_t(\lambda_t)^{-1}} \leq R\sqrt{d}\}$. However, such a bound involves an algorithm-dependent quantity $V_t(\lambda_t)$ and does not provide further insight. We speculate that it might be possible to bound t'_0 by a function of the sequence of the arm sets $\{\mathcal{X}_t\}$ and improve the bound in Theorem 3. Such an investigation is beyond the scope of our paper, which we leave as future work.

4.1 Proof of Theorem 3

We provide a proof sketch here and defer the full proof to our appendix. The proof of NAOFUL differs from that of Abbasi-Yadkori et al. (2011) in two ways. First, we separate the analysis by whether the confidence bound is true at time t or not. Second, we leverage the result of Lattimore and Szepesvári (2020, Exercise 19.3), which we call the *elliptical potential count* lemma and present its improved and generalized version in the appendix. This way, we do not introduce a factor of $1 \vee S_*$ in the leading term of the regret when dropping the assumption of Abbasi-Yadkori et al. (2011) that $\langle x_t, \theta^* \rangle \in [-1, 1], \forall t$.

Let $r_t := \max_{x \in \mathcal{X}_t} \langle x, \theta^* \rangle - \langle x_t, \theta^* \rangle$ and $t_0 = \min\{t : \lambda_t^{1/2} S_* \leq R\sqrt{d}\}$. Then, using Abbasi-Yadkori et al. (2011, Theorem 2) with a union bound over $t \geq 1$, the confidence bound at $t \geq t_0$ is true; see Theorem 6 in

our appendix. Then,

$$\begin{aligned} \text{Reg}_T &= \sum_{t=1}^{t_0-1} r_t + \sum_{t=t_0}^T r_t \\ &\leq t_0 2S_* + \sum_{t=t_0}^T r_t \quad (\|x\| \leq 1, \forall x \in \mathcal{X}_t, \forall t) \end{aligned}$$

It remains to bound the second term. Define $H_T := \{t \in \{t_0, \dots, T\} : \|x_t\|_{V_{t-1}^{-1}(\lambda_T)}^2 > 1\}$ and $\bar{H}_T := \{t_0, \dots, T\} \setminus H_T$. Then,

$$\begin{aligned} &\sum_{t=t_0}^T r_t \\ &= \sum_{t=t_0}^T r_t \mathbf{1}\{\|x_t\|_{V_{t-1}^{-1}(\lambda_T)}^2 \leq 1\} + \sum_{t=t_0}^T r_t \mathbf{1}\{\|x_t\|_{V_{t-1}^{-1}(\lambda_T)}^2 > 1\} \\ &\leq \sum_{t=t_0}^T r_t \mathbf{1}\{\|x_t\|_{V_{t-1}^{-1}(\lambda_T)}^2 \leq 1\} + 2S_* |H_T| \\ &\leq \sqrt{|\bar{H}_T| \sum_{t \in \bar{H}_T} r_t^2} + 2S_* |H_T| \end{aligned} \quad (7)$$

where the last line is due to Cauchy-Schwarz. The standard analysis of linear bandits (Abbasi-Yadkori et al., 2011) implies that (i) $r_t \leq 2\sqrt{\beta_{t-1}} \|x_t\|_{V_{t-1}^{-1}(\lambda_t)} \leq 2\sqrt{\beta_{t-1}} \|x_t\|_{V_{t-1}^{-1}(\lambda_T)}$ and (ii) $\sqrt{\beta_{t-1}} \leq R\sqrt{d \log(1 + \frac{t-1}{d\lambda_t})} + 2\log(\frac{1}{\delta_t}) + R\sqrt{d} = \mathcal{O}(R\sqrt{d \log(T(1 + \lambda_T^{-1}))})$ where the last inequality holds as $t \leq T$ and both $\{\lambda_t\}$ and $\{\delta_t\}$ are nonincreasing. (Recall that we ignore the dependence on δ in the order notation.) Then,

$$\begin{aligned} \sum_{t \in \bar{H}_T} r_t^2 &\leq \sum_{t \in \bar{H}_T} 4\bar{\beta}_{t-1} \|x_t\|_{V_{t-1}^{-1}(\lambda_T)}^2 \\ &\leq \mathcal{O}(R^2 d \log(T(1 + \lambda_T^{-1}))) \cdot \sum_{t \in \bar{H}_T} \|x_t\|_{V_{t-1}^{-1}(\lambda_T)}^2 \\ &= \mathcal{O}\left((R^2 d \log(T(1 + \lambda_T^{-1}))) \cdot d \log(1 + \frac{T}{d\lambda_T})\right) \end{aligned}$$

where the last inequality is due to the standard elliptical potential lemma (e.g., Abbasi-Yadkori et al. (2011, Lemma 11)). It remains to bound $2S_* |H_T|$. The elliptical potential count lemma (Lemma 7 in our appendix) implies that $|H_T| = \mathcal{O}(d \log(1 + \frac{1}{\lambda_T}))$. Applying $|\bar{H}_T| \leq T$ to Eq. (7) concludes the proof.

5 THE FIXED ARM SET

We now consider the fixed arm set case, a special case of changing arm set where $\mathcal{X}_t = \mathcal{X}, \forall t \geq 1$, for some \mathcal{X} . In this setting, the algorithm knows \mathcal{X} from the beginning, which means that it is possible to plan in an informed way so that we can estimate the mean reward of each arm in a stable manner. Motivated by

Algorithm 3 OLSOFUL

Input: Arm set $\mathcal{X} \subset \mathbb{R}^d$

Obtain $\mathcal{X}_0 := \{x_1, \dots, x_{|\mathcal{X}_0|}\}$ by KY Sampling (Algorithm 5).

$t \leftarrow |\mathcal{X}_0| + 1$

while $\max_{x \in \mathcal{X}} \|x\|_{\bar{V}_{t-1}}^2 > 1$ **do**

$x_t = \arg \max_{x \in \mathcal{X}} \|x\|_{\bar{V}_{t-1}}^2$.

Pull the arm x_t and receive reward $y_t \in \mathbb{R}$.

$t \leftarrow t + 1$

end while

$\tau \leftarrow t - 1$

for $t = \tau + 1, \dots$ **do**

Compute the MLE $\hat{\theta}_{t-1,0}$ by Eq. (3) with $\lambda = 0$.

$$\sqrt{\beta_{t-1}^{\text{OLS}}} := R \left(\sqrt{\log\left(\frac{|\bar{V}_{t-1}|}{|\bar{V}_\tau| \delta^2}\right)} + \sqrt{\log(4)d + 4 \log\left(\frac{1}{\delta}\right)} \right)$$

Pull $x_t = \arg \max_{x \in \mathcal{X}} \max_{\theta \in C_{t-1}} \langle x, \theta \rangle$ where

$$C_{t-1} = \left\{ \theta \in \mathbb{R}^d : \|\hat{\theta}_{t-1,0} - \theta\|_{\bar{V}_{t-1}} \leq \sqrt{\beta_{t-1}^{\text{OLS}}} \right\}.$$

Receive reward $y_t \in \mathbb{R}^d$.

end for

this observation, we propose a new algorithm OLSOFUL (Ordinary Least Squares OFUL) that performs a two-step warmup procedure in order to establish good estimates of the mean rewards before starting an OFUL-like arm selection strategy.

The full pseudocode of OLSOFUL can be found in Algorithm 3. OLSOFUL first obtain a set of arms $\mathcal{X}_0 := \{x_1, \dots, x_{|\mathcal{X}_0|}\}$ from KY sampling described in Algorithm 4. Let $\bar{V}_t = \sum_{s=1}^t x_s x_s^\top$. KY sampling selects at most $2d$ arms from \mathcal{X} so that the determinant of the matrix $\bar{V}_{|\mathcal{X}_0|}$ is sufficiently large (details in Section 5.1). Let $\bar{V}_t := \sum_{s=1}^t x_s x_s^\top$. The second warmup sampling simply samples arms that maximize $\|x\|_{\bar{V}_{t-1}}^2$ until $\|x\|_{\bar{V}_{t-1}}^2 \leq 1$ for every $x \in \mathcal{X}$. The second sampling is an auxiliary device that allows the regret analysis, although we speculate that it may not be necessary. For the rest of the time steps, we perform the usual OFUL-like sampling but now with the MLE estimator $\hat{\theta}_{t-1,0}$ computed by Eq. (3) with $\lambda = 0$, which does not use a regularizer. The confidence radius $\sqrt{\beta_{t-1}}$ is thus adjusted accordingly. The difference is that we no longer have the usual $\sqrt{\lambda} S_*$ term but rather a term of $\mathcal{O}(\sqrt{d + \log(1/\delta)})$, which makes it norm-agnostic.

OLSOFUL enjoys the following regret guarantee.

