Exploration in Linear Bandits with Rich Action Sets and its Implications for Inference

Debangshu Banerjee
Indian Institute of Science
Bangalore, India

Avishek Ghosh
Indian Institute of Technology
Bombay, India

Sayak Ray Chowdhury
Microsoft Research
Bangalore, India

Aditya Gopalan
Indian Institute of Science
Bangalore, India

Abstract

We present a non-asymptotic lower bound on the spectrum of the design matrix generated by any linear bandit algorithm with sub-linear regret when the action set has well-behaved curvature. Specifically, we show that the minimum eigenvalue of the expected design matrix grows as $\Omega(\sqrt{n})$ whenever the expected cumulative regret of the algorithm is $O(\sqrt{n})$, where n is the learning horizon, and the action-space has a constant Hessian around the optimal arm. This shows that such action-spaces force a polynomial lower bound on the least eigenvalue, rather than a logarithmic lower bound as shown by Lattimore and Szepesvari (2017) for discrete (i.e., well-separated) action spaces. Furthermore, while the latter holds only in the asymptotic regime ($n \to \infty$), our result for these "locally rich" action spaces is any-time. Additionally, under a mild technical assumption, we obtain a similar lower bound on the minimum eigen value holding with high probability. We apply our result to two practical scenarios – *model selection* and *clustering* in linear bandits. For model selection, we show that an epoch-based linear bandit algorithm adapts to the true model complexity at a rate exponential in the number of epochs, by virtue of our novel spectral bound. For clustering, we consider a multi agent framework where we show, by leveraging the spectral result, that no forced exploration is necessary—the agents can run a linear bandit algorithm and estimate their underlying parameters at once, and hence incur a low regret.

Proceedings of the 26[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

1 INTRODUCTION

Bandit optimisation traditionally focuses on the problem of minimising cumulative *regret*, or the shortfall in reward incurred along the trajectory of learning. To this end, it has yielded optimal, low-regret strategies such as UCB and Thompson sampling (Abbasi-yadkori et al., 2011; Abeille and Lazaric, 2017). On the other hand, there is also the important *inference* goal in bandit problems, in which the experimenter, having access to arms or alternatives, wishes to infer some useful properties of, or estimate a quantity related to, the system by sequential sampling. Perhaps the most well-known example is the problem of best arm identification in multi-armed bandits (Even-Dar et al., 2006; Soare et al., 2014), which is essentially a sequential hypothesis testing problem. It is well known that for optimal error rates in identifying the best arm, it is necessary to sample arms with at least *constant* frequencies, which is vastly more exploratory than the frequencies required for regret minimization (Bubeck et al., 2011).

In this paper, we are interested in identifying settings in which both objectives – sublinear regret and fast inference (estimation error) – are simultaneously possible, opening the door to many useful applications that combine both reward optimisation and statistical inference. As we shall see, this is possible with standard bandit algorithms such as (linear) UCB provided there is sufficient 'local richness' of actions or arms in the problem (think 'suitably continuous arm set'), which is often a reasonable structure encountered in many bandit optimization problems with continuous spaces of alternatives (power control, dynamic pricing, etc.) To introduce ideas, consider standard linear bandit with additive Gaussian noise. A key trajectory-dependent quantity that connects the two goals of regret minimisation and parameter estimation is the minimum singular value of the design matrix, $\lambda_{\min}(V_n)$. To see why, notice that any typical bandit algorithm for regret minimization forms, either explicitly or otherwise, confidence sets for the unknown linear model parameter θ^*, of the form $\|\theta - \widehat{\theta}\|_{V_n} \leqslant c\sqrt{\ln(n)}$, where $c > 0$ is a constant. Thus, getting a lower bound on the growth of $\lambda_{\min}(V_n)$, would help in determining how fast the confidence sets shrink and

in return would help to infer the true bandit parameter θ^*.

The closest related work that sheds light on the singular value of the design matrix is by Lattimore and Szepesvari (2017). The authors show that for a discrete action-space bandit and any bandit algorithm with sub-polynomial regret (e.g., UCB), the minimum singular value of the expected design matrix ($\mathbb{E}[V_n]$) must grow at a logarithmic rate over time. However, their analysis holds true only in the asymptotic regime, with no information available on what happens in a finite time horizon.

We show that in action-spaces with "nice" local curvature properties, bandit algorithms which have inherently good regret properties (at most of the order of $\sqrt{n}$), the minimum singular value of the expected design matrix grows at least as order of $\sqrt{n}$. This is accomplished by a novel use of matrix perturbation techniques (Weyl's inequality and the Davis-Kahan sin-θ theorem) together with the information-theoretic data-processing inequality. Moreover, this result holds true in a finite-time horizon and not just in the asymptotic regime: for all time horizons n larger than a baseline value n_0, we show $\lambda_{\min}\mathbb{E}[V_n] \geqslant \gamma\sqrt{n}$, where γ is a positive constant which depends upon the local curvature and the algorithm being used. A key implication of this result is that in bandits with continuous action-spaces, low-regret algorithms also offer significant exploration in the sense of estimating all 'directions' in the parameter space. This not only extends the work of Lattimore and Szepesvari (2017) to the continuous action setting but also strengthens it to a finite time result from an asymptotic one.

We conclude with two illustrative applications of our theory under a mild assumption that the same $\Omega(\sqrt{n})$ growth of the minimum eigenvalue holds also in high probability.[1] Specifically, we consider the model-selection (Foster et al., 2019) and clustering (Gentile et al., 2014a) problems in linear bandits to apply our result. In the model selection application, the norm of the unknown parameter θ is viewed as a measure of complexity of the problem. We show that a variant of the well-known Optimism in the Face of Uncertainty (OFU) algorithm can adapt to $\|\theta\|$. This is achieved by a careful application of our result to control the rate at which norm estimates converge to the the true norm. It is important to note that a similar result is achieved by Ghosh et al. (2021a) in the different but related setting of *stochastic contextual bandits*, albeit with restrictive assumptions on the contexts. We are able to proceed without such restrictions by virtue of our result. In the clustering setup, we consider several linear bandit agents partitioned into k clusters. We propose a clustering algorithm without any explicit exploration, where the agents simultaneously estimate their (linear model) parameter and attempt to play low-regret actions. When the clusters are separated, our algorithm obtains the correct clustering with high probability. Note that in Gentile et al. (2014b); Ghosh et al. (2021b), the framework of clustered bandits was considered in a contextual framework with several strong assumptions on the context distribution. Our work demonstrates that similar guarantees are attainable (in the context-free linear bandit setup) without additional assumptions via exploiting the rich-action-set inference result (Theorem 2.2). Finally, we empirically validate that the minimum eigenvalue of the design matrix generated by the well-known Thompson sampling algorithm (Agrawal and Goyal, 2013) indeed grows at a rate larger than $\sqrt{n}$.

[1] In the appendix, we provide a technical condition on the trajectory of linear bandit algorithms under which this holds.

Related work. The linear bandit problem has been studied extensively in a large body of work starting from the classic work of Auer et al. (2002). In this model, algorithms based on the celebrated optimism in the face of uncertainty principle has been designed and analyzed by several authors (Chu et al., 2011; Dani et al., 2008; Abbasi-yadkori et al., 2011). A related approach is posterior sampling, also known as Thompson sampling, where sufficient exploration is achieved by randomly sampling a parameter from a posterior distribution over θ (Agrawal and Goyal, 2013; Abeille and Lazaric, 2017). Another related line of work consider the linear reward model in reproducing kernel Hilbert spaces (Srinivas et al., 2009; Valko et al., 2013; Chowdhury and Gopalan, 2017). Spectral properties of the expected design matrix under a discrete action-space has been studied by Lattimore and Szepesvari (2017), while Hao et al. (2020) handles the finitely many contextual case with each context having discrete action space. The framework of model selection has recently gained a lot of momentum in the bandit literature. For example, Chatterji et al. (2020) introduced a hypothesis test based framework to select either the standard bandit or the linear bandit model. Furthermore, in Ghosh et al. (2021a), the authors define parameter norm and sparsity as complexity parameters for stochastic linear bandit and adapt to those without any apriori knowledge, and obtain model selection guarantees. Moreover, Foster et al. (2019) introduces an adaptive algorithm for a similar linear bandit problem with sparsity as a measure of complexity. Additionally, there are different line of works, that uses the the corrall framework of Agarwal et al. (2017) to obtain adaptive algorithms for bandits and reinforcement learning (for example, see Pacchiano et al. (2020)). Very recently, for generic contextual bandits, the adaptation question is also addressed in Krishnamurthy and Athey (2021). Furthermore, on clustering of bandits, Gentile et al. (2014b) proposes an algorithm that works only when the cluster separation is large, which was further improved in Ghosh et al. (2021b), where near optimal regret is obtained even when the clusters are not separable. Moreover, Ghosh et al. (2021b) also proposes a natural personalization framework, which is a generalization of the clustering setup.

Notation. For a positive definite matrix G, (denoted as

$G \succ 0$) and vector x we write $\|x\|_G^2 = x^\top G x$ The euclidean norm of a vector x is denoted as $\|x\|$ and the spectral norm of a matrix G is $\|G\| = \lambda_{\max}(G)$. For hermitian matrices we assume the eigen-decomposition as $G = \sum_{i=1}^d \lambda_i u_i u_i^\top$, with $\lambda_1 > \lambda_d = \lambda_{\min}$. We use standard definitions of Landau Notation when using O, o, Ω and ω notation. We use $\mathbb{E}_\theta[\ .\ ]$, to emphasize that the underlying bandit instance is parameterized by θ.

Problem setting. We consider the linear bandit model of Abbasi-yadkori et al. (2011). Let $\mathcal{X} \subset \mathbb{R}^d$. The learner interact with the environment over n rounds. At each round t, the learner chooses an action $A_t \in \mathcal{X}$ and correspondingly observes a reward $Y_t = \langle A_t, \theta \rangle + \eta_t$, where η_t is a zero-mean Gaussian noise, and $\theta \in \mathbb{R}^d$ is the unknown parameter. The optimal action is $x^* = \arg\max_{x \in \mathcal{X}} \langle x, \theta \rangle$. The performance of the learner is typically measured using its expected regret, defined as

$$\mathbb{E}[R_n(\theta)] = \mathbb{E}\left[\sum_{t=1}^n \langle x^* - A_t, \theta \rangle\right] .$$

Here the expectation is over the action-selection strategy of the learner, denoted, where needed, by π, and over the randomness in observed rewards.

2 MAIN RESULT

In this section, we develop our main result which characterizes the growth of the minimum eigenvalue of the design matrix for linear bandit algorithms run on action spaces with suitable 'local curvature'. By this we mean action spaces which are hyper-surfaces of the form $\{x : f(x) = c\}$, where f is a twice-continuously differentiable function. The following definition expresses the local curvature property needed for our result.

Definition 2.1 (Locally Constant Hessian (LCH) surface)**.** Consider the action space defined by $\mathcal{X} = \{x \in \mathbb{R}^d : f(x) = c\}$, where $f : \mathbb{R}^d \to \mathbb{R}$ is a C^2 function (i.e., all second-order partial derivatives of f exist and are continuous) and $c \in \mathbb{R}$. Let $\theta \in \mathbb{R}^d$. $\mathcal{X}$ is said to be a LCH surface w.r.t. θ if: (i) there is a unique reward-optimal arm with respect to θ (denoted by $\mathrm{OPT}_{\mathcal{X}}(\theta) = \arg\max_{x \in \mathcal{X}} \langle x, \theta \rangle$), and (ii) there is an open neighborhood $U \subset \mathbb{R}^d$ of $\mathrm{OPT}_{\mathcal{X}}(\theta)$ over which the Hessian of f is constant and positive-definite.

Examples of LCH action spaces. Any ellipsoidal action space $\mathcal{E} = \{x \in \mathbb{R}^d : x^\top A^{-1} x = c\}$, with $c > 0$ and A positive-definite, is an LCH action space w.r.t. *every* $\theta \in \mathbb{R}^d$, as the Hessian of $f(x) = x^\top A^{-1} x$ is the constant p.d. matrix A^{-1}. However, an action space can be LCH just by being 'locally ellipsoidal'. As an example, consider an ellipsoid $\mathcal{E} = \{x \in \mathbb{R}^d : x^\top A^{-1} x = c\}$. Let θ be a bandit parameter and $x^* = \arg\max_{x \in \mathcal{E}} \langle x, \theta \rangle$ be optimal for θ. For some $\delta > 0$, let $\mathcal{B}_\delta$ be an open δ-ball in $\mathbb{R}^d$ containing x^*. Consider an action space $\mathcal{X}$ which coincides with $\mathcal{E}$ in the neighborhood $\mathcal{B}_\delta$ and is arbitrary outside it, i.e., $\mathcal{X} \cap \mathcal{B}_\delta = \mathcal{E} \cap \mathcal{B}_\delta$. It follows that $\mathcal{X}$ is LCH w.r.t. θ.

Theorem 2.2. *Let the action-space $\mathcal{X}$ be a Locally Constant Hessian(LCH) surface in $\mathbb{R}^{d-1}$ w.r.t. a bandit parameter θ. Let $\bar{G}_n = \mathbb{E}_\theta\left[\sum_{s=1}^n A_s A_s^\top\right]$, where A_s are arms in $\mathcal{X}$ drawn according to some bandit algorithm. For any bandit algorithm which suffers expected regret[2] at most $O(\sqrt{n})$,*

$$\lambda_{\min}(\bar{G}_n) = \Omega(\sqrt{n}) .$$

That is, there exists an n_0 and a constant $\gamma > 0$, such that for all $n \geqslant n_0$, $\lambda_{\min}(\bar{G}_n) \geqslant \gamma\sqrt{n}$.

The constant γ depends upon the condition number of the Hessian, the algorithmic constants hidden by $O()$, and the size of the bandit parameter θ. The constant n_0 depends on the algorithmic constants hidden by $O()$, the size of the neighbourbood over which the Hessian is constant, the size of the action domain $\|\mathcal{X}\|$, the singular value of the Hessian and the size of the bandit parameter θ.

Comparison with the result of Lattimore and Szepesvari (2017). The authors show a similar result for asymptotically large n. Specifically, they show that for a linear bandit with a *discrete* action-space (i.e., one for which the arms' suboptimality gaps are at least a positive constant), for any good (i.e., low-regret) bandit algorithm, it holds that

$$\liminf_{n \to \infty} \frac{\lambda_{\min}(\bar{G}_n)}{\log n} > 0.$$

In contrast, we prove a bound for any finite $n > n_0$. Moreover, our result applies to a broad category of action spaces.

Comparison with Bubeck et al. (2011). Bubeck et al. (2011) show that for a *finite-armed* bandit, an optimal cumulative regret ($O(\log T)$) algorithm must suffer $\Omega(\mathrm{poly}(1/T))$ simple regret, or equivalently, a probability of misidentifying the best arm for inference, in T time slots. Optimal inference (best arm) identification algorithms can, in contrast, obtain $e^{-\Omega(T)}$ simple regret at the cost of linear cumulative regret. Our result does not contradict this paper as our action space is *richer* (i.e., continuous) than that of a finite-armed bandit, leading to qualitatively different behavior: (a) On one hand, due to the minimum arm suboptimality gap in this setting being zero, the optimal cumulative regret rate is $O(\sqrt{T})$ (regret minimization is harder), (b) On the other hand, inference is easier with an optimal cumulative regret strategy with estimation error (for estimating θ^*) decaying as $O(T^{-1/4})$. Thus, for action-spaces which are locally "nice" enough as described above, any good regret algorithm must induce a well conditioned expected design matrix, and this has implications for parameter recovery as illustrated later.