Theorem 5. *Suppose we run OLSOFUL (Algo-*

Algorithm 4 BH Algorithm (Betke and Henk, 1993; Kumar and Yildirim, 2005)

Input: the set of points $\mathcal{X} \subset \mathbb{R}^d$
 If $|\mathcal{X}| \leq 2d$, then $\mathcal{X}_0 \leftarrow \mathcal{X}$. Return.
 $\Psi \leftarrow 0$, $\mathcal{X}_0 \leftarrow \emptyset$, $i \leftarrow 0$.
while $\mathbb{R}^d \setminus \Psi \neq \emptyset$ **do**
 $i \leftarrow i + 1$
 Pick an arbitrary direction $b^i \in \mathbb{R}^d$ in the orthogonal complement of Ψ .
 $p \leftarrow \arg \max_{x \in \mathcal{X}} \langle b^i, x \rangle$; $\mathcal{X}_0 \leftarrow \mathcal{X}_0 \cup \{p\}$
 $q \leftarrow \arg \min_{x \in \mathcal{X}} \langle b^i, x \rangle$; $\mathcal{X}_0 \leftarrow \mathcal{X}_0 \cup \{q\}$
 $\Psi \leftarrow \text{Span}(\Psi, \{p - q\})$
end while

rithm 3). Then, with probability at least $1 - 2\delta$,

$$\text{Reg}_T = \mathcal{O} \left(Rd \sqrt{T \log^2 \left(\frac{T}{d} \right)} + S_* d \log(d) \right).$$

It is easy to see that the regret of OLSOFUL matches that of OFUL (Theorem 1). This implies that there is no cost of not knowing the norm of θ^* in the worst-case regret scenario. Note that KY sampling can be implemented so the overall time complexity is $\mathcal{O}(d^2|\mathcal{X}|)$, meaning that the time complexity of OLSOFUL is at most that of OFUL. In this respect, we present a computationally efficient version of KY sampling in the appendix. Furthermore, OLSOFUL allows us not to regularize the intercept term, which is a standard practice in statistics and ML for supervised learning. Thus, the rewards can be shifted by an arbitrary but fixed constant and the regret bound does not change. Such a property is not available to OFUL and other existing algorithms, except for elimination algorithms (Lattimore and Szepesvári, 2020, Section 21). While, elimination algorithms have superior regret bounds when $|\mathcal{X}| = o(e^d)$, in practical regimes they have large regret due to its unique behavior of discarding samples.

Remark 2. One can replace our two-step warmup procedure with a design of experiment approach (specifically G-optimal design) and obtain a similar regret bound. That is, solve $\min_{\pi \in \Delta^{|\mathcal{X}|}} \max_x \|x\|_{\bar{V}(\pi)}^2$ where $\bar{V}(\pi) = \sum_{x \in \mathcal{X}} \pi_x x x^\top$ and use a rounding procedure to select samples in a way that the fraction of samples spent on each arm is approximately π_x ; see Fiez et al. (2019, Appendix B) for a discussion of rounding procedures. However, such an approach is computationally more expensive than ours as it requires iterative algorithms for solving the optimization problem. In contrast, our simple warm-up procedure consists of closed-form expressions and is computationally efficient. Alternatively, one can use the approximate barycentric spanner (Awerbuch and Kleinberg, 2004) that consists

of closed form operations but its time complexity is at least $\Omega(d^3 \log(d)|\mathcal{X}|)$ compared to $\mathcal{O}(d^2|\mathcal{X}|)$ of our efficient implementation in the appendix.

5.1 Proof of Theorem 5

We now provide a sketch of the proof, leaving the full proof to the appendix. The proof has two main parts. First, we need to find an upper bound on τ so that the regret up to time τ , which is the length of the warmup phase, can be bounded by $2S_*\tau$. Second, we need to bound the regret after τ where the main novelty is our confidence set that replaces $|\lambda I|$ in the confidence width $\sqrt{\beta_{t-1}}$ of OFUL with $|\bar{V}_\tau|$. We provide the main ideas only and leave the full proof to the appendix. Let us first describe the property of the warmup procedure. For $\pi \in \Delta^{|\mathcal{X}|}$, define

$$\bar{V}(\pi) = \sum_{x \in \mathcal{X}} \pi_x x x^\top. \quad (8)$$

First, we claim that

$$\tau = \mathcal{O}(d \log(d)).$$

To see this, let $\pi^* = \arg \min_{\pi \in \Delta^{|\mathcal{X}|}} \max_{x \in \mathcal{X}} \|x\|_{\bar{V}(\pi)}^2$, the solution of the G-optimal design problem. By Kiefer and Wolfowitz (1960), $\pi^* = \arg \max_{\pi \in \Delta^{|\mathcal{X}|}} \max_{x \in \mathcal{X}} |\bar{V}(\pi)|$ as well. This implies that $|\bar{V}_\tau| = \tau^d |\frac{1}{\tau} \bar{V}_\tau| \leq \tau^d |\bar{V}(\pi^*)|$. Let $\omega = |\mathcal{X}_0|$ be the final time step spent on the first stage of the warmup. Then, $|\bar{V}_\omega| = \omega^d |\frac{1}{\omega} \bar{V}_\omega|$. These bounds on the determinants together imply

$$\begin{aligned} & \log |\bar{V}_\tau| - \log |\bar{V}_\omega| \\ & \leq d \log(\tau) + \log(|\bar{V}(\pi^*)|) - \log \left| \frac{1}{\omega} \bar{V}_\omega \right| - d \log(\omega) \\ & \leq d \log(\tau) + \mathcal{O}(d \log(d)) - d \log(\omega) \end{aligned}$$

where the last inequality is due to Kumar and Yildirim (2005, Theorem 4.2), meaning that the first stage of the warmup allocates arms so the resulting (normalized) determinant is sufficiently close to the ideal allocation of π^* . On the other hand,

$$\begin{aligned} \log |\bar{V}_\tau| - \log |\bar{V}_\omega| &= \sum_{t=\omega+1}^{\tau} \log(1 + \|x_t\|_{\bar{V}_{t-1}}^2) \\ &> (\tau - \omega) \log(2) \end{aligned}$$

where the last inequality is due to the definition of τ . Using $\omega \leq 2d$, one can solve the sandwich bound formed by the two displays above, which concludes the proof of the claim above. This, in turn, proves that the warmup procedure results in $\mathcal{O}(S_* d \log(d))$ regret.

Let us now derive our confidence bound used after the warmup. Let $X_{\leq \tau} \in \mathbb{R}^{\tau \times d}$ ($X_{> \tau} \in \mathbb{R}^{(t-\tau) \times d}$) be the design

matrix consisting of the pulled arms during the warmup phase (after the warmup up time step t , respectively). Define $\eta_{\leq \tau} = (\eta_1, \dots, \eta_\tau)^\top$ and $\eta_{> \tau} = (\eta_{\tau+1}, \dots, \eta_t)^\top$. We first note that the MLE $\hat{\theta}_{t,0}$ satisfies

$$\bar{V}_t(\hat{\theta}_{t,0} - \theta^*) = X_{> \tau}^\top \eta_{> \tau} - \lambda \bar{V}_\tau(\theta^* - \hat{\theta}_{\tau,0})$$

Then,

$$\begin{aligned} x^\top(\hat{\theta}_{t,0} - \theta^*) \\ = x^\top \bar{V}_t^{-1} X_{> \tau}^\top \eta_{> \tau} - \lambda x^\top \bar{V}_t^{-1} \bar{V}_\tau(\theta^* - \hat{\theta}_{\tau,0}) \end{aligned}$$

With probability at least $1 - \delta$, we can bound the first term by $\|x\|_{\bar{V}_t^{-1}} \|X_{> \tau}^\top \eta_{> \tau}\|_{\bar{V}_t^{-1}} \leq \|x\|_{\bar{V}_t^{-1}} R \sqrt{\log(\frac{|\bar{V}_t|}{|\bar{V}_\tau| \delta^2})}$ for all $t \geq \tau + 1$, which is formally stated in the appendix. Such a result is a consequence of applying the standard self-normalized confidence bound of [Abbasi-Yadkori et al. \(2011, Theorem 1\)](#) where the parameter V therein for the prior is set as \bar{V}_τ . Furthermore, with probability at least $1 - \delta$, one can bound

$$\begin{aligned} -x^\top \bar{V}_t^{-1} \bar{V}_\tau(\theta^* - \hat{\theta}_{\tau,0}) &= x^\top \bar{V}_t^{-1} X_{\leq \tau}^\top \eta_{\leq \tau} \\ &\leq \|x\|_{\bar{V}_t^{-1}} \|X_{\leq \tau}^\top \eta_{\leq \tau}\|_{\bar{V}_\tau^{-1}} \\ &\leq R \sqrt{2 \log(2) d + 4 \log(1/\delta)} \end{aligned}$$

where the last inequality is a consequence of applying [Abbasi-Yadkori et al. \(2011, Theorem 1\)](#) in a similar way as above.

To bound the regret after τ , the standard analysis goes through. The only part that remains is that the usual term $\log(|V_t(\lambda)|/|\lambda I|)$ is replaced by $\log(|\bar{V}_t|/|\bar{V}_\tau|)$, which can be bounded by $\mathcal{O}(d \log(t/d))$ using the fact that $|\bar{V}_\tau|$ is sufficiently large.

6 EXPERIMENTS

In this section, validate the need for norm-agnostic algorithms and their effectiveness via a numerical evaluation in synthetic environments.

Algorithms. We test three algorithms: OFUL (Algorithm 1 with $\lambda = 1$ and $S = 1$), NAOFUL (Algorithm 2, and OFUL₀ (described below). We run two versions of NAOFUL with $\lambda_t = R^2 d/t^\alpha$: NAOFUL1 for $\alpha = 1$ and NAOFUL2 for $\alpha = 2$. OFUL₀ is OFUL without a regularization. OFUL₀ is a reasonable attempt to be norm-agnostic. Specifically, we use the confidence set from [Lattimore and Szepesvári \(2020, Eq. \(20.3\)\)](#) for the ordinary least squares estimator $\hat{\theta}_{t,0}$ that is the solution of Eq. (3) with $\lambda = 0$ while breaking ties by taking the minimum norm solution:

$$\mathbb{P} \left(\exists t \geq 1 : \|\hat{\theta}_{t,0} - \theta^*\|_{V_t(0)} \right)$$

| Inst. | a_0 | b | u | θ^* | Best arm |
|-------|------------|-----|------|-----------------|------------------------------|
| (a) | 10^{-10} | 1.5 | 1.0 | $(1, 0)^\top$ | $(1, 0)^\top$ |
| (b) | 10^{-10} | 1.5 | 1.0 | $(1, 100)^\top$ | $(1, 0)^\top$ if $t \leq 45$ |
| (c) | 10^{-10} | 1.5 | 0.05 | $(1, 100)^\top$ | $(1, 0)^\top$ if $t \leq 45$ |

Table 2: Our synthetic problem instances.

$$\geq R \sqrt{8 \left(d \log(6) + \log\left(\frac{\pi^2 t^2}{6\delta}\right) \right)} \leq \delta. \quad (9)$$

We then construct a confidence set with the constraint above to choose the optimistic arm as in Eq. (5). Such a confidence set does not involve S_* and thus is norm-agnostic like NAOFUL, but OFUL₀ does not have a known regret bound.¹

Synthetic data. To evaluate the performance of the norm-agnostic algorithm for the changing arm set setting, we consider three problem instances. We will consider a common arm set

$$\mathcal{X}_t := \{(1, 0)^\top, (0, \min\{u, a_0 \cdot b^t\})^\top\}, \forall t.$$

When arm x_t is pulled, the rewards are generated by $y_t = x_t^\top \theta^* + \eta_t$ where $\eta_t \sim \mathcal{N}(0, 1)$. For ease of exposition, we refer to the arm $(1, 0)^\top$ as A1 and $(0, \min\{u, a_0 \cdot b^t\})^\top$ as A2. We describe the choices of the instance parameters in Table 2. The intention for instance (a) in Table 2, we have $\|S_*\| \leq 1$, so OFUL's regret bound is valid, OFUL should work well whereas for instances (b) and (c) we have $\|S_*\| \gg 1$, so OFUL is expected to perform badly. We run all the algorithms up to $T = 100$ on each instance and measure their regret, which we repeat 10 times to obtain mean and standard deviation.

Results. We plot the mean and standard deviation of the regret of each algorithm in Figure 1, which shows that OFUL and OFUL₀ have almost linear regret for some instances whereas NAOFUL1 and NAOFUL2 maintain low regret for all. Specifically, for instance (a), one can see that OFUL₀ is significantly worse than the rest, showing an almost linear regret. In fact, given T , one can construct an instance where OFUL₀ suffers a linear regret up to time T by setting a_0 arbitrarily small. This is because its upper confidence bound (UCB), though valid, does not fall sufficiently low to allow the algorithm to pull the true best arm due to the fact that the second dimension of A2 is increasing geometrically

¹In fact, the proof of Eq. (9) requires that the chosen arms $\{x_1, \dots, x_t\}$ do not depend on $\{y_1, \dots, y_t\}$, which is violated in OFUL₀. While this is not ideal, OFUL₀ serves our purpose because OFUL₀'s confidence bounds were never wrong in our experiments. Alternatively, one could use OFUL with an extremely small λ (e.g., 10^{-16}) and obtain similar results, but its estimator tends to be numerically unstable.

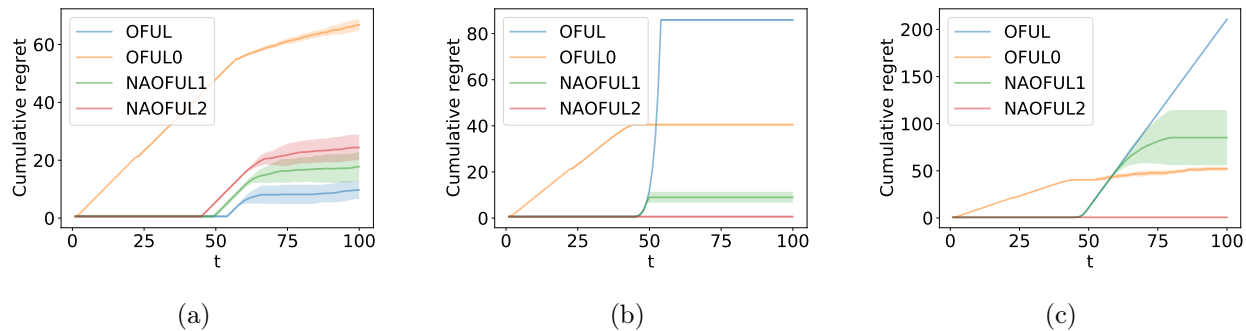


Figure 1: Experimental results. The plots (a), (b), and (c) corresponds to the instance (a), (b), and (c) in Table 2. Our methods consistently perform the best or near best in all the environments whereas OFUL and OFUL₀ can suffers an almost linear regret in some of the instances.

over time. In contrast, NAOFUL1, NAOFUL2, and OFUL are able to ignore A2 initially but start to pull it later in order to avoid the risk of A2 becoming the best arm. As noted in Table 2, instance (b) switches the best arm at $t = 46$. This forces the regret of OFUL₀, which mostly pulls A2, to stop growing, and the regret of OFUL, whose UCB on A2 is underestimated, to start exploding. In contrast, both NAOFUL1 and NAOFUL2 are able to quickly start pulling A2 when the best arm switches, and NAOFUL2 is particularly good at it. Note that OFUL, though late, is still able to pull A2 and stop growing its regret. However, the results for instance (c) shows that by adjusting u , one can force OFUL to have linear regret; what happens is that the second dimension of A2 is just large enough to make it the best arm but not large enough to force OFUL’s UCB of A2 to be the UCB of A1.

Remark 3. Readers may wonder why the regret plot for NAOFUL does not seem to follow a \sqrt{T} growth rate. The regret bound we show in this paper is the *worst-case* regret, so individual regret curves for each instance may behave differently. For instance-optimal regret bounds, we refer to Lattimore and Szepesvári (2017).

7 CONCLUSION

In this paper, we have removed the common assumption in linear bandits that we have access to an upper bound S on the norm $\|\theta^*\|$ and/or the range of the mean rewards. We believe our work has implications for problems beyond bandits because the standard linear bandit algorithm like OFUL or similar versions are abundant in online selective sampling (Dekel et al., 2010; 2012; Gentile and Orabona, 2014) and reinforcement learning (Zhou et al., 2021; Du et al., 2021).

Our work opens up numerous future research directions. First, we believe that our regret upper bound for the

changing arm set setting might be loose. It would be interesting to identify the optimal regret for norm-agnostic algorithms by finding a matching lower and upper bound on the regret. Second, extending our work to kernels would be interesting. Kernels are nontrivial to apply our techniques since their regret bounds scale with the effective dimension that in turn depends on the regularizer. Thus, using a decreasing regularizer would increase the effective dimension over time, so a careful or even data-dependent scheduling of $\{\lambda_t\}$ might be needed in order to keep the regret close to existing bounds obtained with knowledge of S . Next, it would be interesting to also remove the assumption that the arm vector’s norm is bounded by 1. Note that this is not a problem for the fixed arm set setting since one can scale the given arm set and scale their norm to be less than or equal to 1. The interesting case is the changing arm set setting where the norm of the arms may grow indefinitely over time. Finally, we believe a more extensive empirical evaluation of our algorithms on a number of differing real world data sets would help inform the risk of misspecifying S .

References

- Y. Abbasi-Yadkori, D. Pal, and C. Szepesvari. Improved Algorithms for Linear Stochastic Bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1–19, 2011.
- N. Abe and P. M. Long. Associative reinforcement learning using linear probabilistic concepts. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 3–11, 1999.
- A. Antos, V. Grover, and C. Szepesvári. Active learning in heteroscedastic noise. *Theoretical Computer Science*, 411(29-30):2712–2728, 2010.
- P. Auer. Using Confidence Bounds for Exploitation-Exploration Trade-offs. *Journal of Machine Learning Research*, 3:397–422, 2002.