[2]Our big-Oh and Omega notations throughout omit polylogarithmic dependencies for ease of presentation.

Dependence on dimension. Though our result does not explicitly indicate the dependence on the ambient (feature) dimension d, it is sensitive to it via the regret of the linear bandit algorithm in question. For example, for the OFUL algorithm (Abbasi-yadkori et al., 2011), we have a regret bound varying with the dimension as $O(d)$, while for Thompson Sampling (TS) we have a regret bound depending on d as $O(d^{3/2})$ (Agrawal and Goyal, 2013; Abeille and Lazaric, 2017). The constants n_0 and γ of Theorem 2.2 depend on d as $\Omega(d^2)$ and $\Omega(\frac{1}{d})$ for OFUL and as $\Omega(d^3)$ and $\Omega(\frac{1}{d^{3/2}})$ for TS respectively for a spherical action space. We provide more remarks on this in the experiment section.

2.1 Key Technique: Overview

In this section, for simplicity and insights, we provide a proof sketch for Theorem 2.2 for the spherical action space $\mathcal{X} = \mathcal{S}^{d-1}$. In Appendix 7, we first generalize these ideas for general ellipsoidal results, and finally prove for Locally Constant Hessian surfaces.

Proof Sketch. We start with the following information inequality for linear bandits (see Lemma 12.1 in appendix and Lattimore and Szepesvári (2020)):

$$\frac{1}{2}\|\theta - \theta'\|^2_{\bar{G}_n} \geqslant \mathrm{KL}\left(\mathrm{Ber}(\mathbb{E}_\theta[Z])||\mathrm{Ber}(\mathbb{E}_{\theta'}[Z])\right) \quad (1)$$

for any $\theta' \in \mathbb{R}^d$ and a measurable random variable $Z \in (0,1)$. As $\mathrm{span}(\mathcal{S}^{d-1}) = \mathbb{R}^d$, $\bar{G}_n$ will eventually be non-singular (Lattimore and Szepesvari, 2017). Hence, let the eigendecomposition of $\bar{G}_n$ be $\sum_{i=1}^d \lambda_i u_i u_i^\top$ with λ_i's arranged in descending order. Let us choose θ' to be $\theta + \alpha u_d$, where α is a step size to be determined. This gives the L.H.S. of (1) as

$$\|\theta - \theta'\|^2_{\bar{G}_n} = \alpha^2 \|u_d\|^2_{\bar{G}_n} = \alpha^2 \lambda_d \,. \quad (2)$$

Let us define an ε-neighbourhood about the optimal arm $\mathrm{OPT}(\theta)$ for a bandit parameter θ as

$$\mathrm{OPT}_\varepsilon(\theta) \triangleq \{x \in \mathcal{X} \mid x^\top \theta \geqslant \sup_x x^\top \theta - \varepsilon\} \,.$$

The step size α needs to be chosen such that θ and θ' are close in norm. At the same time, we need to ensure that the optimal arm for θ is sub-optimal for θ' and vice-versa. This motivates us to find an α such that

$$\mathrm{OPT}_\varepsilon(\theta) \cap \mathrm{OPT}_\varepsilon(\theta + \alpha u_d) = \emptyset \,. \quad (3)$$

We denote the number of times an arm in the ε-neighbourhood of θ is played as

$$\mathrm{N}_{\varepsilon,n}(\theta) \triangleq \sum\nolimits_{s=1}^{n} 1\{A_s \in \mathrm{OPT}_\varepsilon(\theta)\}$$

Now, we define Z in (1) as the fraction of times in n rounds that an arm in the ε-neighbourhood of θ is played, i.e., $Z \triangleq \frac{\mathrm{N}_{\varepsilon,n}(\theta)}{n}$. Then, as $\mathrm{OPT}_\varepsilon(\theta)$ and $\mathrm{OPT}_\varepsilon(\theta')$ are disjoint, $\mathbb{E}_\theta[Z]$ and $\mathbb{E}_{\theta'}[Z]$ will be different (in fact this is what a sub-linear regret algorithm is expected to do), and hence the right hand side of (1) will be positive. To choose the step size α, we exploit the geometry of the action-space. We refer to the Figure 1 from which it is clear that the amount α needed to perturb θ to θ' such that the disjoint condition (3) holds would depend upon the orientation of u_d with respect to $\mathrm{OPT}(\theta)$, which is θ itself, in our geometry. [3]

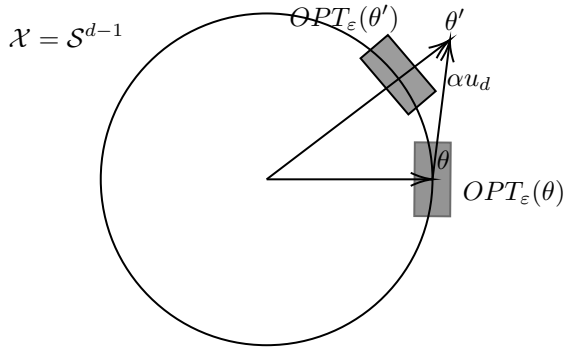


Figure 1: The arm set $\mathcal{S}^{d-1}$ with the bandit parameter $\theta = v$, the optimal arm. The direction u_d is approximately orthogonal to v. The task is to find α, the amount of perturbation of θ to θ' in the direction of u_d, so that the ε neighbourhood of both θ and θ' is disjoint

We use the Davis-Kahan matrix perturbation theorem (see Lemma 12.3) to show $\mathrm{OPT}(\theta)$ (denoted here on as v for notational simplicity) and u_d are approximately orthogonal. In its simplest form, the Theorem states that given two matrices A and H (symmetric), if $\lambda_1(A)$ and all of $\lambda_2(A+H), \cdots, \lambda_d(A+H)$ are separated by δ, then the eigenvector v corresponding to $\lambda_1(A)$ and the eigenvectors $u_2, \cdots, u_d$ corresponding to $\lambda_2(A+H), \cdots, \lambda_d(A+H)$, make an inner product of at most $\|H\|/\delta$, i.e., $|\max_{i=2,\cdots,d} v^\top u_i| \leqslant \|H\|/\delta$.

We use A as $nvv^\top$ and H as $\bar{G}_n - nvv^\top$ such that v is eigen-vector of $nvv^\top$ corresponding to $\lambda_1(A) = n$ and $u_2, \cdots, u_d$ are the eigenvectors of $A + H = \bar{G}_n$ corresponding to the eigenvalues $(\lambda_2(\bar{G}_n), \cdots, \lambda_d(\bar{G}_n))$. To find an upper bound of $\langle v, u_d \rangle$, we get an upper bound on the spectral norm of $\bar{G}_n - nvv^\top$ and the eigen-gap between $\lambda_1(nvv^\top) = n$ and $\lambda_2(\bar{G}_n), \cdots, \lambda_d(\bar{G}_n)$.

We know from Weyl's Lemma (see Lemma 12.4) that $\lambda_i(\bar{G}_n) \leqslant \lambda_i(nvv^\top) + \left\|\bar{G}_n - nvv^\top\right\| = \left\|\bar{G}_n - nvv^\top\right\|$, for $i \in \{2, \cdots, d\}$ as $\lambda_i(nvv^\top) = 0$ for all $i \in \{2, \cdots, d\}$. Thus it suffices to upper bound the spectral norm, $\left\|\bar{G}_n - nvv^\top\right\|$. To do so we decompose it into two cases: when the action played (A_s) belongs to the optimal set $\mathrm{OPT}_\varepsilon(\theta)$, and when it does not. This yields the follow-

[3] For tackling the ellipsoidal case, we need to change variables ('whitening') to bridge to the sphere case; see Appendix 7.

ing

$$\|\bar{G}_n - nvv^\top\| \leqslant \sup_{A_s \in \mathrm{OPT}_\varepsilon(\theta)} 2\|A_s - v\|\,\mathbb{E}\left[\mathrm{N}_{\varepsilon,n}(\theta)\right] + 4\mathbb{E}\left[n - \mathrm{N}_{\varepsilon,n}(\theta)\right], \tag{4}$$

where we have used the fact that arms in $\mathcal{S}^{d-1}$ has maximum norm of 1. We use the geometry of the action-space to determine that for any $A_s \in \mathrm{OPT}_\varepsilon(\theta)$, we have $\|A_s - v\| \leqslant O(\sqrt{\varepsilon})$ (see Appendix 7 for details). Now, we use the regret property of algorithm to show $\mathbb{E}_\theta[\mathrm{N}_{\varepsilon,n}(\theta)] \geqslant n - c\sqrt{n}/\varepsilon$, as

$$\begin{aligned} c\sqrt{n} &\geqslant \mathbb{E}\left[\sum_{s=1}^{n} \max_x x^\top\theta - A_s^\top\theta\right] \\ &\geqslant \mathbb{E}\left[\sum_{s=1}^{n} 1\{A_s \notin \mathrm{OPT}_\varepsilon(\theta)\}(\max_x x^T\theta - A_s^T\theta)\right] \\ &\geqslant \varepsilon(\mathbb{E}\left[n - \mathrm{N}_{\varepsilon,n}(\theta)\right]). \end{aligned} \tag{5}$$

Thus, we have an upper bound on the spectral norm of $\bar{G}_n - nvv^\top$ as $\|\bar{G}_n - nvv^\top\| \leqslant O\big(\sqrt{\varepsilon}n + \frac{\sqrt{n}}{\varepsilon}\big)$. Now choosing ε to be of the order $\frac{1}{\sqrt{n}}$ with a sufficiently large constant factor, we ensure the norm of $\|\bar{G}_n - nvv^\top\| \leqslant 0.01n$ for all n more than some finite n_0. Thus the eigen-gap is at least $0.99n$ and $|\langle v, u_d\rangle| \leqslant 1/99 \approx 0$. With this orientation of u_d, and the choice of ε, we prove that α has to be of the order of $\frac{1}{n^{1/4}}$ for $\mathrm{OPT}_\varepsilon(\theta) \cap \mathrm{OPT}_\varepsilon(\theta + O(1/n^{1/4})u_d) = \emptyset$. The details of this result are provided in the Appendix 7. Now, using our estimates of $\mathbb{E}_\theta[N_{\varepsilon,n}(\theta)]$ (see equation (5)), and by our choice of $\varepsilon = O(1/\sqrt{n})$ we have $\mathbb{E}_\theta[Z]$ close to 1 and $\mathbb{E}_{\theta'}[Z]$ close to 0. (See the previous discussion for our motivation to choose disjoint ε neighbourhood sets). This gives the right hand side of equation (1) a constant c (See appendix 7 for full details). Therefore combining with equation (2) we have $\lambda_d \geqslant c/\alpha^2 = \Omega(\sqrt{n})$. □

Remark 2.3. We observe in the above proof that the eigenvector corresponding to the minimum eigenvalue of $\bar{G}_n$ is approximately orthogonal to the direction of optimal arm. In fact, we show an even stronger fact: *every* eigenvector corresponding to each of the eigenvalues, starting from the second largest to the minimum, must lie in an approximately orthogonal space of the optimal direction.

3 MORE GENERAL ACTION SPACES

3.1 Hyper-Surfaces With Continuous Hessian

The basic conclusion of Theorem 2.2 (growth of the minimum eigenvalue of the design matrix) can be extended beyond LCH action spaces to hyper-surfaces of the form $\{x : f(x) = c\}$ where f is just C^2 and locally convex, without requiring a constant Hessian in a neighborhood. However, this comes potentially at the cost of the $\sqrt{n}$ growth rate of the minimum eigenvalue.

Definition 3.1 (Locally Convex surface)**.** Consider an action space $\mathcal{X} = \{x \in \mathbb{R}^d : f(x) = c\}$, where $f : \mathbb{R}^d \to \mathbb{R}$ is a C^2 function (i.e., all second order partial derivatives exist and are continuous). With $\mathrm{OPT}_\mathcal{X}(\theta)$ being the optimal arm defined as before, let the Hessian of f at $\mathrm{OPT}_\mathcal{X}(\theta)$, denoted as $\nabla^2 f(\mathrm{OPT}_\mathcal{X}(\theta))$, be positive definite. Then, $\mathcal{X}$ is said to be a Locally Convex surface.

Remark 3.2. Locally Convex action spaces are more general than LCH action spaces because they satisfy the property of the Hessian being positive definite and continuous at the optimal arm, and do away with the additional requirement of being constant in a neighbourhood of the optimal arm. This definition can also be extended to cover the situation when f is a convex function and the action space is defined as a sub-level set $\{x : f(x) \leqslant c\}$ as described in the next subsection.

For such action-spaces we show that the minimum eigenvalue still enjoys a polynomial growth rate albeit potentially at a rate less than $\sqrt{n}$.

Theorem 3.3. *Let $\mathcal{X}$ be a Locally Convex action space and $\bar{G}_n$, be the expected design matrix. For any bandit algorithm which suffers expected regret at most $O(\sqrt{n})$, there exists a real number s in the half-open interval $(0, \frac{1}{2}]$ such that*

$$\lambda_{\min}(\bar{G}_n) = \Omega(n^s).$$

The idea of the proof is to approximate a local neighbourhood around the optimal arm by an LCH surface, and to argue sufficient separation of optimal neighbourhoods in the LCH surface to ensure that the optimal neighbourhoods in the original neighbourhood are separated as well. The rest of the argument remains the same as for LCH spaces, and can be found in full detail in Appendix 8.

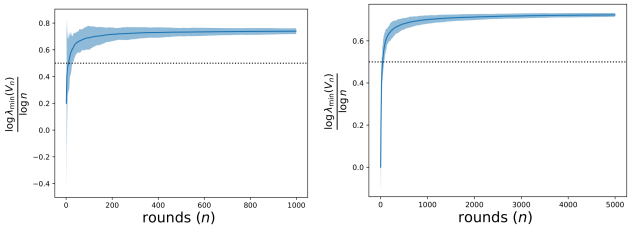
Remark 3.4. The exponent s defined in Theorem 3.3 depends explicitly on the geometry of the action-space. In general it depends on how well the surface approximates a LCH surface and in the Appendix 8 we provide a full methodology on how to calculate s. Here we just remark that for LCH action surfaces we recover the $\sqrt{n}$ growth rate.

3.2 Action Spaces With "Volume"

Our results of Theorem 2.2 and Theorem 3.3 continue to hold even if we replace the action space $\{x : f(x) = c\}$ by action spaces which are defined as sub-level sets for a C^2 convex function function $\{x : f(x) \leqslant c\}$. Note that popular bandit algorithms, like UCB or Thompson Sampling, takes the greedy action with respect to an optimistic estimate of the bandit parameter θ. This effectively reduces the playable action space to that on the surface (maximization of a linear function over a convex domain occurs at the boundary) and we can take the action space to be the surface. We can also prove this using the basic principles used in the proof of Theorem 2.2 and by extension

Theorem 3.3. For example, let $\mathcal{X}$ be a ball in $\mathbb{R}^d$, and θ be a bandit parameter. Without loss in generality, we can take $\|\theta\| = 1$ (otherwise, we can make a change of variable argument like in the ellipsoidal case to show that the result holds, now including a factor of $\|\theta\|$). The crux of the proof relies on two factors. (a) First, to show that u_d, the eigen-vector, corresponding to the minimum eigen-value is approximately perpendicular to $\mathrm{OPT}_{\mathcal{X}}(\theta)$. (b) Second to show that $\mathrm{OPT}_{\varepsilon,\mathcal{X}}(\theta) \cap \mathrm{OPT}_{\varepsilon,\mathcal{X}}(\theta + \alpha u_d) = \emptyset$. For the first part, the main idea is to bound the norm $\|A_s - \mathrm{OPT}_{\mathcal{X}}(\theta)\|$ for any $A_s \in \mathrm{OPT}_{\varepsilon,\mathcal{X}}(\theta)$. To do so, note that we had used $\|A_s\| = 1$ for the surface action spaces. Now, when action space is a ball, we can take an upper bound on $\|A_s\| \leqslant 1$ while the remainder of the proof follows as before. For the second part, note that the boundary of $\mathrm{OPT}_{\varepsilon,\mathcal{X}}(\theta)$ for a ball is simply $\mathrm{OPT}_{\varepsilon,\mathcal{S}^{d-1}}(\theta)$ as defined earlier. Thus ensuring the ε-optimal neighbourhoods of the sphere are separated ensures the separation of the ε-optimal neighbourhoods of the ball. Now we can proceed as before and find the order of the separation to get the same conclusions as in Theorem 2.2 and Theorem 3.3.

4 NUMERICAL RESULTS



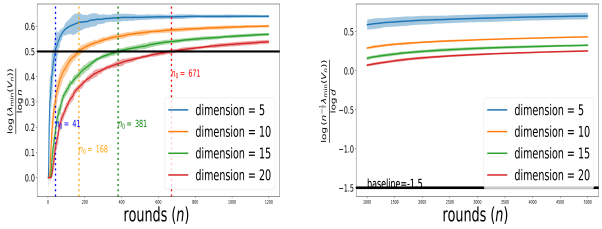
(a) $d = 3$ (b) $d = 5$

(c) $d = 10$ (d) $d = 20$

Figure 2: Scaling of the minimum eigenvalue of design matrix generated by the Thompson sampling algorithm with time. The plots represent averages over 20 independent trials. The X axis denotes the number of rounds n and the Y-axis denotes $\frac{\log \lambda_{\min}(V_n)}{\log n}$. The dotted black line is the constant (exponent) $1/2$. Note how $\frac{\log \lambda_{\min}(V_n)}{\log n}$ crosses and settles to a value above $1/2$ in each case.

[4]In this section, we carry out experiments to understand the rate of growth of the minimum eigenvalue. We find, using a "good" bandit algorithm and a "good" action space, that the minimum eigen-value of the design matrix grows at a rate of more than $\sqrt{n}$, with high probability. We use the well-known linear Thompson Sampling (TS) algorithm (Agrawal and Goyal, 2013) as a representative algorithm for linear bandits. It is well known that the regret of TS is $\tilde{O}(\sqrt{n})$ with high probability, where the $\tilde{O}$, hides logarithmic factors. We use a spherical action space $\mathcal{X} = \mathcal{S}^{d-1}$ as a candidate for the LCH action surfaces and we fix the unknown bandit parameter θ at e_1. We use a regularization parameter of $\lambda = 1.0$.

We observe from Figure 2 the dispersion and mean trend of the sample trajectory dependent quantity $\log \lambda_{\min}(V_n)/\log n$. We expect to see to see this term cross the benchmark line of $1/2$ for all n more than a finite n_0. In order to demonstrate the high probability phenomenon, we form a high confidence band of the mean observation of $\log \lambda_{\min}(V_n)/\log n$ with three standard deviations of width. We plot this confidence band and we show that the lower envelope of the band remains more than $1/2$. It could be observed that the trend of $\log \lambda_{\min}(V_n)/\log n$ remains increasing, but this could be accounted for the hidden logarithmic factors in the regret of the TS algorithm itself.



(a) n_0 trend with dimension d (b) γ trend with dimension d

Figure 3: Scaling of the theoretical constants γ and n_0 with the dimension d. Note that n_0 increases with the dimension d whereas γ decreases with the dimension d. Since the algorithmic constants varies appropriately with the dimension d, in the case of TS, our experimental results corroborates with our theory.

We note that for the spherical action space, the constants n_0 and γ of Theorem 2.2 would vary with the dimension d as $\Omega(d^3)$ and $\Omega(1/d^{3/2})$ for TS (see proof of Theorem 2.2). In Figure 3 we explicitly calculate the the mean n_0, as the time above and over which $\log \lambda_{\min}(V_n)/\log n$ crosses the 0.5 bench mark, and observe that our theoretical lower bound of $n_0 = \Omega(d^3)$ is a loose lower bound and that in practical settings the n_0 can be much less than the order d^3. Similarly for γ, defined as $\liminf \frac{\lambda_{\min}(V_n)}{\sqrt{n}}$, is plotted as the mean trend of $\frac{\log \lambda_{\min}(V_n)/\sqrt{n}}{\log d}$ for rounds n more than 1000, as a reasonable upper estimate of n_0 and see that this quantity remains well above the benchmark of -1.5, thus showing that $\gamma = \Omega(1/d^{3/2})$, is indeed a loose lower bound.

5 APPLICATIONS

We now give two instances where the eigenvalue bound obtained in Theorem 2.2 can be used to its advantage due to

[4]All experiments can be found in `https://github.com/Debangshu93/analytic`

its implication of fast inference about the unknown parameter. We first start by noting that for parameter estimation we need a high probability version of the Theorem 2.2.

In Theorem 2.2, the eigenvalue bound is shown on the expected design matrix. In the simulations (Section 4), we observe (over multiple trials) that the lower envelope of $\lambda_{\min}(V_n)$ scales as $\Omega(\sqrt{n})$, which hints towards a high probability result. It turns out that under an appropriately defined *stability* condition, the expectation bound can be translated to a high probability bound. In Appendix 11 we formally define this stability condition and its consequences. However, in this section, for clarity of exposition, we take the high probability bound as granted, and using this, we consider a couple of applications in bandit model selection and clustering. Recall that we define the Gram matrix as $V_n = \sum_{s=1}^{n} A_s A_s^\top$.

Assumption 5.1. Given an action-set $\mathcal{X}$ which belongs to the class of *Locally Constant Hessian* surfaces (see Definition 2.1) for a bandit parameter θ, for any bandit algorithm which suffers regret, $R_n(\theta)$, at most $O(\sqrt{n})$, there exists a $\delta \in (0, 1]$, such that

$$\lambda_{\min}(V_n) = \Omega(\sqrt{n}) \ ,$$

with probability at least $1 - \delta$.

Remark 5.2. Using the aforementioned *stability* condition we can get similar high probability versions for *Locally Convex* action spaces. However as discussed earlier this could potentially loose the $\sqrt{n}$ growth rate of the minimum eigen-value. For ease of exposition, we shall assume in this section that the action-space in consideration is an LCH surface unless mentioned otherwise.

5.1 Model Selection in Linear Bandits

A natural measure of complexity of linear bandits is the norm of the parameter θ. Typically it is assumed that θ lies in a norm ball with known radius, i.e., $\|\theta\| \leqslant b$ (Abbasi-yadkori et al., 2011; Chatterji et al., 2020). This leads to non-adaptive algorithms and the algorithms use b as a proxy for the problem complexity, which can be a huge over-estimate. We analyze a linear bandit algorithm, that adapts to the problem complexity $\|\theta\|$, and as a result, the regret obtained will depend on $\|\theta\|$. Specifically, we start with an over estimate of $\|\theta\|$, and successively refine this estimate over multiple epochs. We show that this refinement strategy yields a consistent sequence of estimates of $\|\theta\|$, and as a consequence, our regret bound depends on $\|\theta\|$, but not on its upper bound b. Ghosh et al. (2021a) considers a similar model selection problem in a related setting of *stochastic contextual bandits*, with restrictive assumptions on the contexts. We show that model selection guarantees continue to hold even in the context free setting without any additional assumption by using the result in Theorem 2.2 and the subsequent Assumption 5.1. Before that, we restate the algorithm of Ghosh et al. (2021a).

Algorithm 1 Adaptive Linear Bandit (ALB)

1: **Input:** An upper bound b of $\|\theta\|$, initial epoch length n_1, confidence level $\delta \in (0, 1]$
2: Initialize estimate of $\|\theta\|$ as $b^{(1)} = b$, set $\delta_1 = \delta$
3: **for** epochs $i = 1, 2 \ldots$ **do**
4: Play OFUL with norm estimate $b^{(i)}$ for n_i rounds with confidence level δ_i
5: Refine estimate of $\|\theta\|$ as $b^{(i+1)} = \max_{\theta \in \mathcal{B}_{n_i}} \|\theta\|$
6: Set $n_{i+1} = 2n_i$, $\delta_{i+1} = \frac{\delta_i}{2}$
7: **end for**

5.1.1 Algorithm and its Regret Bound

The algorithm – called Adaptive Linear Bandits (ALB) – uses the OFUL algorithm of Abbasi-yadkori et al. (2011) as a black-box. The learning proceeds in epochs – at each epoch $i \geqslant 1$, OFUL is run for $n_i = 2^{i-1} n_1$ episodes with confidence level $\delta_i = \frac{\delta}{2^{i-1}}$ and norm estimate $b^{(i)}$, where the initial epoch length n_1 and confidence level δ are parameters of the algorithm. We begin with b as an initial (over) estimate of $\|\theta\|$, and at the end of i-th epoch, based on the confidence set $\mathcal{B}_{n_i}$ build by OFUL, we choose the new estimate as $b^{(i+1)} = \max_{\theta \in \mathcal{B}_{n_i}} \|\theta\|$. We argue that this sequence of estimates is indeed consistent, and as a result, the regret depends on $\|\theta\|$.

Now, we present the main result of this section. We show that, by virtue of Theorem 2.2 and the subsequent Assumption 5.1, the norm estimates $b^{(i)}$ computed by ALB (Algorithm 1) indeed converges to the true norm $\|\theta\|$ at an exponential rate with high probability.

Lemma 5.3 (Convergence of norm estimates)**.** *Suppose Assumption 5.1 holds. Also, suppose that, for any $\delta \in (0, 1]$, the length n_1 of the initial epoch satisfies*

$$n_1 \geqslant \max\left\{ n_0, C\, d^2 \left(\max\{p, q\}\, b^{(1)} \right)^4 \right\},$$

where $p = O\left(\frac{1}{\sqrt{\gamma_0}}\right)$, $q = O\left(\sqrt{\frac{\log(n_1/\delta)}{\gamma_0}}\right)$, and $C > 1$ is some sufficiently large universal constant. Then, with probability exceeding $1 - 4\delta$, the sequence $\{b^{(i)}\}_{i=1}^{\infty}$ converges to $\|\theta\|$ at a rate $\mathcal{O}\left(\frac{i}{2^{i/4}}\right)$, and $b^{(i)} \leqslant c_1 \|\theta\| + c_2$, where $c_1, c_2 > 0$ are universal constants.

Comparison with prior work. Ghosh et al. (2021a) consider the setting stochastic contextual linear bandits, and show that the norm estimates converge at a rate $\mathcal{O}\left(\frac{i}{2^{i/2}}\right)$. This seemingly better rate, however, comes at the cost of restrictive assumptions on the contexts. Specifically, they assume that the minimum eigenvalue of the conditional covariance matrix (given observations up to time t) is bounded away from zero. This restriction on the contexts severely limits the applicability of their result. In contrast, we obtain a slightly worse, but still exponential, conver-

gence rate $\mathcal{O}\left(\frac{i}{2^{i/4}}\right)$, albeit with a milder assumption in the context of our main result (Theorem 2.2).

Armed with the above result, we can now prove a sublinear regret bound of ALB.

Corollary 5.4 (Cumulative regret of ALB)**.** *Fix any $\delta \in (0,1]$, and suppose that the hypothesis of Lemma 5.3 holds. Then, with probability exceeding $1-6\delta$, ALB enjoys the regret bound*

$$R(n) = \tilde{\mathcal{O}}\Big(\|\theta\|\sqrt{dn\log n_1} + d\sqrt{n\log n_1 \log(n_1/\delta)}\Big) \leqslant \tilde{\mathcal{O}}\left((\|\theta\|+1)\, d\sqrt{n\log(n/\delta)}\right),$$

where $\tilde{\mathcal{O}}$ hides a $\mathsf{polylog}(n/n_1)$ *factor.*

Remark 5.5. Note that the above regret depends on $\|\theta\|$ and hence is adaptive to the norm complexity. Furthermore, the term $(\|\theta\|+1)$ can be reduced to $(\|\theta\|+\varepsilon)$ for an arbitrary $\varepsilon > 0$ by increasing the length of initial epoch in ALB.

Lemma 5.3 and Corollary 5.4 together highlight the advantage of having an eigenvalue bound as in Theorem 2.2 in this specific setting of model selection due to its implication in fast inference and regret minimization, simultaneously.

5.2 Clustering without Exploration in Linear Bandits

We consider a multi-agent system with N users, which communicate with a "center". Moreover, the agents are partitioned into k clusters. We aim to identify the cluster identity of each user so that collaborative learning is possible within a cluster. Note that previously (Gentile et al., 2014a; Ghosh et al., 2021b) tackled the problem of online clustering in a stochastic contextual bandit framework, where the players have additional context information to make a decision. More importantly, in the mentioned works, several restrictive assumptions were made on the behavior of the stochastic context, such as, based on the observation upto time t, (a) the conditional mean is 0, (b) the conditional covariance matrix is positive definite, time uniform with the minimum eigenvalue bounded away from 0 and (c) the conditional variance of the contexts projected in a fixed direction is bounded. Apart from these three, there are a few additional technical assumptions made, see (Gentile et al., 2014a, Lemma 1) for example. We believe it is difficult to find natural examples of contexts where all these assumptions are satisfied simultaneously, and the authors of the aforementioned papers also do not provide any.

To this end, we provide clustering guarantee without these restrictive assumptions. We only require the action space satisfying Definition 2.1, which includes standard spaces like spheres and ellipsoids. Furthermore, we obtain the underlying clustering without *pure (forced) exploration*—thanks to the lower bound of Theorem 2.2 and the subsequent Assumption 5.1. This is quite useful in recommendation systems and advertisement placement, where forced exploration is often very tricky and not at all desired. We stick to the framework of standard linear bandits, and we assume a clustering framework, where the user parameters $\{\theta_1^*, \ldots, \theta_N^*\}$ are partitioned into k groups. All users in a cluster have the same preference parameter, and hence, without loss of generality, for all $j \in [k]$, we denote θ_j^* as the preference parameter for cluster j. We employ the standard OFUL of Abbasi-yadkori et al. (2011) as our learning algorithm. Note that, by virtue of Theorem 2.2 and the subsequent Assumption 5.1, we can estimate the underlying parameter for any agent $i \in [N]$ in ℓ_2 norm. We use this information to do the clustering task. Our goal here is to propose an algorithm that finds the correct clustering of all N users while simultaneously obtain low regret in the process. Before providing the algorithm and regret guarantee for clustering let us define the (minimum) separation parameter as $\Delta = \min_{i,j;i\neq j} \|\theta_i^* - \theta_j^*\|$. Note that if Δ is large, the problem is easier to solve, and vice versa. Hence, Δ is also called the SNR (signal to noise ratio, since we assume that the noise variance is unity) of the problem.