- B. Awerbuch and R. D. Kleinberg. Adaptive Routing with End-to-End Feedback: Distributed Learning and Geometric Approaches. In *Proceedings of the ACM Symposium on Theory of Computing (STOC)*, pages 45–53, 2004.
- U. Betke and M. Henk. Approximating the volume of convex bodies. *Discrete & Computational Geometry*, 10(1):15–21, 1993.
- V. Dani, T. P. Hayes, and S. M. Kakade. Stochastic Linear Optimization under Bandit Feedback. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 355–366, 2008.
- O. Dekel, C. Gentile, and K. Sridharan. Robust selective sampling from single and multiple teachers. In *Proceedings of the Conference on Learning Theory (COLT)*, 2010.
- O. Dekel, C. Gentile, and K. Sridharan. Selective sampling and active learning from single and multiple teachers. *Journal of Machine Learning Research*, 13:2655–2697, 2012.
- S. S. Du, S. M. Kakade, J. D. Lee, S. Lovett, G. Mahajan, W. Sun, and R. Wang. Bilinear Classes: A Structural Framework for Provable Generalization in RL. *arXiv preprint arXiv:2103.10897*, 2021.
- T. Fiez, L. Jain, K. G. Jamieson, and L. Ratliff. Sequential Experimental Design for Transductive Linear Bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.
- C. Gentile and F. Orabona. On Multilabel Classification and Ranking with Bandit Feedback. *Journal of Machine Learning Research*, 15:2451–2487, 2014.
- A. Ghosh, A. Sankararaman, and R. Kannan. Problem-complexity adaptive model selection for stochastic linear bandits. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1396–1404. PMLR, 2021.
- J. Kiefer and J. Wolfowitz. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366, 1960.
- Y. Kim, I. Yang, and K.-S. Jun. Improved Regret Analysis for Variance-Adaptive Linear Bandits and Horizon-Free Linear Mixture MDPs, 2021.
- T. Kocák, R. Munos, B. Kveton, S. Agrawal, and M. Valko. Spectral bandits. *Journal of Machine Learning Research*, 2020.
- P. Kumar and E. A. Yildirim. Minimum-volume enclosing ellipsoids and core sets. *Journal of Optimization Theory and applications*, 126(1):1–21, 2005.
- T. Lattimore and C. Szepesvári. The end of optimism? An asymptotic analysis of finite-armed linear bandits. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.
- T. Lattimore and C. Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.
- L. Li, W. Chu, J. Langford, and R. E. Schapire. A Contextual-Bandit Approach to Personalized News Article Recommendation. *Proceedings of the International Conference on World Wide Web (WWW)*, pages 661–670, 2010.
- F. Orabona and N. Cesa-Bianchi. Better Algorithms for Selective Sampling. In *Proceedings of the International Conference on International Conference on Machine Learning (ICML)*, pages 433–440, 2011.
- D. Russo and B. V. Roy. Learning to Optimize via Posterior Sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.
- D. Russo and B. Van Roy. Eluder dimension and the sample complexity of optimistic exploration. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2256–2264, 2013.
- W. R. Thompson. On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika*, 25(3/4):285, 1933.
- Z. Wang, P. Awasthi, C. Dann, A. Sekhari, and C. Gentile. Neural active learning with performance guarantees. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
- D. Zhou, Q. Gu, and C. Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 4532–4576. PMLR, 2021.
- Y. Zhu and R. Nowak. On regret with multiple best arms. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:9050–9060, 2020.
- Y. Zhu and R. Nowak. Pareto Optimal Model Selection in Linear Bandits. *arXiv preprint arXiv:2102.06593*, 2021.

Supplementary Material: Norm-Agnostic Linear Bandits

Table of Contents

| | |
|---|-----------|
| A Minor error of Abbasi-Yadkori et al. (2011, Theorem 3) | 10 |
| B Proof of Theorem 3 | |
| B.1 Elliptical Potential Count Lemma | 13 |
| C Proof of Theorem 5 | |
| C.1 Computationally Efficient Version of Algorithm 4 | 19 |

A Minor error of Abbasi-Yadkori et al. (2011, Theorem 3)

We claim that the upper bound reported in Theorem 1 should have an additive lower order term $\Omega(dS^*)$. To see this, imagine the arm set is $\mathcal{X}_t = \{e_1, \dots, e_d\}$ where e_i is the i -th indicator vector. Suppose $\theta^* = S^* \cdot e_d$. This way, pulling a suboptimal arm incurs instantaneous regret of S^* . Assume that OFUL breaks the tie chooses an arm $i \neq i^*$, which is an allowed behavior for OFUL. Assume that there is no noise; i.e., $R = 0$, which makes $\sqrt{\beta_{t-1}} = \sqrt{\lambda}S^*$. Quick calculation shows that the UCB for the pulled arm becomes $\sqrt{\beta_{t-1}} \frac{1}{1+\lambda}$ and that for the unpulled arm becomes $\sqrt{\beta_{t-1}} \frac{1}{\lambda}$. So, OFUL will play $x_t = e_t$ up to time step d , incurring the regret of $S^*(d-1)$.

The source of such a mismatch comes from a minor error in the proof of OFUL regret bound. In the proof of Theorem 3 of Abbasi-Yadkori et al. (2011), they show that

$$r_t \leq 2 \min\{\sqrt{\beta_{t-1}(\delta)} \|X\|_{\bar{V}_{t-1}^{-1}}, 1\} \leq 2\sqrt{\beta_{t-1}(\delta)} \min\{\|X\|_{\bar{V}_{t-1}^{-1}}, 1\}$$

The second inequality is not true in general and becomes true only if $\beta_{t-1}(\delta) \geq 1$. The analysis of OFUL in (Lattimore and Szepesvári, 2020) goes around the issue by assuming $R = 1$.

B Proof of Theorem 3

Define the instantaneous regret as $r_t := \max_{x \in \mathcal{X}_t} \langle x, \theta^* \rangle - \langle x_t, \theta^* \rangle$, then

Proof. Consider the following,

$$\begin{aligned} \text{Reg}_T &= \sum_{t=1}^T r_t \\ &= \sum_{t=1}^T \max_{x \in \mathcal{X}_t} \langle x, \theta^* \rangle - \langle x_t, \theta^* \rangle \\ &= \sum_{t=1}^T \mathbb{1}\{\lambda_t^{1/2} S_* > R\sqrt{d}\} r_t + \sum_{t=1}^T \mathbb{1}\{\lambda_t^{1/2} S_* \leq R\sqrt{d}\} r_t \\ &= \sum_{t=1}^{t_0-1} r_t + \sum_{t=t_0}^T r_t \end{aligned}$$

where $t_0 = \min\{t \geq 1 : \lambda_t^{1/2} S_* \leq R\sqrt{d}\}$. Therefore consider the case $\lambda_t^{1/2} S_* > R\sqrt{d}$,

$$\begin{aligned} \sum_{t=1}^T \mathbb{1}\{\lambda_t^{1/2} S_* > R\sqrt{d}\} r_t &= \sum_{t=1}^{t_0-1} r_t \\ &= \sum_{t=1}^{t_0-1} \max_{x \in \mathcal{X}_t} \langle x - x_t, \theta^* \rangle \\ &\leq \sum_{t=1}^{t_0-1} \max_{x \in \mathcal{X}_t} \|x - x_t\|_2 \|\theta^*\|_2 \\ &\leq \sum_{t=1}^{t_0-1} 2S_* \\ &= 2(t_0 - 1)S_*. \end{aligned}$$

Now consider the second term for $t \geq t_0$. Since $\lambda_t^{1/2} S_* \leq R\sqrt{d}$, we can adapt the analysis of Theorem 3 from (Abbasi-Yadkori et al., 2011). Adapting Theorem 2 of (Abbasi-Yadkori et al., 2011) for our problem we have the following

Theorem 6. Define $t_0 := \min\{t \geq 1 : \lambda_t^{1/2} S_* \leq R\sqrt{d}\}$, and recall $y_t = \langle x_t, \theta^* \rangle + \eta_t$. Then, for any $\delta > 0$ with $\delta_t = \frac{1}{t^2} \frac{6}{\pi^2} \delta$, we have, with probability at least $1 - \delta$,

$$\forall t \geq t_0, \theta^* \in \overline{C}_t = \left\{ \theta^* \in \mathbb{R}^d : \|\hat{\theta}_t - \theta^*\|_{V_t(\lambda_t)} \leq R\sqrt{2 \log\left(\frac{|V_t(\lambda_t)|^{1/2} |\lambda_t I|^{-1/2}}{\delta_t}\right)} + R\sqrt{d} \right\}$$

where $\hat{\theta}_t = \arg \min_{\theta} \sum_{k=1}^t (y_k - \langle x_k, \theta \rangle)^2 + \lambda_t \|\theta\|_2^2$.