5.2.1 Algorithm and its regret bound

We now present the clustering algorithm, formally given in Algorithm 2. We let all the agents play the learning algorithm OFUL of Abbasi-yadkori et al. (2011) for n rounds. Note that, with a lower bound on the minimum eigenvalue of the Gram matrix (Theorem 2.2, Assumption 5.1), we now can compute how close the estimated parameters $\{\widehat{\theta}^{(i)}\}_{i=1}^N$ are to the true parameters in ℓ_2 norm. The choice of norm is important here. Note that OFUL also obtains an estimate of the underlying parameter. However, that closeness is only guaranteed in a problem dependent matrix norm, which can not be directly used for inference problems like clustering. Note that Theorem.. basically converts this problem dependent norm to (an universal) ℓ_2 norm closeness guarantee, which enables us to perform clustering. After n rounds, the center performs a pairwise clustering with threshold γ. If the pairwise distance between two estimates are less than γ they are estimated to belong to the same cluster, otherwise different. This procedure is given in the EDGE-CLUSTER subroutine. With properly chosen threshold, we show that Algorithm 2 clusters the agents correctly with high probability.

Lemma 5.6. *Suppose we choose the threshold $\eta = \frac{4}{n^{1/4}}\sqrt{\frac{2d\log(n/\delta)}{\gamma\log(d/\delta)}}$, and the separation Δ satisfies $\Delta > 2\eta = \frac{8}{n^{1/4}}\sqrt{\frac{2d\log(n/\delta)}{\gamma\log^{1/2}(d/\delta)}}$. Then, Algorithm 2 clusters all N agents correctly with probability at least $1-4\binom{N}{2}\delta$. Furthermore, the regret of any agent $i \in [N]$ is given by $R_i \leqslant \mathcal{O}(d\sqrt{n}\log(1/\delta))$ with probability at least $1-\delta$.*

Note that the separation decays with n, and for a large n, this is just a mild requirement. Also, our clustering algorithm does not incur regret from pure exploration, and the regret is just from playing the OFUL algorithm.

Algorithm 2 Cluster without Explore

Input: No. of users N, learning horizon n, high probability parameter δ, threshold η

Individual Learning Phase

All agents play $\texttt{OFUL}(\delta)$ independently for n rounds

$\{\widehat{\theta}^{(i)}\}_{i=1}^{N} \leftarrow$ All agents' estimates at the end of round n and send to the center

Cluster the Users at center

$\texttt{User-Cluster} \leftarrow \texttt{EDGE-CLUSTER}(\{\widehat{\theta}^{(i)}\}_{i=1}^{N}, \eta)$

EDGE-CLUSTER

Input: All estimates $\{\widehat{\theta}^{(i)}\}_{i=1}^{N}$, threshold $\eta \geqslant 0$.

Construct an undirected Graph G on N vertices as follows: $||\widehat{\theta}_i^* - \widehat{\theta}_j^*|| \leqslant \eta \Leftrightarrow i \sim_G j$

$\mathcal{C} \leftarrow \{C_1, \cdots, C_k\}$ all strongly connected components of G

Return : $\mathcal{C}$

Remark 5.7 (Clustering gain). We run Algorithm 2 upto to the instant where all users are clustered correctly wit high probability. In particular, we do not characterize the clustering gain of Algorithm 2 since this is not the main focus of the paper. In order to see the gain, one needs to do the following: after the agents are clustered, the center treats each cluster as a single agent, and averages reward from all users in the same cluster. This ensures standard deviation of the resulting noise goes down by a factor of $1/\sqrt{N'}$ (clustering gain, similar to Gentile et al. (2014b); Ghosh et al. (2021b)), where N' is the number of users in the cluster.

Clustering without separation assumption on Δ. In the above result, we assume that the separation Δ satisfies the above condition. Instead, if the learner knows Δ, one can set

$$\frac{n}{\log^2(n/\delta)} \geqslant C\frac{d^2}{\gamma^2\,\Delta^4}\left(\frac{1}{\log(d/\delta)}\right) = \tilde{\Omega}(\frac{d^2}{\gamma^2\,\Delta^4}),$$

and the threshold $\eta = \Delta/2$ to obtain the same result.

6 CONCLUSION

We present a minimum eigenvalue bound on the expected design matrix—a theoretical extension to what is already known for discrete action space. In particular, we show that in action spaces with locally "nice" curvature properties, the above-mentioned minimum eigenvalue grows at the order of $\sqrt{n}$. We show that this eigenvalue bound enables us to obtain inference and minimize regret simultaneously, in the linear bandit setup. We then apply our findings in two practical applications, *bandit clustering* and *model selection*. We emphasize these results pave the way for new research in the area of stability and robustness of practical bandit algorithms. Such ideas are not new and the study of differentially private algorithms (Dwork et al., 2014) focuse on this area precisely. Another line of research is to study the question of asymptotic optimality for algorithms based on the principle of optimism in the continuous action domain.

References

Yasin Abbasi-yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24, pages 2312–2320. Curran Associates, Inc., 2011.

Marc Abeille and Alessandro Lazaric. Linear thompson sampling revisited. In *Artificial Intelligence and Statistics*, pages 176–184. PMLR, 2017.

Alekh Agarwal, Haipeng Luo, Behnam Neyshabur, and Robert E Schapire. Corralling a band of bandit algorithms. In *Conference on Learning Theory*, pages 12–38. PMLR, 2017.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pages 127–135. PMLR, 2013.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science*, 412(19):1832–1852, 2011.

Niladri Chatterji, Vidya Muthukumar, and Peter Bartlett. Osom: A simultaneously optimal algorithm for multi-armed and linear contextual bandits. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1844–1854. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/chatterji20b.html.

Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. In *International Conference on Machine Learning*, pages 844–853. PMLR, 2017.

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

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

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407, 2014.

Eyal Even-Dar, Shie Mannor, Yishay Mansour, and Sridhar Mahadevan. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(6), 2006.

Dylan J Foster, Akshay Krishnamurthy, and Haipeng Luo. Model selection for contextual bandits. *Advances in Neural Information Processing Systems*, 32:14741–14752, 2019.

Claudio Gentile, Shuai Li, and Giovanni Zappella. Online clustering of bandits. In *International Conference on Machine Learning*, pages 757–765. PMLR, 2014a.

Claudio Gentile, Shuai Li, and Giovanni Zappella. Online clustering of bandits. In *International Conference on Machine Learning*, pages 757–765. PMLR, 2014b.

Avishek Ghosh, Abishek Sankararaman, and Ramchandran Kannan. Problem-complexity adaptive model selection for stochastic linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 1396–1404. PMLR, 2021a.

Avishek Ghosh, Abishek Sankararaman, and Kannan Ramchandran. Collaborative learning and personalization in multi-agent stochastic linear bandits. *arXiv preprint arXiv:2106.08902*, 2021b.

Botao Hao, Tor Lattimore, and Csaba Szepesvari. Adaptive exploration in linear contextual bandit. In *International Conference on Artificial Intelligence and Statistics*, pages 3536–3545. PMLR, 2020.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17(1):1–42, 2016.

Sanath Kumar Krishnamurthy and Susan Athey. Optimal model selection in contextual bandits with many classes via offline oracles. *arXiv preprint arXiv:2106.06483*, 2021.

Tor Lattimore and Csaba Szepesvari. The end of optimism? an asymptotic analysis of finite-armed linear bandits. In *Artificial Intelligence and Statistics*, pages 728–737. PMLR, 2017.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Aldo Pacchiano, Christoph Dann, Claudio Gentile, and Peter Bartlett. Regret bound balancing and elimination for model selection in bandits and rl. *arXiv preprint arXiv:2012.13045*, 2020.

Marta Soare, Alessandro Lazaric, and Rémi Munos. Best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 27:828–836, 2014.

Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.

Gilbert W Stewart. Matrix perturbation theory. 1990.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12 (4):389–434, 2012.

Joel A Tropp. An introduction to matrix concentration inequalities. *arXiv preprint arXiv:1501.01571*, 2015.

Michal Valko, Nathaniel Korda, Rémi Munos, Ilias Flaounas, and Nelo Cristianini. Finite-time analysis of kernelised contextual bandits. *arXiv preprint arXiv:1309.6869*, 2013.

7 LOCALLY CONSTANT HESSIAN SURFACES (PROOF OF THEOREM 2.2)

In this section we shall prove Theorem 2.2. We shall divide the proof of the theorem in three parts. In Section 7.1 we shall prove it for spherical surface action sets. In Section 7.2 we shall extend the proof for ellipsoidal surface action sets and finally in Section 7.3 we prove the result for the generalization to action spaces which are LCH surfaces.

7.1 Spherical Action Sets

Theorem 7.1. *Let the set of arms be $\mathcal{X} = \mathcal{S}^{d-1} := \{x \in \mathbb{R}^d : \|x\| = 1\}$, the surface of the d dimensional unit sphere. Let $\bar{G}_n = \mathbb{E}_\theta\left[\sum_{s=1}^n A_s A_s^\top\right]$, where θ is a bandit parameter and A_s are arms in $\mathcal{X}$ drawn according to some bandit algorithm. For any bandit algorithm which suffers expected regret, $R_n(\theta)$, at most $O(\sqrt{n})$,*

$$\lambda_{min}(\bar{G}_n) = \Omega(\sqrt{n}) \ .$$

That is there exists constants $\gamma > 0$ and a finite time n_0 such that for all $n \geqslant n_0$, we have $\lambda_{min}(\bar{G}_n) \geqslant \gamma\sqrt{n}$.

Proof. We start with a result which is standard while proving such lower bounds (Lattimore and Szepesvari (2017); Lattimore and Szepesvári (2020); Hao et al. (2020)) which is using the measure change inequality of Kaufmann et al. (2016) (see Lemma 12.1) combined with the Divergence Decomposition Lemma under the Linear Bandit Setup Lattimore and Szepesvári (2020) (see Lemma 12.2) to get

$$\frac{1}{2}\|\theta - \theta'\|^2_{\bar{G}_n} \geqslant \mathrm{KL}(\mathrm{Ber}(\mathbb{E}_\theta[Z])||\mathrm{Ber}(\mathbb{E}_{\theta'}[Z])) \ . \tag{6}$$

For any time n we have an eigenvalue decomposition of $\bar{G}_n$ as

$$\bar{G}_n = \sum_{i=1}^d \lambda_i u_i u_i^\top \ , \tag{7}$$

where $\lambda_1 \geqslant \lambda_2 \geqslant \cdots, \geqslant \lambda_d \triangleq \lambda_{min}(\bar{G}_n)$. From the regret property we have, for all M, such that for any bandit parameter θ satisfying $\|\theta\| \leqslant 2M$, there exists a constant $c > 0$ such that expected regret $\mathbb{E}\left[R_n(\theta)\right] \leqslant c\sqrt{n}$ for any n. Without loss of generality let us fix the bandit parameter $\theta = e_1$ (Otherwise we can always rename our coordinate system s.t. in the new system $\theta = e_1$) and $\theta' = \theta + \alpha u_d$ for some α to be decided later. This gives,

$$\|\theta - \theta'\|^2_{\bar{G}_n} = \alpha^2 \lambda_d \ . \tag{8}$$

ε-neighbourhood optimal arms We shall define $\mathrm{OPT}_\varepsilon(\theta)$ as the set of $\varepsilon-$ suboptimal arms with respect to θ

$$\mathrm{OPT}_\varepsilon(\theta) \triangleq \{x \in \mathcal{X} \mid x^T\theta \geqslant \sup_x x^T\theta - \varepsilon\}$$

We shall with abuse of notation denote $\mathrm{OPT}_0(\theta) = \mathrm{OPT}(\theta) = e_1$. Let us also define $\mathrm{N}_{\varepsilon,n}(\theta)$ to be the number of times in time n that an arm in $\mathrm{OPT}_\varepsilon(\theta)$ has been played.

$$\mathrm{N}_{\varepsilon,n}(\theta) \triangleq \sum_{s=1}^n 1\{A_s \in \mathrm{OPT}_\varepsilon(\theta)\}.$$

Under the hypothesis that for any bandit parameter θ satisfying $\|\theta\| \leqslant 2$, there exists a constant $c > 0$ such that $\mathbb{E}R_n(\theta) \leqslant c\sqrt{n}$, we have

$$c\sqrt{n} \geqslant \mathbb{E}R_n(\theta) \geqslant \mathbb{E}[\sum_{s=1}^n 1\{A_s \notin \mathrm{OPT}_\varepsilon(\theta)\}(\max_x x^T\theta - A_s^T\theta)] \geqslant \varepsilon(\mathbb{E}[n - \mathrm{N}_{\varepsilon,n}(\theta)]) \ .$$

Rearranging, we get $\mathbb{E}_\theta[\mathrm{N}_{\varepsilon,n}(\theta)] \geqslant n - \frac{c\sqrt{n}}{\varepsilon}$.

Showing that u_d is approximately perpendicular to $\mathrm{OPT}(\theta)$. Let $v \equiv \mathrm{OPT}(\theta)$. Then, by definition $v = \theta = e_1$. Let us decompose $\bar{G}_n$ into $nvv^\top$, the unperturbed component, A, and $\bar{G}_n - nvv^\top$ the perturbation, H, as per our notation of the perturbation lemmas Davis-Kahan(see Lemma 12.3) and Weyl's lemma (see Lemma 12.4).

$$\bar{G}_n = \underbrace{nvv^\top}_{A} + \underbrace{\bar{G}_n - nvv^\top}_{H,\text{"perturbation"}} \tag{9}$$

Note that both the unperturbed matrix A, the perturbation matrix H are hermitian. Thus we can use Weyl's Lemma (Lemma 12.4). This gives us for all $i = 2, 3, \cdots, d$

$$\lambda_i(\bar{G}_n) \leqslant \underbrace{\lambda_i(A)}_{=0} + \lambda_{\max}(H), \tag{10}$$

where we have used $\lambda_i(e_1 e_1^\top) = 0$ for all $i = 2, 3, \cdots, d$. Now note from the definition of the perturbation matrix H and our notation for the spectral norm

$$\lambda_{\max}(H) = \|\bar{G}_n - nvv^\top\| = \left\| \mathbb{E}\left[\sum_{s=1}^{n} A_s A_s^\top - vv^\top \right] \right\|,$$

where in the last equation we have expanded the definition of $\bar{G}_n$. We now decompose the last sum into two parts, all arms, A_s which belong in the group $\mathrm{OPT}_\varepsilon(\theta)$ and those that do not. We are doing this because we shall see that all arms that belong in the group of $\mathrm{OPT}_\varepsilon(\theta)$ will have norm $\|A_s - v\|$ small than those which do not. We get

$$\lambda_{max}(H) = \left\| \mathbb{E}\left[\sum_{s: A_s \in \mathrm{OPT}_\varepsilon(\theta)} (A_s A_s^\top - vv^\top) + \sum_{s: A_s \notin \mathrm{OPT}_\varepsilon(\theta)} (A_s A_s^\top - vv^\top) \right] \right\|. \tag{11}$$

Now, by interchanging norm and expectation and using triangle inequality for norms we have that

$$\lambda_{max}(H) \leqslant \sum_{s: A_s \in \mathrm{OPT}_\varepsilon(\theta)} \mathbb{E}\left[\left\| A_s A_s^\top - vv^\top \right\| \right] + \sum_{s: A_s \notin \mathrm{OPT}_\varepsilon(\theta)} \mathbb{E}\left[\left\| (A_s A_s^\top - vv^\top) \right\| \right] . \tag{12}$$

Thus, we see that we need to control the norm of $A_s A_s^\top - vv^\top$ for two separate cases. Before we do that, let us first rearrange $A_s A_s^\top - vv^\top$ to get

$$\|A_s A_s^\top - vv^\top\| = \|(A_s - v)A_s^\top + v(A_s - v)^\top\| \leqslant \|A_s - v\|(\|A_s\| + \|v\|) \leqslant 2\|A_s - v\| \tag{13}$$

where in the last equation we have used the fact that both A_s and v belongs to $\mathcal{S}^{d-1}$. Note that here we need to have kept $\mathcal{S}^{d-1}$ and a generic upper bound on the arm set $\mathcal{X}$ would have sufficed. Thus, we need to control $\|A_s - v\|$ in the two cases. When $A_s \notin \mathrm{OPT}_\varepsilon(\theta)$ we can use a generic upper bound on the bounded property of the Action Space to get $\|A_s - v\| \leqslant 2$. Now, let us consider the case when $A_s \in \mathrm{OPT}_\varepsilon(\theta)$. Then by definition of $\mathrm{OPT}_\varepsilon(\theta)$ we have

$$A_s^\top \theta \geqslant v^\top \theta - \varepsilon = 1 - \varepsilon \tag{14}$$

where we have again used that $v = e_1 = \theta$ because of the action-space $\mathcal{X}$. Note here again that for a general geometry $v^\top \theta$ would still be some constant depending upon the geometry of the action-space.

Now let us consider the 2 orthogonal components of such an A_s, $A_s - (A_s^\top v)v$ and $(A_s^\top v)v$. We have by Pythagorean Theorem

$$\|A_s\|^2 = \|(A_s^\top v)v\|^2 + \|A_s - (A_s^\top v)v\|^2.$$

Therefore again utilizing that $\|\|A_s\|^2 = 1$ (although a generic bound would have sufficed) and rearranging

$$\|A_s - (A_s^\top v)v\|^2 = 1 - \|(A_s^\top v)v\|^2.$$

Now we utilize the fact that in the spherical geometry we can exactly compute $A_s^\top v = A_s^\top \theta \geqslant 1 - \varepsilon$, (see Equation (14)), to get

$$1 - \|(A_s^\top v)v\|^2 \leqslant 1 - (1 - \varepsilon)^2 = (2\varepsilon - \varepsilon^2) ,$$

where we have used homogenity of norms in the inequality. For other geometry we need to use the geometry of the set $\mathrm{OPT}_\varepsilon(\theta)$ to estimate $A_s^\top v$.

Now, we can conclude that for any $A_s \in \mathrm{OPT}_\varepsilon(\theta)$, by virtue of orthogonality between $A_s - (A_s^\top v)v$ and $v - (A_s^\top v)v$, we have using the Pythagorean theorem

$$\|A_s - v\|^2 = \|A_s - (A_s^\top v)v\|^2 + \|v - (A_s^\top v)v\|^2.$$

Now using the fact that $v - (A_s^\top v)v = (1 - A_s^\top v)v \leqslant \varepsilon v$, we get

$$\|A_s - (A_s^\top v)v\|^2 + \|v - (A_s^\top v)v\|^2 \leqslant (2\varepsilon - \varepsilon^2) + \varepsilon^2 = 2\varepsilon \ .$$

Thus we have for any $A_s \in \mathrm{OPT}_\varepsilon(\theta)$ the following estimate

$$\|A_s - v\| \leqslant \sqrt{2\varepsilon} \ . \tag{15}$$

Note that by changing the geometry we will have different constants but still of the same $\sqrt{\varepsilon}$ order.

Thus we have from Equations (12), (13) and (15) along with the trivial bound of $\|A_s - v\| \leqslant 2$, the following estimate on the spectral norm of the perturbation matrix H :

$$\lambda_{max}(H) \leqslant 2\sqrt{2\varepsilon}\mathbb{E}_\theta[\mathrm{N}_{\varepsilon,n}(\theta)] + 4\mathbb{E}_\theta[n - \mathrm{N}_{\varepsilon,n}(\theta)] \ ,$$

where we have used the definitions of $\mathrm{N}_{\varepsilon,n}(\theta) = \sum_{s=1}^n 1\{A_s \in \mathrm{OPT}_\varepsilon(\theta)\}$ and $n - \mathrm{N}_{\varepsilon,n}(\theta) = \sum_{s=1}^n 1\{A_s \notin \mathrm{OPT}_\varepsilon(\theta)\}$. Then using a crude upper bound of n on $\mathbb{E}_\theta[\mathrm{N}_{\varepsilon,n}(\theta)]$ and using the expression for the lower bound of $\mathbb{E}_\theta[\mathrm{N}_{\varepsilon,n}(\theta)]$ to get an upper bound on $\mathbb{E}_\theta[n - \mathrm{N}_{\varepsilon,n}(\theta)]$, we get

$$\leqslant 2\sqrt{2\varepsilon}n + 4\frac{c\sqrt{n}}{\varepsilon} \ . \tag{16}$$

Till now ε was a free parameter. We choose $\varepsilon = \frac{c}{0.01\sqrt{n}}$ such that, $\frac{c\sqrt{n}}{\varepsilon} = 0.01n$. This gives from Equation (16)

$$\lambda_{max}(H) \leqslant 2\sqrt{2}\sqrt{\frac{c}{0.01}}n^{\frac{3}{4}} + 4 * 0.01 * n \ . \tag{17}$$

Thus for a large but finite n, depending upon c and also on the geometry of the arm set, we have

$$\lambda_{max}(H) \leqslant 0.1n \ . \tag{18}$$

Hence, from Equation (10), we have

$$\lambda_i(\bar{G}_n) \leqslant 0.1n \ \forall i = 2, 3, \ldots, d \ . \tag{19}$$

Thus we have from Equation (18) the norm of the perturbation matrix; from equation (7) we have the eigen decompostion of $\bar{G}_n$; we also have the eigen-decomposition of $nvv^\top$ as $nvv^\top = nvv^\top + \sum_{i=2}^d 0\tilde{u}_i\tilde{u}_i^\top$ and from equation (19) we have the eigen-gap $\delta = \lambda_1(nvv^\top) - \lambda_2(\bar{G}_n) \geqslant n - 0.1n = 0.9n$.

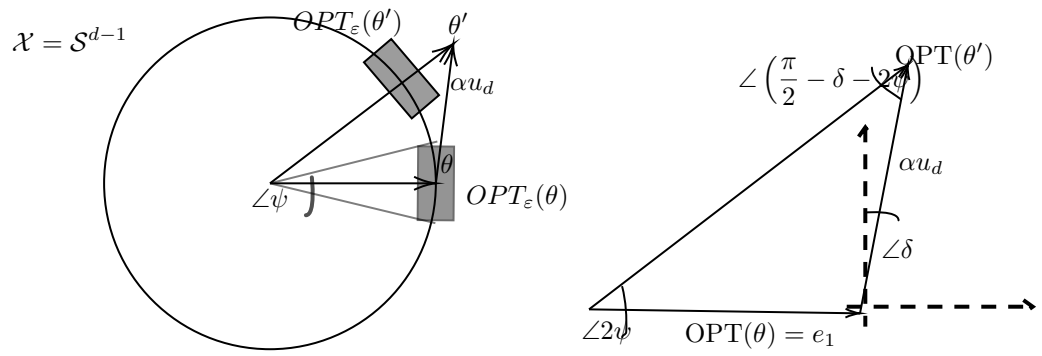
Therefore the Davis Kahan Theorem, (see Theorem 12.3) gives us

$$\|v^T \begin{bmatrix} u_2 & u_3 & \cdots & u_d \end{bmatrix}\| \leqslant \frac{\|H\|}{n - 0.1n} \leqslant \frac{0.1n}{0.9n} = \frac{1}{9} \ ,$$

which implies

$$\max_{i \in 2,\ldots,d} v^T u_i \leqslant \frac{1}{9},$$

which we interpret as the eigenvector corresponding to $\lambda_{min}(\bar{G}_n)$ is sufficiently orthogonal to $\mathrm{OPT}(\theta)$ when the optimal set decreases as roughly $\frac{1}{\sqrt{n}}$.



(a) The ε-nbd of optimal arms with respect to θ subtends an angle ψ in the centre. So in order for the ε-nbd of optimal arms about θ' to be disjoint, the angular displacement of θ' must be at least 2ψ

(b) αu_d makes an angle of $\pi/2 - \delta$ to θ

Figure 4: Proof of Lemma 7.2

The amount α to perturb θ to θ' In this penultimate section, we will try to find α such that the ε-optimal arms for θ and θ' are necessarily disjoint.

Formally we would want to find the smallest α s.t.

$$\mathrm{OPT}_\varepsilon(\theta) \cap \mathrm{OPT}_\varepsilon(\theta + \alpha u_d) = \emptyset$$

subject to the constraint $|\langle u_d, \mathrm{OPT}(\theta)\rangle| \leqslant \frac{1}{9}$

Lemma 7.2. *There exists a constant β such that $\mathrm{OPT}_\varepsilon(\theta) \cap \mathrm{OPT}_\varepsilon(\theta + \frac{\beta}{n^{1/4}} u_d) = \emptyset$ with ε defined as above with u_d is subject to the constraint $|\langle u_d, \mathrm{OPT}(\theta)\rangle| \leqslant \frac{1}{9}$.*

Proof. We will be focusing on 3 points θ, $\theta + \alpha u_d$ and the centre of the sphere. Since a plane can be drawn passing through any 3 non-collinear points we will be working in terms of the triangle formed by these three points. We will refer to Figure 4 for the proof.

Let us first note that for any given arbitrary ε, the $\mathrm{OPT}_\varepsilon(\theta)$ subtends an angle ψ about the centre. (See left image of Figure 4). In order to ensure that $\mathrm{OPT}_\varepsilon(\theta)$ and $\mathrm{OPT}_\varepsilon(\theta + \alpha u_d)$ are disjoint, we need to ensure that angular displacement of θ' is at least twice that is 2ψ from θ (see left image of Figure 4) and thus choose the step size α based on the angular displacement ψ.

Note that even in continuous action-sets with arbitrary geometry, if the geometry is smooth, the idea still remains the same. Thus we have,

$$\psi = \cos^{-1}(1 - \varepsilon) \tag{20}$$

(see the left image of Figure 4 for better clarity $\langle x, \theta\rangle = \cos\psi \geqslant 1 - \varepsilon$ for any $x \in \mathrm{OPT}_\varepsilon(\theta)$) and hence the displacement must be at least $2\cos^{-1}(1 - \varepsilon)$.

Now from the constraint condition that $v^\top u_d \leqslant 1/9$, we refer to the right hand side of figure 4, where δ is the complementary angle between by v (and hence θ) and u_d. Thus

$$\sin\delta = \cos(\pi/2 - \delta) = \langle v, u_d\rangle \leqslant 1/9 \tag{21}$$

(see the right image of Figure 4) and hence $\cos\delta \geqslant \sqrt{\frac{80}{81}}$. Therefore applying the law of sines we have (see the right image of Figure 4)

$$\frac{\alpha}{\sin 2\psi} = \frac{1}{\sin(\pi/2 - \delta - 2\psi)} \implies \alpha = \frac{\sin(2\psi)}{\sin(\pi/2 - \delta - 2\psi)} = \frac{\sin(2\psi)}{\cos(\delta + 2\psi)} \tag{22}$$

where α is the step-size to be determined.

Now note from left image of Figure 4 that we have (see Equation (20))

$$\cos\psi = 1-\varepsilon \implies \sin\psi = \sqrt{2\varepsilon-\varepsilon^2} \implies \cos^{-1}(1-\varepsilon) = \sin^{-1}(\sqrt{2\varepsilon-\varepsilon^2}) \approx \sqrt{2\varepsilon-\varepsilon^2} \approx \sqrt{2\varepsilon}$$

for small angles and therefore $\sin 2\psi \approx 2\psi = 2\cos^{-1}(1-\varepsilon) \approx 2\sqrt{2\varepsilon}$ and $\cos(\delta+2\psi) \approx \cos\delta \geqslant \sqrt{\frac{80}{81}}$ (see Equation (21)). Therefore from Equation (22) we have

$$\alpha \lessapprox \frac{2\sqrt{2\varepsilon}}{\sqrt{\frac{80}{81}}} = c_2\sqrt{\varepsilon} = \beta\frac{1}{n^{\frac{1}{4}}},$$

and hence the claim is proved. □

Finishing the proof Let us now define the random variable Z in Equation (6) as the fraction of times an arm in the ε-optimal set has been played, that is $Z \triangleq \frac{\mathrm{N}_{\varepsilon,n}(\theta)}{n}$. Then , we have

$$\mathbb{E}_\theta[Z] \geqslant \frac{n-\frac{c\sqrt{n}}{\varepsilon}}{n} = \frac{0.99n}{n} = 0.99,$$

where we have used the definition of Z, the expression for the lower bound on $\mathbb{E}_\theta[\mathrm{N}_{\varepsilon,n}(\theta)]$ and the value of ε. On the other hand, because of the disjoint ε-optimal sets of θ and θ' by construction, we have

$$\mathbb{E}_{\theta'}[Z] = \frac{\mathbb{E}_{\theta'}[\mathrm{N}_{\varepsilon,n}(\theta)]}{n} = \frac{n-\mathbb{E}_{\theta'}[\mathrm{N}_{\varepsilon,n}(\theta')]}{n} \leqslant \frac{\frac{c\sqrt{n}}{\varepsilon}}{n} = \frac{0.01n}{n} = 0.01$$

where we have used the regret property of the algorithm of being at most $c\sqrt{n}$ for any bandit parameter satisfying $\|\theta\| \leqslant 2$ by having n large enough such that $\|\theta'\| = \|\theta+\alpha u_d\| \leqslant 2$. Thus using Equations (6) and (8) and using the estimates of $\mathbb{E}_{\theta'}[Z]$ and $\mathbb{E}_\theta[Z]$ we have

$$\frac{\alpha^2}{2}\lambda_d \geqslant \mathrm{KL}(\mathrm{Ber}(\mathbb{E}_\theta[Z])||\mathrm{Ber}(\mathbb{E}_{\theta'}[Z])) \geqslant 2(\mathbb{E}_\theta[Z]-\mathbb{E}_{\theta'}[Z])^2 \geqslant 2(0.99-0.01)^2 = c_4$$

where for the second inequality we have used Pinsker's Inequality. Finally using Lemma 7.2 to find α which also guarantees our estimates of $\mathbb{E}_{\theta'}[Z]$ and $\mathbb{E}_\theta[Z]$, we have

$$\lambda_d(\bar{G}_n)) \geqslant 2c_4\frac{1}{\alpha^2} = c_5\frac{1}{(\frac{\beta}{n^{\frac{1}{4}}})^2} = c_6\sqrt{n}$$

This completes the proof with the positive constant c_6 being redefined as γ. □

Remark 7.3 (Dependence on dimension d). Though our result does not explicitly depend on the dimension d, it can depend on it through the regret upper bound of underlying linear bandit algorithm. Hypothetically, if there exists an algorithm which does not depend upon d, our result shows that the growth of minimum eigenvalue of design matrix is also independent of d. That being said, the dimensional dependency enters through the constant c and is different for different algorithms. For example, for the OFUL algorithm (Abbasi-yadkori et al., 2011), we have $c = O(d)$. Similarly, for Thompson Sampling algorithm (Abeille and Lazaric, 2017), we get $c = O(d^{3/2})$.

We can show that $n_0 = \Omega(c^2)$ and $\gamma = \Omega(\frac{1}{c})$. Let us compute this for the spherical action set. From Equation (17), we have

$$\lambda_{\max}(H) \leqslant 2\sqrt{2}\sqrt{\frac{c}{0.01}}n^{\frac{3}{4}} + 4\times 0.01\times n.$$

Now, for a large but finite n, we have the right hand side to be less than $0.1n$. This finite n is the n_0 as defined in Thm 2.3. An easy calculation shows that $n_0 = \Omega(c^2)$. Similarly, γ can be computed as follows. Note that $\lambda_d \geqslant \frac{3.8416}{\alpha^2} = \frac{3.8416}{\beta^2}\sqrt{n}$, where $\beta = \frac{2\sqrt{\frac{2c}{0.01}}}{\sqrt{\frac{80}{81}}}$ and $\gamma = \frac{3.8416}{\beta^2}$. This implies $\gamma = \Omega(\frac{1}{c})$.