Proof. We know from Theorem 3 from (Abbasi-Yadkori et al., 2011) that for any fixed λ_t and $\delta_t = \frac{1}{t^2} \frac{6}{\pi^2} \delta > 0$ that with probability at least $1 - \delta$, for all $t \geq t_0$

$$\theta^* \in \overline{C}_t = \left\{ \theta^* \in \mathbb{R}^d : \|\hat{\theta}_t - \theta^*\|_{V_t(\lambda_t)} \leq R\sqrt{2 \log\left(\frac{|V_t(\lambda_t)|^{1/2} |\lambda_t I|^{-1/2}}{\delta_t}\right)} + R\sqrt{d} \right\}.$$

Define $E = \cap_{t=t_0}^T \overline{C}_t$. We will show that with probability at least $1 - \delta$ and for any $t \geq t_0$ and for a decreasing sequence of λ_t and $\delta_t = \frac{1}{t^2} \frac{6}{\pi^2} \delta$ that $P(E) \geq 1 - \delta$. Indeed,

$$\begin{aligned} P(E^c) &= P(\cup_{t=t_0}^T \overline{C}_t^c) \\ &\leq \sum_{t=t_0}^T P(\overline{C}_t^c) && \text{(By the Union Bound)} \\ &\leq \sum_{t=t_0}^T \delta_t \\ &\leq \sum_{t=t_0}^T \frac{1}{t^2} \frac{6}{\pi^2} \delta \\ &\leq \delta && \text{(Since } \sum_{i=1}^{\infty} \frac{1}{i^2} = \frac{\pi^2}{6} \text{)} \end{aligned}$$

□

Therefore, with $\lambda_t^{1/2} S_* \leq R\sqrt{d}$, using Theorem 6 where $\sqrt{\beta_{t-1}(\delta)} = R\sqrt{2 \log\left(\frac{|V_t(\lambda_t)|^{1/2} |\lambda_t I|^{-1/2}}{\delta_t}\right)} + R\sqrt{d}$

$$\begin{aligned} r_t &= \max_{x \in \mathcal{X}_t} \langle x, \theta^* \rangle - \langle x_t, \theta^* \rangle \\ &\leq \langle x_t, \hat{\theta}_t \rangle + \sqrt{\beta_{t-1}(\delta)} \|x_t\|_{V_{t-1}^{-1}(\lambda_t)} - \langle x_t, \theta^* \rangle \end{aligned}$$

$$\begin{aligned}
 & \stackrel{(a)}{\leq} \langle x_t, \hat{\theta}_t \rangle + \sqrt{\bar{\beta}_{t-1}(\delta)} \|x_t\|_{V_{t-1}^{-1}(\lambda_t)} - \langle x_t, \hat{\theta}_t \rangle + \sqrt{\bar{\beta}_t(\delta)} \|x_t\|_{V_{t-1}^{-1}(\lambda_t)} \\
 & \leq 2\sqrt{\bar{\beta}_{t-1}(\delta)} \|x_t\|_{V_{t-1}^{-1}(\lambda_t)} \\
 & \stackrel{(b)}{\leq} 2\sqrt{\bar{\beta}_{t-1}(\delta)} \|x_t\|_{V_{t-1}^{-1}(\lambda_T)}.
 \end{aligned}$$

(a) comes from using the confidence ellipsoid such that $\langle x_t, \hat{\theta}_t \rangle \leq \langle x_t, \theta^* \rangle + \sqrt{\bar{\beta}_t(\delta)} \|x_t\|_{V_{t-1}^{-1}(\lambda_t)}$. (b) follows since λ_t is a decreasing sequence in t , this implies that $\lambda_t \geq \lambda_T > 0$ for all t and in turn, $\|x_t\|_{V_{t-1}^{-1}(\lambda_t)} \leq \|x_t\|_{V_{t-1}^{-1}(\lambda_T)}$ and $\bar{\beta}_T(\delta) \geq 1 + \beta_t(\delta) \geq 1$. For $t_0 \geq 1$, let $H_T = \{t \in \{t_0, \dots, T\} : \|x_t\|_{V_{t-1}^{-1}(\lambda_T)}^2 > 1\}$ and let $\bar{H}_T := \{t_0, \dots, T\} \setminus H_T$ and $(T - t_0) = |\bar{H}_T| + |H_T|$, it follows

$$\begin{aligned}
 \sum_{t=t_0}^T r_t & \leq \sum_{t=t_0}^T r_t \{\|x_t\|_{V_{t-1}^{-1}(\lambda_T)}^2 \leq 1\} + \sum_{t=t_0}^T r_t \{\|x_t\|_{V_{t-1}^{-1}(\lambda_T)}^2 > 1\} \\
 & \leq \sqrt{|\bar{H}_T|} \sum_{t \in \bar{H}_T} r_t^2 + \sum_{t \in H_T} 2S_* \\
 & \leq 2\sqrt{|\bar{H}_T|} \sum_{t \in \bar{H}_T} \bar{\beta}_T(\delta) \|x_t\|_{V_{t-1}^{-1}(\lambda_T)}^2 + 2S_* |H_T| \\
 & \stackrel{(c)}{\leq} 2\sqrt{(T - (t_0 - 1))\bar{\beta}_T(\delta)} \sqrt{2d \log \left(1 + \frac{T - (t_0 - 1)}{d\lambda_T}\right)} + 4S_* \frac{d}{\log(2)} \log \left(1 + \frac{1}{\lambda_T \log(2)}\right)
 \end{aligned}$$

where $\sqrt{\bar{\beta}_T(\delta)} = R\sqrt{2d \log \left(\frac{T^2 \pi^2 (1+T/\lambda_T)}{6\delta}\right)} + R\sqrt{d}$ and (c) follows from Lemma 7 and the elliptical potential lemma. Therefore, summing the bounds, $\text{Reg}_T = \sum_{t=1}^T r_t = \sum_{t=1}^{t_0-1} r_t + \sum_{t=t_0}^T r_t$ we obtain our desired results. \square

Proof of Corollary 4. For each case let $t_0 = \min\{t \geq 1 : \lambda_t^{1/2} S \leq R\sqrt{d}\}$.

1. Consider $\lambda_t = \frac{R^2 d}{\log^\gamma(1+t)}$, for $\gamma > 0$, $t \geq 0$ then,

$$t_0 = \exp(S_*^{2/\gamma}) - 1$$

Now we insert into the regret bound from Theorem 6 and it follows that w.p at least $1 - \delta$,

$$\begin{aligned}
 \text{Reg}_T & \leq 2\exp(S_*^{2/\gamma}) S_* \\
 & + \sqrt{T+2} \left(R \sqrt{2d \log \left(\frac{T^2 \pi^2 (1 + \frac{T \log^\gamma(1+T)}{dR^2})}{6\delta}\right)} + R\sqrt{d} \right) \sqrt{2d \log \left(1 + \frac{\log^\gamma(1+T)(T - \exp(S_*^{2/\gamma}) + 2)}{R^2 d^2}\right)} \\
 & + 4S_* \frac{d}{\log(2)} \log \left(1 + \frac{\log^\gamma(1+T)}{d^2 R^2 \log(2)}\right) \\
 & \leq \tilde{\mathcal{O}}(Rd\sqrt{T} + S_*(d + \exp(S_*^{2/\gamma}))).
 \end{aligned}$$

2. Consider $\lambda_t = \frac{R^2 d}{t^\alpha}$ where for this case

$$t_0 = S_*^{2/\alpha}.$$

Now we insert into the regret bound from Theorem 6 and it follows that w.p at least $1 - \delta$,

$$\begin{aligned}
 \text{Reg}_T & \leq 2S_*^{2/\alpha+1} + \sqrt{T - S_*^{2/\alpha} + 1} \left(R \sqrt{2d \log \left(\frac{T^2 \pi^2 (1 + T^{\alpha+1}/(dR^2))}{6\delta}\right)} + R\sqrt{d} \right) \sqrt{2d \log \left(1 + \frac{T^\alpha (T - S_*^{2/\alpha} + 1)}{R^2 d^2}\right)} \\
 & + 4S_* \frac{d}{\log(2)} \log \left(1 + \frac{T^\alpha}{R^2 d^2 \log(2)}\right) \\
 & \leq \tilde{\mathcal{O}}(Rd\sqrt{(\alpha^2 + 1)T} + S_*(d\alpha + S_*^{2/\alpha}))
 \end{aligned}$$

3. Consider $\lambda_t = \frac{R^2 d}{\exp(t^q)}$ for $q \in (0, 1)$, then

$$t_0 = \log^{1/q}(S_*^2).$$

Now we insert into the regret bound from Theorem 6 and it follows that w.p at least $1 - \delta$,

$$\begin{aligned} & \text{Reg}_T \\ & \leq 2 \log^{1/q}(S_*^2) S_* \\ & \quad + \sqrt{T} \left(R \sqrt{2d \log \left(\frac{T^2 \pi^2 (1 + T \exp(T^q)/(dR^2))}{6\delta} \right)} + R\sqrt{d} \right) \sqrt{2d \log \left(1 + \frac{\exp(T^q)(T - \log^{1/q}(S_*^2) + 1)}{R^2 d^2} \right)} \\ & \quad + 4S_* \frac{d}{\log(2)} \log \left(1 + \frac{\exp(T^q)}{R^2 d^2 \log(2)} \right) \\ & \leq \tilde{\mathcal{O}}(Rd\sqrt{T^{1+q}} + S_*(dT^q + \log^{1/q}(S_*^2))). \end{aligned}$$

□

B.1 Elliptical Potential Count Lemma

We present our elliptical potential count lemma in Lemma 7 below, which can also be found in a parallel work of Kim et al. (2021). We also show a tight and easy derivation for bounding implicit inequalities like $X \leq A \log(1 + BX)$

The original form of Lemma 7 appears in Lattimore and Szepesvári (2020, Exercise 19.3), but Lemma 7 has an improved leading constant and has an arguably more streamlined proof. We remark that a similar result appeared earlier in Russo and Van Roy (2013, Proposition 3) as well.