7.2 Ellipsoidal Action-Set

Theorem 7.4. *Let the set of arms be $\mathcal{X} =:= \left\{x \in \mathbb{R}^d : x^\top A^{-1} x = 1\right\}$, the surface of the d dimensional ellipsoid (A is symmetric and positive definite). Let $\bar{G}_n = \mathbb{E}_\theta \left[\sum_{s=1}^n A_s A_s^\top\right]$, where θ is a bandit parameter and A_s are arms in $\mathcal{X}$ drawn according to some bandit algorithm.*

For any bandit algorithm which suffers expected regret, $R_n(\theta)$, at most $O(\sqrt{n})$,

$$\lambda_{min}(\bar{G}_n) = \Omega(\sqrt{n}) \ .$$

That is there exists constants $\gamma > 0$ and a finite time n_0 such that for all $n \geqslant n_0$, we have $\lambda_{min}(\bar{G}_n) \geqslant \gamma\sqrt{n}$.

Remark 7.5. The constants n_0 and γ depend upon the algorithm constants hidden by $O()$, the condition number of A and the size of the bandit parameter θ.

Proof. Let $\mathcal{X} = \{x \in \mathbb{R}^d : x^\top A^{-1} x = 1\}$ be an ellipsoidal action set where A is symmetric positive definite and θ be the bandit parameter.

Let us make a change of variable $y = A^{-1/2}x$ for all $x \in \mathcal{X}$. Thus we have $\|y\|^2 = 1$. We define this new transformed action space $\mathcal{Y}$ and note that $\mathcal{Y}$ is the unit sphere.

Before we proceed further we make the following observations :

$$\mathrm{OPT}_{\mathcal{X}}(\theta) = \arg\max_{x \in \mathcal{X}} \langle x, \theta\rangle = \arg\max_{x \in \mathcal{X}} \langle A^{-1/2}x, A^{1/2}\theta\rangle \tag{23}$$

$$= A^{1/2} \arg\max_{y \in A^{-1/2}\mathcal{X}} \langle y, A^{1/2}\theta\rangle = A^{1/2} \arg\max_{y \in \mathcal{Y}} \langle y, A^{1/2}\theta\rangle = A^{1/2}\mathrm{OPT}_{\mathcal{Y}}(A^{1/2}\theta).$$

Similarly we have

$$\mathrm{OPT}_{\varepsilon,\mathcal{X}}(\theta) = A^{1/2}\mathrm{OPT}_{\varepsilon,\mathcal{Y}}(A^{1/2}\theta) \tag{24}$$

Thus let us consider the following linear bandit problem where the action set is $\mathcal{Y} = \mathcal{S}^{d-1}$, and the bandit parameter is $A^{1/2}\theta$. Let the actions sampled be $\{y_i\}_{i=1}^n$ by any bandit algorithm with regret at most $O(\sqrt{n})$. Let the design matrix be $V_n = \mathbb{E}\left[\sum_{i=1}^n y_i y_i^\top\right]$, and the corresponding eigenvector decomposition be $\sum_{i=1}^d \lambda_i u_i u_i^\top$.

From the previous section (see Lemma 7.2) we know,

$$\mathrm{OPT}_{\varepsilon,\mathcal{Y}}\left(\frac{A^{1/2}\theta}{\left\|A^{1/2}\theta\right\|}\right) \cap \mathrm{OPT}_{\varepsilon,\mathcal{Y}}\left(\frac{A^{1/2}\theta}{\left\|A^{1/2}\theta\right\|} + \alpha u_d\right) = \emptyset \ , \tag{25}$$

for $\alpha = O(1/n^{1/4})$. Thus multiplying the two disconnected sets by $\left\|A^{1/2}\theta\right\|$ is still going to keep them disconnected and hence,

$$\mathrm{OPT}_{\varepsilon,\mathcal{Y}}(A^{1/2}\theta) \cap \mathrm{OPT}_{\varepsilon,\mathcal{Y}}\left(A^{1/2}\theta + \alpha\left\|A^{1/2}\theta\right\| u_d\right) = \emptyset. \tag{26}$$

This implies,

$$\mathrm{OPT}_{\varepsilon,\mathcal{Y}}(A^{1/2}\theta) \cap \mathrm{OPT}_{\varepsilon,\mathcal{Y}}\left(A^{1/2}(\theta + \alpha\left\|A^{1/2}\theta\right\| A^{-1/2}u_d)\right) = \emptyset \tag{27}$$

and thus from the change of variable relation,(See equation (24)), we have

$$A^{-1/2}\mathrm{OPT}_{\varepsilon,\mathcal{X}}(\theta) \cap A^{-1/2}\mathrm{OPT}_{\varepsilon,\mathcal{X}}(\theta + \alpha\left\|A^{1/2}\theta\right\| A^{-1/2}u_d) = \emptyset. \tag{28}$$

Now because continuous functions preserve disjoint sets, we have

$$\mathrm{OPT}_{\varepsilon,\mathcal{X}}(\theta) \cap \mathrm{OPT}_{\varepsilon,\mathcal{X}}(\theta + \alpha\left\|A^{1/2}\theta\right\| A^{-1/2}u_d) = \emptyset. \tag{29}$$

Now let us define the design matrix $\bar{G}_n$ for the action set $\mathcal{X}$ as $\bar{G}_n = \mathbb{E}\left[\sum_{i=1}^n x_i x_i^\top\right]$. With the change of variables we have

$$\bar{G}_n = \mathbb{E}\left[\sum_{i=1}^n x_i x_i^\top\right] = A^{1/2}\mathbb{E}\left[\sum_{i=1}^n y_i y_i^\top\right] A^{1/2} = A^{1/2}\sum_{i=1}^d \lambda_i u_i u_i^\top A^{1/2}\,, \tag{30}$$

where u_i are the eigen-vectors as defined for the action space $\mathcal{Y}$. Now defining the perturbed $\theta' = \theta + \alpha \left\|A^{1/2}\theta\right\| A^{-1/2}u_d$, we have

$$\|\theta - \theta'\|^2_{\bar{G}_n} = \alpha^2 \left\|A^{1/2}\theta\right\|^2 u_d^\top A^{-1/2}\Big(A^{1/2}\sum_{i=1}^d \lambda_i u_i u_i^T A^{1/2}\Big)A^{-1/2}u_d = \alpha^2 \left\|A^{1/2}\theta\right\|^2 \lambda_d \tag{31}$$

Now, from the methodology discussed in the previous section using the Kauffman measure change inequality 12.1(see the paragraph 7.1), we have

$$\lambda_d \geqslant \frac{c}{\alpha^2\left\|A^{1/2}\theta\right\|^2} = \frac{c\sqrt{n}}{\left\|A^{1/2}\theta\right\|^2}\,, \tag{32}$$

for some positive c. To conclude the proof we need to find a relation between the eigenvalues lowest eigenvalue μ_d of $\bar{G}_n$ and the lowest eigenvalue λ_d of V_n. For this note, from equation (30)

$$\bar{G}_n = A^{1/2}V_n A^{1/2} = A^{1/2}U\Lambda U^\top A^{1/2}. \tag{33}$$

From spectral theory we have

$$\mu_d(\bar{G}_n) = \frac{1}{\left\|\bar{G}_n^{-1}\right\|}. \tag{34}$$

Now from equation (33), and definition of matrix norms and orthogonality of eigenctors U, we have

$$\left\|G_n^{-1}\right\| \leqslant \left\|A^{-1/2}\right\|^2 \left\|\Lambda^{-1}\right\|, \tag{35}$$

and therefore,

$$\mu_d(\bar{G}_n) \geqslant \frac{1}{\left\|A^{-1/2}\right\|^2\left\|\Lambda^{-1}\right\|} = \frac{\lambda_d}{\left\|A^{-1/2}\right\|^2} \geqslant \frac{c\sqrt{n}}{\left\|A^{-1/2}\right\|^2\left\|A^{1/2}\theta\right\|^2} \geqslant \frac{c\sqrt{n}}{\left\|A^{-1/2}\right\|^2\left\|A^{1/2}\right\|^2\|\theta\|^2}. \tag{36}$$

This concludes the proof. □

In the next subsection we illustrate that the same proof for ellipsoids can be done through first principles.

7.2.1 Ellipsoidal Sets : Proof by first principles

In this section we highlight the main ideas of the proof and leave the gaps as an exercise. The proof follows verbatim from the proof for the spherical case (see subsection 7.1).

Remark 7.6. We present this alternative proof to highlight the main takeaway of the proof technique, namely to construct an alternative bandit parameter θ', for which the actions played by the algorithm would significantly differ from the original bandit parameter and yet for both parameters, the algorithm plays optimally.

Remark 7.7. The idea of this proof is also to highlight that what really is needed is information about neighbourhood of the optimal arm, and the global action space does not influence the exploration strategy of good regret algorithms. This would be useful for the proof of the LCH surfaces as given in the next Section 7.3.

Proof. We fix a θ and start with the Garivier-Kauffman inequality

$$\frac{1}{2}\|\theta - \theta'\|^2_{\bar{G}_n} \geqslant \mathrm{KL}(\mathrm{Ber}(\mathbb{E}_\theta[Z])||\mathrm{Ber}(\mathbb{E}_{\theta'}[Z]))\,. \tag{37}$$

and do an eigen-decomposition of $\bar{G}_n$ as

$$\bar{G}_n = \sum_{i=1}^{d} \lambda_i u_i u_i^\top \ , \tag{38}$$

where $\lambda_1, \cdots, \lambda_d$ and $u_1, \cdots, u_d$ have the usual meaning. The crux of the proof remains that we design the perturbed θ' as before, namely as $\theta' = \theta + \alpha u_d$ for some α to be computed such that $\mathrm{OPT}_\varepsilon(\theta)$ and $\mathrm{OPT}_\varepsilon(\theta')$ are disjoint. With this choice of θ' we have from the measure change inequality as,

$$\|\theta - \theta'\|^2_{\bar{G}_n} = \alpha^2 \lambda_d \ . \tag{39}$$

As before, we have from the definitions of $\mathrm{OPT}_\varepsilon(\theta)$ and $\mathrm{N}_{\varepsilon,n}(\theta)$,

$$\mathbb{E}_{\theta'}[\mathrm{N}_{\varepsilon,n}(\theta')] \geqslant n - \frac{c\sqrt{n}}{\varepsilon} \ , \tag{40}$$

for all θ' satisfying $\|\theta'\| \leqslant 2\|\theta\|$ such that $\mathbb{E}[R_n(\theta')] \leqslant c\sqrt{n}$, for some $c > 0$. For ellipse we have the $\mathrm{OPT}(\theta) = \frac{A\theta}{\|\theta\|_A}$, which as before we denote as v.

Showing that v and u_d are roughly orthogonal to each other As before we decompose $\bar{G}_n$ as

$$\bar{G}_n = nvv^\top + \bar{G}_n - nvv^\top \tag{41}$$

and note from Weyl's Lemma and definition of λ_i, that

$$\lambda_i(\bar{G}_n) \leqslant \left\|\bar{G}_n - nvv^\top\right\| , \tag{42}$$

for $i \in \{2, 3, \cdots, d\}$. Now decomposing $\left\|\bar{G}_n - nvv^\top\right\|$ into the two sets, one containing arms belonging to the $\mathrm{OPT}_\varepsilon(\theta)$ set and another where arms do not belong to $\mathrm{OPT}_\varepsilon(\theta)$, and further using a generic upper bound of $\sqrt{\lambda_{\max}(A)}$ for the size of all arms in $\mathcal{X}$, we have,

$$\left\|\bar{G}_n - nvv^\top\right\| \leqslant \sum_{s:A_s \in \mathrm{OPT}_\varepsilon(\theta)} \sqrt{\lambda_{\max}(A)}\mathbb{E}\left[\|A_s - v\|\right] + \sum_{s:A_s \notin \mathrm{OPT}_\varepsilon(\theta)} \sqrt{\lambda_{\max}(A)}\mathbb{E}\left[\|(A_s - v)\|\right] \tag{43}$$

(See previous section of the spherical set section 7.1 for details of this decomposition). For $A_s \in \mathrm{OPT}_\varepsilon(\theta)$, we need to find an upper bound for $\|A_s - v\|$. Now recall the change of variable formula (Equations 23 and 24) from the last section, namely

$$\mathrm{OPT}_{\mathcal{X}}(\theta) = A^{1/2}\mathrm{OPT}_{\mathcal{Y}}(A^{1/2}\theta)$$

and

$$\mathrm{OPT}_{\varepsilon,\mathcal{X}}(\theta) = A^{1/2}\mathrm{OPT}_{\varepsilon,\mathcal{Y}}(A^{1/2}\theta) \ ,$$

where $\mathcal{X}$ and $\mathcal{Y}$ are the ellipse and unit sphere respectively.

Thus for $A_s \in \mathrm{OPT}_{\varepsilon,\mathcal{X}}(\theta)$, there exists a $x \in \mathrm{OPT}_{\varepsilon,\mathcal{Y}}(A^{1/2}\theta)$, such that $A_s = A^{1/2}x$. Therefore, for any $A_s \in \mathrm{OPT}_{\varepsilon,\mathcal{X}}(\theta)$

$$\|A_s - v\| = \left\|A^{1/2}x - A^{1/2}\mathrm{OPT}_{\mathcal{Y}}(A^{1/2}\theta)\right\| \tag{44}$$

for some $x \in \mathrm{OPT}_{\varepsilon,\mathcal{Y}}(A^{1/2}\theta)$. Now from the section on the sphere it follows, that for any $x \in \mathrm{OPT}_{\varepsilon,\mathcal{Y}}(A^{1/2}\theta)$, we have

$$\left\|x - \mathrm{OPT}_{\mathcal{Y}}(A^{1/2}\theta)\right\| \leqslant \sqrt{\frac{2\varepsilon}{\left\|A^{1/2}\theta\right\|}} \tag{45}$$

(see the section on the sphere for the full detail 7.1). Therefore for any $A_s \in \mathrm{OPT}_{\varepsilon,\mathcal{X}}(\theta)$ we have

$$\|A_s - v\| \leqslant \left\|A^{1/2}\right\| \sqrt{\frac{2\varepsilon}{\left\|A^{1/2}\theta\right\|}} \ .$$

Thus we have

$$\left\|\bar{G}_n - nvv^\top\right\| \leqslant \sqrt{\lambda_{\max}(A)}\left\|A^{1/2}\right\|\sqrt{\frac{2\varepsilon}{\left\|A^{1/2}\theta\right\|}}\mathbb{E}_\theta[\mathrm{N}_{\varepsilon,n}(\theta)] + \lambda_{\max}(A)\mathbb{E}_\theta[n - \mathrm{N}_{\varepsilon,n}(\theta)] \ .$$

Now using the estimates of $\mathbb{E}_\theta[\mathrm{N}_{\varepsilon,n}(\theta)]$ (Equation 40) we have

$$\left\|\bar{G}_n - nvv^\top\right\| \leqslant \sqrt{\lambda_{\max}(A)}\left\|A^{1/2}\right\|\sqrt{\frac{2\varepsilon}{\left\|A^{1/2}\theta\right\|}}n + \lambda_{\max}(A)\frac{c\sqrt{n}}{\varepsilon} \ . \tag{46}$$

The rest of the proof remains the same as in the spherical case (see subsection 7.1), but now n_0 would depend upon the matrix A as well.

To show that the sets $\mathrm{OPT}_\varepsilon(\theta)$ and $\mathrm{OPT}_\varepsilon(\theta + O(1/n^{1/4})u_d)$ are disjoint. Let us first rewrite what we need to show. Namely we want to find an α such that

$$\mathrm{OPT}_{\varepsilon,\mathcal{X}}(\theta) \cap \mathrm{OPT}_{\varepsilon,\mathcal{X}}(\theta + \alpha u_d) = \emptyset \ . \tag{47}$$

Now using the change of variables (Equations (23) and (24)) this is equivalent to showing

$$A^{1/2}\mathrm{OPT}_{\varepsilon,\mathcal{Y}}(A^{1/2}\theta) \cap A^{1/2}\mathrm{OPT}_{\varepsilon,\mathcal{Y}}(A^{1/2}(\theta + \alpha u_d)) = \emptyset \ , \tag{48}$$

where $\mathcal{Y}$ is the sphere. By the bijection of $A^{1/2}$, the above is equivalent to showing

$$\mathrm{OPT}_{\varepsilon,\mathcal{Y}}(A^{1/2}\theta) \cap \mathrm{OPT}_{\varepsilon,\mathcal{Y}}(A^{1/2}\theta + \alpha A^{1/2}u_d) = \emptyset \ . \tag{49}$$

Now we know $\langle v, u_d\rangle \approx 0$ where v is the optimal arm for the ellipsoid (see the previous paragraph). But for the ellipsoid we know $v = \frac{A\theta}{\|\theta\|_A}$. Thus $\langle \frac{A\theta}{\|\theta\|_A}, u_d\rangle \approx 0$. This implies $\langle A^{1/2}\theta, A^{1/2}u_d\rangle \approx 0$. Thus now we can use the result for the spherical action set (Lemma 7.2), namely,

$$\mathrm{OPT}_{\varepsilon,\mathcal{Y}}\Big(\frac{A^{1/2}\theta}{\left\|A^{1/2}\theta\right\|}\Big) \cap \mathrm{OPT}_{\varepsilon,\mathcal{Y}}\Big(\frac{A^{1/2}\theta}{\left\|A^{1/2}\theta\right\|} + \alpha'\frac{A^{1/2}u_d}{\left\|A^{1/2}u_d\right\|}\Big) = \emptyset \tag{50}$$

where $\alpha' = O(\frac{1}{n^{1/4}})$. Thus multiplying by $\left\|A^{1/2}\theta\right\|$, we have

$$\mathrm{OPT}_{\varepsilon,\mathcal{Y}}\Big(A^{1/2}\theta\Big) \cap \mathrm{OPT}_{\varepsilon,\mathcal{Y}}\Big(A^{1/2}\theta + \alpha'\frac{\left\|A^{1/2}\theta\right\|}{\left\|A^{1/2}u_d\right\|}A^{1/2}u_d\Big) = \emptyset \ . \tag{51}$$

Thus the required $\alpha = \alpha'\frac{\left\|A^{1/2}\theta\right\|}{\left\|A^{1/2}u_d\right\|}$.

Finishing the proof The rest of the proof follows as the ones shown in the previous two sections (paragraph 7.1). By ensuring that for a large enough n, such that $\|\theta'\| \leqslant 2\|\theta\|$, and using the implication of the dijointedness of the ε-optimal sets for θ and θ' in the Garivier-Kauffman inequality, we get

$$\lambda_d \geqslant \frac{c_1}{\alpha^2} = \frac{c_1\sqrt{n}\left\|A^{1/2}u_d\right\|^2}{\left\|A^{1/2}\theta\right\|^2} \geqslant \frac{c_1\sqrt{n}\lambda_{\min}(A)}{\left\|A^{1/2}\theta\right\|^2} = \frac{c_1\sqrt{n}}{\left\|A^{-1/2}\right\|^2\left\|A^{1/2}\theta\right\|^2} \ , \tag{52}$$

for some positive constant c_1. This completes the proof. □

Non centred ellipsoids. Even though we are considering centred ellipsoids, our proofs readily extend to non-centred ellipsoids. Consider $\mathcal{X} = \{x : (x-a)^\top M^{-1}(x-a)\}$ a non-centred ellipsoid. Let $y = x - a$, Then for all such x in $\mathcal{X}$, y satisfies the equation of the following centred ellipsoid $\mathcal{Y} = \{y : y^\top M^{-1}y = 1\}$. Now with the usual definitions of $\mathrm{OPT}_{\mathcal{X}}(\theta)$, $\mathrm{OPT}_{\mathcal{Y}}(\theta)$, $\mathrm{OPT}_{\varepsilon,\mathcal{X}}(\theta)$ and $\mathrm{OPT}_{\varepsilon,\mathcal{Y}}(\theta)$, observe that the following relations hold true

$$\mathrm{OPT}_{\mathcal{X}}(\theta) = \mathrm{OPT}_{\mathcal{Y}}(\theta) + a$$
$$\mathrm{OPT}_{\varepsilon,\mathcal{X}}(\theta) = \mathrm{OPT}_{\varepsilon,\mathcal{Y}}(\theta) + a \ .$$

With these relations and using arguments similar to this section (subsection 7.2), we have our result readily.

7.3 Locally Constant Hessian (LCH) Action Spaces

In this section we prove Theorem 2.2. We begin by recalling our definition of LCH surfaces (Definition 2.1) and Theorem 2.2:

Definition 7.8 (Locally Constant Hessian (LCH) surface)**.** Consider the action space defined by $\mathcal{X} = \{x \in \mathbb{R}^d : f(x) = c\}$, where $f : \mathbb{R}^d \to \mathbb{R}$ is a C^2 function (i.e., all second-order partial derivatives of f exist and are continuous) and $c \in \mathbb{R}$. Let $\theta \in \mathbb{R}^d$. $\mathcal{X}$ is said to be a LCH surface w.r.t. θ if: (i) there is a unique reward-optimal arm with respect to θ (denoted by $\mathrm{OPT}_{\mathcal{X}}(\theta) = \arg\max_{x\in\mathcal{X}}\langle x, \theta\rangle$), and (ii) there is an open neighborhood $U \subset \mathbb{R}^d$ of $\mathrm{OPT}_{\mathcal{X}}(\theta)$ over which the Hessian of f is constant and positive-definite.

We are now ready to prove our main result.

Theorem 7.9. *Let the action-space $\mathcal{X}$ be a Locally Constant Hessian(LCH) surface in $\mathbb{R}^{d-1}$ w.r.t. a bandit parameter θ. Let $\bar{G}_n = \mathbb{E}_\theta\left[\sum_{s=1}^n A_s A_s^\top\right]$, where A_s are arms in $\mathcal{X}$ drawn according to some bandit algorithm. For any bandit algorithm which suffers expected regret at most $O(\sqrt{n})$,*

$$\lambda_{\min}(\bar{G}_n) = \Omega(\sqrt{n})\ .$$

That is, there exists an n_0 and a constant $\gamma > 0$, such that for all $n \geqslant n_0$, $\lambda_{\min}(\bar{G}_n) \geqslant \gamma\sqrt{n}$.

Remark 7.10. The constant γ depends upon the condition number of the Hessian, the algorithmic constants hidden by $O()$, and the size of the bandit parameter θ. The constant n_0 depends on the algorithmic constants hidden by $O()$, the size of the neighbourbood over which the Hessian is constant, the size of the action domain $\|\mathcal{X}\|$, the singular value of the Hessian and the size of the bandit parameter θ.

Remark 7.11. The proof essentially follows the proof of Section 7.2.1. We present it here for the sake of completeness.

Proof. Consider, the action space given by the curve $\mathcal{X} = \{x \in \mathbb{R}^d : f(x) = 1\}$. Assume that $\mathcal{X}$ is upper bounded in norm by a constant represented as $\lambda(\mathcal{X})$. Let us define $\mathrm{OPT}_{\mathcal{X}}(\theta)$ and $\mathrm{OPT}_{\varepsilon,\mathcal{X}}(\theta)$ as before for a bandit parameter θ. Let U be an open-neighbourhood of $\mathrm{OPT}_{\mathcal{X}}(\theta)$, over which the Hessian of f, $\nabla^2 f$, is a constant H, and is positive-definite.

We shall show that in this setup the conclusion of $\lambda_{\min}(\bar{G}_n)$ holds true.

As before, we define $\bar{G}_n = \mathbb{E}\left[\sum_{s=1}^n A_s A_s^\top\right]$ where $A_s \in \mathcal{X}$ and we have a eigenvalue decomposition of $\sum_{i=1}^d \lambda_i u_i u_i^\top$.

We make a Taylor Series approximation of f about $\mathrm{OPT}_{\mathcal{X}}(\theta)$ (denoted here as x^* for notational ease) in the neighbourhood of U.

$$f(x) = f(x^*) + (x - x^*)^\top \nabla f(x^*) + (x - x^*)^\top \nabla^2 f(x^*)(x - x^*). \tag{53}$$

(Note that we have assumed the Hessian is constant in U, so there are no third order terms.) Simplifying the above the expression, we get the following quadratic term,

$$(x - x^*)^\top \nabla^2 f(x^*)(x - x^*) + (x - x^*)^\top \nabla f(x^*) = 0\ . \tag{54}$$

Completing the squares we get the following equation for the ellipsoid,

$$(x - a)^\top M^{-1}(x - a) = 1 \tag{55}$$

where $M^{-1} = \frac{H}{4b^\top H^{-1} b}$, $a = x^* - 1/2H^{-1}b$ and $b = \nabla f(x^*)$, for all $x \in U$.

Thus for all $x \in U \subset \mathcal{X}$, x satisfies the equation of the ellipse $\mathcal{Y} = \{x : (x - v)^\top M^{-1}(x - v) = 1\}$. The following claims are immediately clear $\mathrm{OPT}_{\mathcal{X}}(\theta) = \mathrm{OPT}_{\mathcal{Y}}(\theta)$ and for small ε such that $\mathrm{OPT}_{\varepsilon,\mathcal{X}}(\theta) \subset U$, we have $\mathrm{OPT}_{\varepsilon,\mathcal{X}}(\theta) = \mathrm{OPT}_{\varepsilon,\mathcal{Y}}(\theta)$.

Let us then choose an ε small enough so that $\mathrm{OPT}_{\varepsilon,\mathcal{X}}(\theta) = \mathrm{OPT}_{\varepsilon,\mathcal{Y}}(\theta)$.

As a first step we see that the estimates of $\mathbb{E}_{\theta'}[\mathrm{N}_{\varepsilon,n}(\theta')] \geqslant n - \frac{c\sqrt{n}}{\varepsilon}$ remain true, for the entire action space $\mathcal{X}$, for all θ' such that $\|\theta'\| \leqslant 2\|\theta\|$ for which $\mathbb{E}[R_n(\theta')] \leqslant c\sqrt{n}$ for some positive constant c.

As the next order of business we need to show $\mathrm{OPT}_{\mathcal{X}}(\theta)$ is approximately orthogonal to u_d. For this first we note that $\mathrm{OPT}_{\mathcal{X}}(\theta) = \mathrm{OPT}_{\mathcal{Y}}(\theta)$. Thus it suffices to show that $\mathrm{OPT}_{\mathcal{Y}}(\theta)$ is approximately orthogonal to u_d.

As before denote $\mathrm{OPT}_{\mathcal{Y}}(\theta) = v$ and decompose $\bar{G}_n$ as

$$\bar{G}_n = nvv^\top + \bar{G}_n - nvv^\top \tag{56}$$

and note from Weyl's Lemma and definition of λ_i, that

$$\lambda_i(\bar{G}_n) \leqslant \left\|\bar{G}_n - nvv^\top\right\|, \tag{57}$$

for $i \in \{2, 3, \cdots, d\}$. Now decomposing $\left\|\bar{G}_n - nvv^\top\right\|$ into the two sets, one containing arms belonging to the $\mathrm{OPT}_{\varepsilon,\mathcal{X}}(\theta)$ set and another where arms do not belong to the $\mathrm{OPT}_{\varepsilon,\mathcal{X}}(\theta)$ set and further using the generic upper bound of $\lambda_{\max}(\mathcal{X})$ for the size of all arms we have,

$$\left\|\bar{G}_n - nvv^\top\right\| \leqslant \sum_{s:A_s \in \mathrm{OPT}_\varepsilon(\theta)} \lambda(\mathcal{X})\mathbb{E}\left[\|A_s - v\|\right] + \sum_{s:A_s \notin \mathrm{OPT}_\varepsilon(\theta)} \lambda(\mathcal{X})\mathbb{E}\left[\|(A_s - v)\|\right] . \tag{58}$$

Note as before, this can be be upper bounded using the estimates of $\mathbb{E}_\theta[\mathrm{N}_{\varepsilon,n}(\theta)]$ as

$$\left\|\bar{G}_n - nvv^\top\right\| \leqslant \frac{c\sqrt{n}\lambda(\mathcal{X})^2}{\varepsilon} + n\lambda(\mathcal{X}) \sup_{A_s \in \mathrm{OPT}_{\varepsilon,\mathcal{X}}(\theta)} \|A_s - v\| . \tag{59}$$

Now for small ε such that $\mathrm{OPT}_{\varepsilon,\mathcal{X}}(\theta) = \mathrm{OPT}_{\varepsilon,\mathcal{Y}}(\theta)$ we have $\sup_{A_s \in \mathrm{OPT}_{\varepsilon,\mathcal{X}}(\theta)} \|A_s - v\| = \sup_{A_s \in \mathrm{OPT}_{\varepsilon,\mathcal{Y}}(\theta)} \|A_s - v\|$ for which we know $\|A_s - v\| \leqslant O(\sqrt{\varepsilon})$. Thus by choosing ε as $O(1/\sqrt{n})$ for sufficiently large n, such that $\mathrm{OPT}_{\varepsilon,\mathcal{X}}(\theta) = \mathrm{OPT}_{\varepsilon,\mathcal{Y}}(\theta)$, we have (by choosing appropriate constant for reference see the section on the sphere)

$$\left\|\bar{G}_n - nvv^\top\right\| \leqslant 0.1n . \tag{60}$$

Thus we have the separation of eigen-values between $\lambda_1(nvv^\top)$ and $\lambda_2(\bar{G}_n)$ as at least $0.9n$. This gives us from the Davis-Kahan Theorem

$$|v^\top u_d| \leqslant 1/9 \tag{61}$$

(See the previous sections for details). Thus we have $\mathrm{OPT}_{\mathcal{X}}(\theta)$ is approximately orthogonal to u_d.