Lemma 7. (*Elliptical potential count*) Let $x_1, \dots, x_T \in \mathbb{R}^d$ be a sequence of vectors with $\|x_t\|_2 \leq 1$ for all $t \in [T]$. Let $V_t(\lambda) = \lambda I + \sum_{s=1}^t x_s x_s^\top$, and define the following, $H_T = \{t \in [T] : \|x\|_{V_{t-1}^{-1}(\lambda)}^2 > 1\}$ and denote the size of H_T by $|H_T|$, then for $t \in [T]$

$$|H_T| \leq \frac{2d}{\log(2)} \log \left(1 + \frac{1}{\lambda \log(2)} \right).$$

Proof. Let $M_T = V_T^{H_T}(\lambda) = \lambda I + \sum_{s=1}^T \mathbb{1}\{\|x_s\|_{V_{s-1}^{-1}} \geq 1\} x_s x_s^\top = \lambda I + \sum_{s \in H_T} x_s x_s^\top$. The determinant of M_T is

$$|M_T| = |\lambda I + \sum_{s \in H_T} x_s x_s^\top| = \prod_{i=1}^d (\lambda + \lambda_i),$$

where λ_i are the eigenvalues of $\sum_{s \in H_T} x_s x_s^\top$. We now bound above,

$$\begin{aligned} |M_T| &= \left| \lambda I + \sum_{s \in H_T} x_s x_s^\top \right| \\ &= \prod_{i=1}^d (\lambda + \lambda_i) \\ &\leq \left(\sum_{i=1}^d \left(\frac{\lambda + \lambda_i}{d} \right) \right)^d && \text{(AM - GM)} \\ &= \left(\frac{1}{d} \text{trace}(M_T) \right)^d \\ &\leq \left(\frac{1}{d} (\text{trace}(V) + \text{trace}(\sum_{s=1}^T \mathbb{1}\{s \in H_T\} x_s x_s^\top)) \right)^d \\ &\leq \left(\frac{1}{d} (d\lambda + \sum_{s=1}^T \mathbb{1}\{s \in H_T\} x_s x_s^\top) \right)^d \end{aligned}$$

$$\leq \left(\frac{d\lambda + |H_T|}{d} \right)^d$$

To get a lower bound, we start with the following form of M_T , as shown in [Abbasi-Yadkori et al. \(2011, proof of Lemma 11\)](#),

$$\begin{aligned} |M_T| &= |\lambda I| \prod_{k \in H_T} (1 + \|x_k\|_{M_{k-1}^{-1}}^2) \\ &\geq |\lambda I| \prod_{k \in H_T} (1 + \|x_k\|_{V_{k-1}^{-1}(\lambda)}^2) \\ &= \lambda^d \prod_{k \in H_T} (1 + \|x_k\|_{V_{k-1}^{-1}(\lambda)}^2) \\ &\geq \lambda^d 2^{|H_T|}. \end{aligned} \quad (\text{Since } \|x_k\|_{V_{k-1}^{-1}(\lambda)} \geq 1 \text{ for all } k \in H_T \text{ and } \prod_{i=1}^n 2 = 2^n)$$

The first inequality is true because $\|x_k\|_{V_{k-1}^{-1}(\lambda)}^2$ is a nonincreasing sequence in k ; i.e., $\|x_k\|_{M_{k-1}^{-1}} \geq \|x_k\|_{V_{k-1}^{-1}(\lambda)}$. Thus, we have

$$\lambda^d 2^{|H_T|} \leq \left(\frac{d\lambda + |H_T|}{d} \right)^d.$$

Taking the log of both sides and rearranging we obtain

$$|H_T| \leq \frac{d}{\log(2)} \log \left(1 + \frac{|H_T|}{\lambda d} \right).$$

Using Lemma 8 with $\eta = 1/2$, $A = \frac{d}{\log(2)}$, $B = \frac{1}{d\lambda}$, and $X = |H_T|$,

$$|H_T| \leq \frac{2d}{\log(2)} \log \left(\frac{2d}{2\log(2)} \left(\frac{1}{|H_T|} + \frac{1}{\lambda d} \right) \right) = \frac{2d}{\log(2)} \log \left(\frac{1}{\log(2)} \left(\frac{d}{|H_T|} + \frac{1}{\lambda} \right) \right).$$

We now consider two cases; when for some value $c > 0$, $\frac{|H_T|}{d} \leq c$ and $\frac{|H_T|}{d} \geq c$.

Case 1: Suppose that $|H_T| < dc$. Then it follows that

$$|H_T| \leq 2d \log \left(1 + \frac{|H_T|}{d\lambda} \right) \leq 2d \log \left(1 + \frac{c}{\lambda} \right)$$

Case 2: Suppose that $|H_T| \geq dc$ then

$$\begin{aligned} |H_T| &\leq \frac{2d}{\log(2)} \log \left(\frac{1}{\log(2)} \left(\frac{d}{|H_T|} + \frac{1}{\lambda} \right) \right) \\ &\leq \frac{2d}{\log(2)} \log \left(\frac{1}{\log(2)} \left(\frac{1}{c} + \frac{1}{\lambda} \right) \right) \end{aligned}$$

Choose $c = \frac{1}{\log(2)}$ and since the right hand side is independent of t , then for up to time T for all cases we have that

$$|H_T| \leq \frac{2d}{\log(2)} \log \left(1 + \frac{1}{\log(2)\lambda} \right).$$

□

We now present Lemma 8 that is a key inequality used in the proof above. We remark that one can make the leading constant 1 by applying the bound to the RHS of $X \leq A \log(1 + BX)$:

$$X \leq A \log \left(1 + B \cdot \left(\inf_{\eta \in (0,1)} \frac{1}{1-\eta} A \log \left(\frac{A}{2\eta} \left(\frac{1}{X} + B \right) \right) \right) \right)$$

Lemma 8. *Let $X, A, B \geq 0$. Then, $X \leq A \log(1 + BX)$ implies that*

$$X \leq \inf_{\eta \in (0,1)} \frac{1}{1-\eta} A \log \left(\frac{A}{2\eta} \left(\frac{1}{X} + B \right) \right)$$

Proof. Let $\eta \in (0,1)$

$$\begin{aligned} X &\leq A \log(1 + BX) \\ \Rightarrow X &\leq A \log\left(\frac{\eta X}{A}\right) + A \log\left(\frac{A}{\eta} \left(\frac{1}{X} + B\right)\right) \\ \Rightarrow X &\leq A\left(\frac{\eta X}{A} - 1\right) + A \log\left(\frac{A}{\eta} \left(\frac{1}{X} + B\right)\right) && \text{(since } \log(x) \leq x - 1\text{)} \\ \Rightarrow X(1 - \eta) &\leq A \log\left(\frac{A}{2\eta} \left(\frac{1}{X} + B\right)\right) \\ X &\leq \frac{1}{1-\eta} A \log\left(\frac{A}{2\eta} \left(\frac{1}{X} + B\right)\right) && \text{(since } \eta < 1\text{)} \end{aligned}$$

□

Note that the technique shown above is more generic and can be used to solve implicit inequalities like $X \leq O(\log(X))$.² For example, it is easy to show that

$$X \leq A \log(X) + C \implies X \leq \inf_{\eta \in (0,1)} \frac{1}{1-\eta} \left(A \log\left(\frac{A}{\eta e}\right) + C \right)$$

In fact, one can apply the same technique to solve $X^p \leq O(\log(X))$ where $p > 0$. The proof simply introduced a free parameter η that is tight with the best choice of η . This proof is much simpler than other known techniques such as [Antos et al. \(2010, Appendix B\)](#).

C Proof of Theorem 5

Lemma 9. *In OLSOFUL (Algorithm 3), the length of the warmup phase τ satisfies*

$$\tau = \mathcal{O}(d \log(d))$$

Proof. Define $\bar{V}(\pi) := \sum_{x \in \mathcal{X}} \pi_x x x^\top$ where $\pi \in \Delta^{|\mathcal{X}|}$. Recall that τ is the length of the warmup phase. Let $\omega := |\mathcal{X}_0|$ be the number of time steps spent on the first stage of the warmup procedure. Let $\pi_\omega \in \Delta^K$ have $1/(2d)$ for $x \in \mathcal{X}_0$ and 0 otherwise. Let

$$\bar{V}_\tau = \bar{V}_\omega + \sum_{s=\omega+1}^{\tau} x_s x_s^\top$$

where $\bar{V}_\omega := 2d \cdot \bar{V}(\pi_\omega) = \sum_{x \in \mathcal{X}_0} x x^\top$. Thus, \bar{V}_ω is the initial matrix for \bar{V}_τ defined by the $2d$ samples used in Algorithm 3.

Let $\pi^* = \arg \min_{\pi \in \Delta^{|\mathcal{X}|}} \max_{x \in \mathcal{X}} \|x\|_{\bar{V}(\pi)}^2$, the solution of the G-optimal design problem. By [Kiefer and Wolfowitz \(1960\)](#), $\pi^* = \arg \max_{\pi \in \Delta^{|\mathcal{X}|}} \max_{x \in \mathcal{X}} |\bar{V}(\pi)|$ as well. This implies that $|\bar{V}_\tau| = \tau^d |\frac{1}{\tau} \bar{V}_\tau| \leq \tau^d |\bar{V}(\pi^*)|$. Then, $|\bar{V}_\omega| = \omega^d |\frac{1}{\omega} \bar{V}_\omega|$. According to proof of Theorem 4.2 of [Kumar and Yildirim, 2005](#), we have

$$\log |\bar{V}(\pi^*)| - \log |\bar{V}(\pi_\omega)| = \mathcal{O}(d \log(d)) \tag{10}$$

which means that the first stage of the warmup allocates arms so the resulting (normalized) determinant is sufficiently close to the ideal allocation of π^* .

We will now show $\tau \leq 2d + \mathcal{O}(d \log(d))$, thus, using the bounds on the determinant of $\log |\bar{V}_\tau|$ and $\log |\bar{V}_\omega|$

$$\begin{aligned} \log |\bar{V}_\tau| - \log |\bar{V}_\omega| &\leq d \log(\tau) + \log(|\bar{V}(\pi^*)|) - \log \left| \frac{1}{\omega} \bar{V}(\omega) \right| - d \log(\omega) \\ &\leq d \log(\tau) + \mathcal{O}(d \log(d)) - d \log(\omega) \end{aligned}$$

where the last inequality is by (10). On the other hand,

$$\log |\bar{V}_\tau| - \log |\bar{V}_\omega| = \sum_{t=\omega+1}^{\tau} \log(1 + \|x_t\|_{\bar{V}_{t-1}}^2) > (\tau - \omega) \log(2)$$

Therefore, we have

$$(\tau - \omega) \log(2) < \mathcal{O}(d \log(d)) + d \log\left(\frac{\tau}{\omega}\right).$$

Dropping $\log(2)$ on the LHS we can find τ . Thus,

$$\begin{aligned} (\tau - \omega) \log(2) &< \mathcal{O}(d \log(d)) + d \log\left(\frac{\tau}{\omega}\right) \\ \Rightarrow \tau &< \mathcal{O}(d \log(d)) + \frac{d}{\log(2)} d \log\left(\frac{2d\tau}{2d\omega}\right) + \omega \\ \Rightarrow \tau &< \mathcal{O}(d \log(d)) + \frac{d}{\log(2)} \log\left(\frac{\tau}{2d}\right) + \frac{d}{\log(2)} \log\left(\frac{2d}{\omega}\right) + 2d && (\omega \leq 2d) \\ \Rightarrow \tau &< \mathcal{O}(d \log(d)) + \frac{d}{\log(2)} \left(\frac{\tau}{2d} - 1\right) + \frac{d}{\log(2)} \log\left(\frac{2d}{\omega}\right) + 2d && (\log(x) \leq x - 1) \\ \Rightarrow \tau \left(\frac{2 \log(2) - 1}{2 \log(2)}\right) &< \mathcal{O}(d \log(d)) + \frac{d}{\log(2)} \log\left(\frac{d}{\eta\omega}\right) + 2d \\ \Rightarrow \tau &< \mathcal{O}(d \log(d)) + \frac{4 \log(2)}{2 \log(2) - 1} d + \frac{4 \log(2)}{2 \log(2) - 1} d \log\left(\frac{d}{\eta\omega}\right) = \mathcal{O}(d \log(d)). \end{aligned}$$

This concludes the proof. \square

Theorem 10. Suppose $\mathcal{X} \subset \mathbb{R}^d$. Define $\bar{V}_\tau := \sum_{s=1}^{\tau} x_s x_s^\top$ where τ is the length of the warmup procedure. Then, for any $\delta > 0$, we have, with at least probability $1 - 2\delta$ and for all $t \geq 0$ that

$$\theta^* \in C_t = \left\{ \theta^* \in \mathbb{R}^d : \|\hat{\theta}_{t,0} - \theta^*\|_{\bar{V}_t} \leq \sqrt{\beta_t^{\text{OLS}}} := R \left(\sqrt{\log\left(\frac{|\bar{V}_t|}{|\bar{V}_\tau| \delta^2}\right)} + \sqrt{2 \log(2) d + 4 \log(1/\delta)} \right) \right\}$$

where $\hat{\theta}_{t,0}$ is the estimator defined for $\tau + 1, \dots, t$, such that

$$\hat{\theta}_{t,0} = \arg \min_{\theta} \sum_{s=1}^t \frac{1}{2} (y_s - x_s^\top \theta)^2. \quad (11)$$

Proof. We start by defining notation. Define τ to be the length of the warmup procedure. Let $\hat{\theta}_{\tau,0}$ be the MLE up to τ . Let $X_{\leq \tau} \in \mathbb{R}^{\tau \times d}$ ($X_{> \tau} \in \mathbb{R}^{(t-\tau) \times d}$) be the design matrix consisting of the pulled arms during the warmup phase (after the warmup up time step t , respectively). Recall $\bar{V}_\tau := \sum_{s=1}^{\tau} x_s x_s^\top$ is the covariance matrix. Let $\eta_{\leq \tau} = (\eta_1, \dots, \eta_\tau)^\top$ and $\eta_{> \tau} = (\eta_{\tau+1}, \dots, \eta_t)^\top$. For the following rounds after τ , define the MLE estimator

$$\hat{\theta}_{t,0} = \arg \min_{\theta} \frac{1}{2} \sum_{s=\tau+1}^t (x_s^\top \theta - y_s)^2 + \frac{1}{2} \|\theta - \hat{\theta}_{\tau,0}\|_{\bar{V}_\tau}^2. \quad (12)$$

Fix $x \in \mathbb{R}^d$. The optimality condition can be written as

$$\bar{V}_t(\hat{\theta}_{t,0} - \theta^*) = \left(X_{>\tau}^\top X_{>\tau} + \bar{V}_\tau \right) (\hat{\theta}_{t,0} - \theta^*) = X_{>\tau}^\top \eta_{>\tau} - \bar{V}_\tau (\theta^* - \hat{\theta}_{\tau,0})$$

In fact, this optimality condition for Estimator (12) is exactly that of using the whole data. That is, Estimator (12) has the same optimality condition as Estimator (11).

Fix $x \in \mathbb{R}^d$, which we will specify later on. Then,

$$x^\top (\hat{\theta}_{t,0} - \theta^*) = x^\top \bar{V}_t^{-1} X_{>\tau}^\top \eta_{>\tau} - x^\top \bar{V}_t^{-1} \bar{V}_\tau (\theta^* - \hat{\theta}_{\tau,0}). \quad (13)$$

Consider the first term on the RHS of Equation (13). We will bound it using Abbasi-Yadkori et al. (2011, Theorem 1). We will use the high probability bound,

$$\mathbb{P} \left(\forall t \geq 0 : \|X_{>\tau}^\top \eta_{>\tau}\|_{\bar{V}_t^{-1}} \leq R \sqrt{\log \left(\frac{|\bar{V}_t|}{|\bar{V}_\tau|} \frac{1}{\delta^2} \right)} \right) \geq 1 - \delta. \quad (14)$$

Therefore, with probability at least $1 - \delta$ for all $t \geq \tau + 1$,

$$\begin{aligned} x^\top \bar{V}_t^{-1} X_{>\tau}^\top \eta_{>\tau} &\leq \|x\|_{\bar{V}_t^{-1}} \|X_{>\tau}^\top \eta_{>\tau}\|_{\bar{V}_t^{-1}} \\ &\stackrel{(a)}{\leq} \|x\|_{\bar{V}_t^{-1}} R \sqrt{\log \left(\frac{|\bar{V}_t|}{|\bar{V}_\tau|} \frac{1}{\delta^2} \right)}. \end{aligned}$$

where (a) is due to the high probability bound in Equation (14).

One can further bound the adaptive design by

$$\begin{aligned} \log \left(\frac{|\bar{V}_t|}{|\bar{V}_\tau| \delta^2} \right) &\leq \log \left(\frac{|\bar{V}_\tau^{-1/2} \bar{V}_t \bar{V}_\tau^{-1/2}|}{\delta^2} \right) \leq d \log \left(\frac{1}{d} \text{tr}(\bar{V}_\tau^{-1/2} \bar{V}_t \bar{V}_\tau^{-1/2}) \frac{1}{\delta^2} \right) \\ &= d \log \left(\frac{1}{d} \sum_{s=1}^t \text{tr}(\bar{V}_\tau^{-1/2} x_s x_s^\top \bar{V}_\tau^{-1/2}) \frac{1}{\delta^2} \right) \\ &\stackrel{(b)}{\leq} d \log \left(\frac{t}{d} \frac{1}{\delta^2} \right) \end{aligned}$$

where (b) is due to the fact that $\max_{x \in \mathcal{X}} \text{tr}(\bar{V}_\tau^{-1/2} x x^\top \bar{V}_\tau^{-1/2}) = \max_{x \in \mathcal{X}} \|x\|_{\bar{V}_\tau^{-1}}^2 \leq 1$.

Consider the second term on the RHS of Equation (13). We use Abbasi-Yadkori et al. (2011, Theorem 1). We have the high probability event

$$\mathbb{P} \left(\forall t \geq 0 : \|X_{\leq \tau}^\top \eta_{\leq \tau}\|_{[(1+\sigma)\bar{V}_\tau]^{-1}} \leq R \sqrt{\log \left(\frac{|(1+\sigma)\bar{V}_\tau|}{|\sigma \bar{V}_\tau|} \frac{1}{\delta^2} \right)} \right) \geq 1 - \delta. \quad (15)$$

Therefore with probability at least $1 - \delta$ for all $t \geq \tau + 1$

$$\begin{aligned} -x^\top \bar{V}_t^{-1} \bar{V}_\tau (\theta^* - \hat{\theta}_{\tau,0}) &= -x^\top \bar{V}_t^{-1} \bar{V}_\tau (\theta^* - \bar{V}_\tau^{-1} X_{\leq \tau}^\top X_{\leq \tau} \theta^* - \bar{V}_\tau^{-1} X_{\leq \tau}^\top \eta_{\leq \tau}) \\ &= -x^\top \bar{V}_t^{-1} \bar{V}_\tau (-\bar{V}_\tau^{-1} X_{\leq \tau}^\top \eta_{\leq \tau}) \\ &= x^\top \bar{V}_t^{-1} X_{\leq \tau}^\top \eta_{\leq \tau} \\ &\leq \|x\|_{\bar{V}_t^{-1}} \|X_{\leq \tau}^\top \eta_{\leq \tau}\|_{\bar{V}_t^{-1}} \\ &\leq \|x\|_{\bar{V}_t^{-1}} \|X_{\leq \tau}^\top \eta_{\leq \tau}\|_{\bar{V}_\tau^{-1}} \\ &= \|x\|_{\bar{V}_t^{-1}} \sqrt{1 + \sigma} \|X_{\leq \tau}^\top \eta_{\leq \tau}\|_{[(1+\sigma)\bar{V}_\tau]^{-1}} \quad (\text{for any } \sigma > 0) \end{aligned}$$

$$\begin{aligned}
 &\stackrel{(c)}{\leq} \|x\|_{\bar{V}_t^{-1}} \sqrt{1+\sigma} R \sqrt{\log\left(\frac{|(1+\sigma)\bar{V}_\tau|}{|\sigma\bar{V}_\tau|} \frac{1}{\delta^2}\right)} \\
 &\leq \|x\|_{\bar{V}_t^{-1}} \sqrt{1+\sigma} R \sqrt{d \log\left(\frac{1+\sigma}{\sigma}\right) + 2\log(1/\delta)}.
 \end{aligned}$$

where (c), is due to the high probability bound in Equation (15).

We now choose $\sigma = 1$, which gives us

$$\sqrt{1+\sigma} R \sqrt{d \log\left(\frac{1+\sigma}{\sigma}\right) + 2\log(1/\delta)} \leq R \sqrt{2\log(2)d + 4\log(1/\delta)}.$$

Finally, we combine the two high probability bounds, Equation's (14) and (15) to bound the RHS of Equation (13). Then we have with probability at least $1 - 2\delta$,

$$|x^\top(\hat{\theta}_{t,0} - \theta^*)| \leq \|x\|_{\bar{V}_t^{-1}} R \left(\sqrt{\log\left(\frac{|\bar{V}_t|}{|\bar{V}_\tau|\delta^2}\right)} + \sqrt{2\log(2)d + 4\log(1/\delta)} \right)$$

Choosing x we obtain our final result. That is, set $x = \bar{V}_t(\hat{\theta}_{t,0} - \theta^*)$. This implies $|x^\top(\hat{\theta}_{t,0} - \theta^*)| = \|\hat{\theta}_{t,0} - \theta^*\|_{\bar{V}_t}^2$ and $\|\bar{V}_t(\hat{\theta}_{t,0} - \theta^*)\|_{\bar{V}_t^{-1}} = \|\hat{\theta}_{t,0} - \theta^*\|_{\bar{V}_t}$. Then with probability at least $1 - 2\delta$,

$$\|\hat{\theta}_{t,0} - \theta^*\|_{\bar{V}_t} \leq R \left(\sqrt{\log\left(\frac{|\bar{V}_t|}{|\bar{V}_\tau|\delta^2}\right)} + \sqrt{2\log(2)d + 4\log(1/\delta)} \right)$$

□

Proof of Theorem 5. Recall $\omega = |\mathcal{X}_0|$ and τ is the length of the total warmup procedure in Algorithm 3. Define $r_t := \max_{x \in \mathcal{X}} \langle x, \theta^* \rangle - \langle x_t, \theta^* \rangle$ and recall $\text{Reg}_T = \sum_{t=1}^T r_t$. Then,

$$\begin{aligned}
 \text{Reg}_T &= \sum_{t=1}^{\tau} r_t + \sum_{t=\tau+1}^T r_t \\
 &\stackrel{(a)}{\leq} 2dS_* + S_* \mathcal{O}(d \log(d)) + \sum_{t=\tau+1}^T r_t
 \end{aligned}$$

where (a) is due to Lemma 9. It remains to bound the last term. On the high probability event in Theorem 10,

$$C_t = \left\{ \theta^* \in \mathbb{R}^d : \|\hat{\theta}_{t,0} - \theta^*\|_{\bar{V}_t} \leq \sqrt{\beta_t^{\text{OLS}}} \right\}$$

we bound the following,

$$\begin{aligned}
 \sum_{t=\tau+1}^T r_t &= \sum_{t=\tau+1}^T \left(\max_{x \in \mathcal{X}} \langle x, \theta^* \rangle - \langle x_t, \theta^* \rangle \right) \\
 &\leq \sum_{t=\tau+1}^T \left(\langle x_t, \tilde{\theta}_{t,0} \rangle - \langle x_t, \theta^* \rangle \right) \\
 &= \sum_{t=\tau+1}^T \left(\langle x_t, \tilde{\theta}_{t,0} - \hat{\theta}_{t,0} \rangle - \langle x_t, \theta^* - \hat{\theta}_{t,0} \rangle \right) \\
 &\leq \sum_{t=\tau+1}^T \|x_t\|_{\bar{V}_{t-1}^{-1}} \left(\|\tilde{\theta}_{t-1,0} - \hat{\theta}_{t-1,0}\|_{\bar{V}_{t-1}} + \|\theta^* - \hat{\theta}_{t-1,0}\|_{\bar{V}_{t-1}} \right)
 \end{aligned}$$

$$\begin{aligned}
 &\stackrel{(b)}{\leq} \sum_{t=\tau+1}^T \|x_t\|_{\bar{V}_{t-1}^{-1}} 2\sqrt{\beta_{t-1}^{\text{OLS}}} \\
 &\leq 2\sqrt{T \sum_{t=\tau+1}^T \|x_t\|_{\bar{V}_{t-1}^{-1}}^2 \beta_{t-1}^{\text{OLS}}} \\
 &\leq 2\sqrt{T \sum_{t=\tau+1}^T 2\log(1 + \|x_t\|_{\bar{V}_{t-1}^{-1}}^2) \beta_{t-1}^{\text{OLS}}} \\
 &= 2\sqrt{2} \sqrt{T \log\left(\frac{|\bar{V}_T|}{|\bar{V}_\tau|}\right) \beta_{t-1}^{\text{OLS}}} \\
 &\leq \mathcal{O}\left(R\sqrt{dT \log\left(\frac{T}{d}\right) \log\left(\frac{T}{d\delta}\right)}\right)
 \end{aligned}$$

where (b) uses the high probability event of C_t . Hence the desired regret bound holds with at least probability $1 - 2\delta$. \square

C.1 Computationally Efficient Version of Algorithm 4

As discussed in Section 5, we present a computationally efficient version of Algorithm 4, which is displayed in Algorithm 5. That is, for choosing the direction, b_i by utilizing the Gram-Schmidt Process to find a vector that lives in the orthogonal complement of Ψ . It is easy to verify that the computational complexity of Algorithm 5 is $\mathcal{O}(d^2|\mathcal{X}|)$.

Algorithm 5 Computationally efficient BH algorithm (Betke and Henk, 1993)

Input: the set of points $\mathcal{X} \subset \mathbb{R}^d$ with $|\mathcal{X}| = K$.

If $K \leq 2d$, then $\mathcal{X}_0 \leftarrow \mathcal{X}$. Return.

$\Psi \leftarrow \{0\}$, $\mathcal{X}_0 \leftarrow \emptyset$, $i \leftarrow 0$, $v_0 \leftarrow (0, \dots, 0)^\top \in \mathbb{R}^d$.

while $\mathbb{R}^d \setminus \Psi \neq \emptyset$ **do**

$i \leftarrow i + 1$

if $i = 1$ **then**

Set $b_i = e_1$

else

Set $v_{i-1}^\perp = v_{i-1} - \sum_{j \leq i-2} \frac{\langle v_j^\perp, v_{i-1} \rangle}{\langle v_j^\perp, v_j^\perp \rangle} v_j^\perp$

Set $b_i = e_i - \sum_{j \leq i-1} \frac{\langle v_j^\perp, e_i \rangle}{\langle v_j^\perp, v_j^\perp \rangle} v_j^\perp$

end if

$p \leftarrow \arg \max_{x \in \mathcal{X}} \langle b_i, x \rangle$; $\mathcal{X}_0 \leftarrow \mathcal{X}_0 \cup \{p\}$

$q \leftarrow \arg \min_{x \in \mathcal{X}} \langle b_i, x \rangle$; $\mathcal{X}_0 \leftarrow \mathcal{X}_0 \cup \{q\}$

$v_i \leftarrow p - q$

$\Psi \leftarrow \text{Span}(\Psi, \{v_i\})$

end while