As the next order of business we need to find a perturbed version of θ, namely $\theta' = \theta + \alpha u_d$ such that $\mathrm{OPT}_{\varepsilon,\mathcal{X}}(\theta)$ and $\mathrm{OPT}_{\varepsilon,\mathcal{X}}(\theta')$ are disjoint. First let us choose ε small enough by sufficiently large n such $\mathrm{OPT}_{\varepsilon,\mathcal{X}}(\theta) = \mathrm{OPT}_{\varepsilon,\mathcal{Y}}(\theta)$. Now as before for the ellipsoid case we have $\mathrm{OPT}_{\varepsilon,\mathcal{Y}}(\theta) \cap \mathrm{OPT}_{\varepsilon,\mathcal{Y}}(\theta + O(1/n^{1/4})u_d) = \emptyset$. Now by the continuity of $\mathrm{OPT}_{\mathcal{Y}}(\theta) = \frac{M\theta}{\|\theta\|_M} + a$, we have for sufficiently large n, $\mathrm{OPT}_{\mathcal{Y}}(\theta + O(1/n^{1/2})u_d) \in U$ and hence for small ε we have $\mathrm{OPT}_{\varepsilon,\mathcal{Y}}(\theta + O(1/n^{1/4})u_d) \subset U$. Thus for large enough n, we have two disjoint sets $\mathrm{OPT}_{\varepsilon,\mathcal{Y}}(\theta)$ and $\mathrm{OPT}_{\varepsilon,\mathcal{Y}}(\theta + O(1/n^{1/4})u_d)$ both contained in U and thus $\mathrm{OPT}_{\varepsilon,\mathcal{X}}(\theta) \cap \mathrm{OPT}_{\varepsilon,\mathcal{X}}(\theta + O(1/n^{1/4})u_d) = \emptyset$. As before, we can now use the Gariver-Kauffman inequality to get

$$\lambda_d(\bar{G}_n) \geqslant \gamma\sqrt{n} \tag{62}$$

for large enough n depending upon U, and some constant γ depending upon the Hessian. This completes our proof. □

8 LOCALLY CONVEX SURFACES (PROOF OF THEOREM 3.3)

In this section we provide a proof of Theorem 3.3. We begin again by by recalling or definition of Locally Convex surfaces (Definition 3.1) and Theorem 3.3 for Locally Convex Action Sets.

Definition 8.1 (Locally Convex surface)**.** Consider an action space $\mathcal{X} = \{x \in \mathbb{R}^d : f(x) = c\}$, where $f : \mathbb{R}^d \to \mathbb{R}$ is a C^2 function (i.e., all second order partial derivatives exist and are continuous). With $\mathrm{OPT}_{\mathcal{X}}(\theta)$ being the optimal arm defined as before, let the Hessian of f at $\mathrm{OPT}_{\mathcal{X}}(\theta)$, denoted as $\nabla^2 f(\mathrm{OPT}_{\mathcal{X}}(\theta))$, be positive definite. Then, $\mathcal{X}$ is said to be a Locally Convex surface.

Theorem 8.2. *Let $\mathcal{X}$ be a Locally Convex action space and $\bar{G}_n$, be the expected design matrix. For any bandit algorithm which suffers expected regret at most $O(\sqrt{n})$, there exists a real number s in the half-open interval $(0, \frac{1}{2}]$ such that*

$$\lambda_{\min}(\bar{G}_n) = \Omega(n^s) .$$

Remark 8.3. The exponent s defined in Theorem 8.2 depends explicitly on the geometry of the action-space. In general it depends on how well the surface approximates a LCH surface. The best we can say when no other information is given about the action space is that the eigenvalue grows polynomially, which still results in a polynomial rate of the parameter estimation.

Proof of Theorem 8.2. The proof utilizes the following lemmas,

Lemma 8.4. *We can construct a quadratic approximation $\mathcal{E}$ about $\mathrm{OPT}_{\mathcal{X}}(\theta)$ similar to the construction for the LCH case.*

Lemma 8.5. *There exists a one-to-one map, $\bar{\mathbf{P}}_{\mathcal{X},\mathcal{E}} : \mathcal{X} \to \mathcal{E}$ and a positive real number p in the interval $(0, 1]$, such that the ε-optimal set in $\mathcal{X}$, $\mathrm{OPT}_{\varepsilon,\mathcal{X}}(\theta)$ is contained in an $O(\varepsilon^p)$-optimal set in $\mathcal{E}$ for small enough ε .*

Lemma 8.6. *There exists a positive real number s in the interval $(0, \frac{1}{2}]$, such that for $\theta' = \theta + O(\sqrt{\frac{1}{n^s}})u_d$, the $O(\varepsilon^{2s})$-optimal sets in the approximation $\mathcal{E}$, $\mathrm{OPT}_{O(\varepsilon^{2s}),\mathcal{E}}(\theta)$ and $\mathrm{OPT}_{O(\varepsilon^{2s}),\mathcal{E}}(\theta')$ are disjoint, and the image of the ε-optimal set $\mathrm{OPT}_{\varepsilon,\mathcal{X}}(\theta')$ under the operator defined in Lemma 8.5 is contained in $\mathrm{OPT}_{O(\varepsilon^{2s}),\mathcal{E}}(\theta')$.*

Thus using Lemma 8.5 and Lemma 8.6, we ensure that the ε-optimal sets of $\mathcal{X}$ for the bandit parameters θ and θ' are disjoint. The remainder of the proof is the same as in the earlier cases. □

Proof of Lemma 8.4. By the C^2 definition, there exists an open set $W \in \mathbb{R}^d$ such that the Hessian of f over $W \cap \mathcal{X} \triangleq U$, $\nabla^2(f)(U)$ exists and is positive definite. Let $\mathcal{E}$ be a quadratic approximation ellipsoid (similar to the construction as in the locally constant hessian (LCH) case) about $\mathrm{OPT}_{\mathcal{X}}(\theta)$. □

Proof of Lemma 8.5. Let $\bar{\mathbf{P}}_{\mathcal{X},\mathcal{E}} : \mathcal{X} \to \mathcal{E}$ be the best approximation operator denoted by

$$\bar{\mathbf{P}}_{\mathcal{X},\mathcal{E}}(y) = \operatorname*{argmin}_{x \in \mathcal{E}} \|x - y\|$$

for any y in $\mathcal{X}$. This operator finds the nearest point to a point in the action set $\mathcal{X}$ to the ellipsoid approximation $\mathcal{E}$.

From continuity of f and construction of $\mathcal{E}$, there exists a small neighbourhood W about $\mathrm{OPT}_{\mathcal{X}}(\theta)$ such that the $\bar{\mathbf{P}}_{\mathcal{X},\mathcal{E}}$ operator is injective. As in the proof of the Locally Constant Hessian (LCH) case, we shall restrict our attention to the neighbourhood of $U \cap W \triangleq V$.

Let $x \in \mathrm{OPT}_{\varepsilon,\mathcal{X}}(\theta)$. By definition,

$$x^\top \theta \geqslant \mathrm{OPT}_{\mathcal{X}}(\theta)^\top \theta - \varepsilon.$$

Now $x = \bar{\mathbf{P}}_{\mathcal{X},\mathcal{E}}(x) + r\vec{u}$ where $\vec{u}$ is the normal at x and r is $\left\|\bar{\mathbf{P}}_{\mathcal{X},\mathcal{E}}(x) - x\right\|$. Note that as $\varepsilon \to 0$, $\bar{\mathbf{P}}_{\mathcal{X},\mathcal{E}}(x) \to \mathrm{OPT}_{\mathcal{X}}(\theta)$ and $x \to \mathrm{OPT}_{\mathcal{X}}(\theta)$. Therefore $r \to 0$. Thus $r = O(\varepsilon^p)$, for some $p > 0$. Then, without loss of generality assuming $\|\theta\| = 1$, we have by Cauchy-Schwartz,

$$\bar{\mathbf{P}}_{\mathcal{X},\mathcal{E}}(x)^\top \theta + r \geqslant \bar{\mathbf{P}}_{\mathcal{X},\mathcal{E}}(x)^\top \theta + r\vec{u}^\top \theta = x^\top \theta \geqslant \mathrm{OPT}_{\mathcal{X}}(\theta)^\top \theta - \varepsilon = \mathrm{OPT}_{\mathcal{E}}(\theta)^\top \theta - \varepsilon \ .$$

Thus $\bar{\mathbf{P}}_{\mathcal{X},\mathcal{E}}(x) \in \mathrm{OPT}_{O(\varepsilon^{\min(1,p)}),\mathcal{E}}(\theta)$.

Let ε_1 be such that for all $\varepsilon \leqslant \varepsilon_1$, $\mathrm{OPT}_{\varepsilon,\mathcal{X}}(\theta) \subset V$, then we have $\bar{\mathbf{P}}_{\mathcal{X},\mathcal{E}}(\mathrm{OPT}_{\varepsilon,\mathcal{X}}(\theta)) \subset \mathrm{OPT}_{O(\varepsilon^{\min(1,p)}),\mathcal{E}}(\theta)$ for all $\varepsilon \leqslant \varepsilon_1$. □

Proof of Lemma 8.6. From the ellipsoidal case, we have by construction, $\theta' = \theta + \alpha u_d$, where $\alpha = O(\sqrt{\varepsilon})$ and u_d is the eigen-vector corresponding to the minimum eigen-value to ensure separation. This required showing that u_d and $\mathrm{OPT}(\theta)$ are roughly perpendicular to each other. For the current case, it is also possible to show that u_d and $\mathrm{OPT}_{\mathcal{X}}(\theta)$ is approximately perpendicular under the polynomial approximation assumption for $\varepsilon = O(\frac{1}{\sqrt{n}})$. We leave this as an easy exercise.

Thus let us construct $\theta' = \theta + \alpha u_d$, where α, would have to be carefully constructed, such that as $\varepsilon \to 0$, $\alpha \to 0$. Let, $x \in \mathrm{OPT}_{\varepsilon,\mathcal{X}}(\theta')$. By definition,

$$x^\top \theta' \geqslant \mathrm{OPT}_{\mathcal{X}}(\theta')^\top \theta' - \varepsilon.$$

For α very small, we have, $\|\theta'\| \approx 1$, and as before using the approximation operator and Cauchy-Schwartz, we have

$$\bar{\mathbf{P}}_{\mathcal{X},\mathcal{E}}(x)^\top \theta' + r_1 \geqslant \bar{\mathbf{P}}_{\mathcal{X},\mathcal{E}}(\mathrm{OPT}_{\mathcal{X}}(\theta'))^\top \theta' + r_2\vec{u}^\top \theta' - \varepsilon \ . \tag{63}$$

where r_1 is the approximation error $\left\|\bar{\mathbf{P}}_{\mathcal{X},\mathcal{E}}(x) - x\right\|$, r_2 is the approximation error $\left\|\bar{\mathbf{P}}_{\mathcal{X},\mathcal{E}}(\mathrm{OPT}_{\mathcal{X}}(\theta')) - \mathrm{OPT}_{\mathcal{X}}(\theta')\right\|$ and $\vec{u}$ is the normal at $\mathrm{OPT}_{\mathcal{X}}(\theta')$.

As $\varepsilon \to 0$, we have $\theta' \to \theta$ (because $\alpha \to 0$) and therefore $x \to \mathrm{OPT}_{\mathcal{X}}(\theta')$. Thus $r_1 = O(\varepsilon^q)$ for some $q > 0$ by the polynomial approximation. Also $\vec{u}$ approaches the normal at $\mathrm{OPT}_{\mathcal{X}}(\theta)$ and thus by continuity of f, there exists an ε_2 such that for all $\varepsilon \leqslant \varepsilon_2$ we have $\vec{u}^\top \theta' \geqslant 0$. Thus for small enough ε, we have for any $x \in \mathrm{OPT}_{\varepsilon,\mathcal{X}}(\theta')$,

$$\bar{\mathbf{P}}_{\mathcal{X},\mathcal{E}}(x)^\top \theta' + O(\varepsilon^q) \geqslant \bar{\mathbf{P}}_{\mathcal{X},\mathcal{E}}(\mathrm{OPT}_{\mathcal{X}}(\theta'))^\top \theta' - \varepsilon \,. \tag{64}$$

Note that as $\varepsilon \to 0$, we have $\bar{\mathbf{P}}_{\mathcal{X},\mathcal{E}}(\mathrm{OPT}_{\mathcal{X}}(\theta')) \to \mathrm{OPT}_{\mathcal{E}}(\theta)$ and $\mathrm{OPT}_{\mathcal{E}}(\theta') \to \mathrm{OPT}_{\mathcal{E}}(\theta)$, and thus we have $\left\|\bar{\mathbf{P}}_{\mathcal{X},\mathcal{E}}(\mathrm{OPT}_{\mathcal{X}}(\theta')) - \mathrm{OPT}_{\mathcal{E}}(\theta')\right\| = O(\varepsilon^r)$ for some $r > 0$. This implies,

$$\mathrm{OPT}_{\mathcal{E}}(\theta')^\top \theta' - \bar{\mathbf{P}}_{\mathcal{X},\mathcal{E}}(\mathrm{OPT}_{\mathcal{X}}(\theta'))^\top \theta' \leqslant \left\|\mathrm{OPT}_{\mathcal{E}}(\theta') - \bar{\mathbf{P}}_{\mathcal{X},\mathcal{E}}(\mathrm{OPT}_{\mathcal{X}}(\theta'))\right\| \|\theta'\| \leqslant O(\varepsilon^r) \,.$$

Using this relation we get,

$$\bar{\mathbf{P}}_{\mathcal{X},\mathcal{E}}(x)^\top \theta' \geqslant \mathrm{OPT}_{\mathcal{E}}(\theta')^\top \theta' - \varepsilon - O(\varepsilon^r) - O(\varepsilon^q).$$

Thus we have $\bar{\mathbf{P}}_{\mathcal{X},\mathcal{E}}(x) \in \mathrm{OPT}_{O(\varepsilon^{\min(1,q,r)}),\mathcal{E}}(\theta')$. Let ε_3 be such that for all $\varepsilon \leqslant \varepsilon_3$, $\mathrm{OPT}_{\varepsilon,\mathcal{X}}(\theta') \subset V$, then we have $\bar{\mathbf{P}}_{\mathcal{X},\mathcal{E}}(\mathrm{OPT}_{\varepsilon,\mathcal{X}}(\theta')) \subset \mathrm{OPT}_{O(\varepsilon^{\min(1,q,r)}),\mathcal{E}}(\theta)$ for all $\varepsilon \leqslant \varepsilon_3$.

Thus for all $\varepsilon \leqslant \min(\varepsilon_1, \varepsilon_2, \varepsilon_3)$, we have $\bar{\mathbf{P}}_{\mathcal{X},\mathcal{E}}(\mathrm{OPT}_{\varepsilon,\mathcal{X}}(\theta)) \subset \mathrm{OPT}_{O(\varepsilon^{\min(1,p)}),\mathcal{E}}(\theta) \subset \mathrm{OPT}_{O(\varepsilon^{\min(1,p,q,r)}),\mathcal{E}}(\theta)$

and $\bar{\mathbf{P}}_{\mathcal{X},\mathcal{E}}(\mathrm{OPT}_{\varepsilon,\mathcal{X}}(\theta')) \subset \mathrm{OPT}_{O(\varepsilon^{\min(1,q,r)}),\mathcal{E}}(\theta') \subset \mathrm{OPT}_{O(\varepsilon^{\min(1,p,q,r)}),\mathcal{E}}(\theta')$.

Now from the LCH case, we have $\mathrm{OPT}_{O(\varepsilon^{\min(1,p,q,r)}),\mathcal{E}}(\theta)$ and $\mathrm{OPT}_{O(\varepsilon^{\min(1,p,q,r)}),\mathcal{E}}(\theta')$ are disjoint if $\theta' = \theta + O(\sqrt{\varepsilon^{\min(1,p,q,r)}})u_d$, and by construction based on the injectivity of the approximation operator, $\mathrm{OPT}_{\varepsilon,\mathcal{X}}(\theta)$ and $\mathrm{OPT}_{\varepsilon,\mathcal{X}}(\theta')$ are as well disjoint.

This gives us by choice of $\varepsilon = O(\frac{1}{\sqrt{n}})$

$$\lambda_d \geqslant \Omega(n^s)$$

where $s \in (0, 1/2]$ and defined as $s = \frac{1}{2}\min(1, p, q, r)$. This completes our proof. □

Remark 8.7. For the LCH case, we have no approximation error and the exponents defined by p, q, r, are strictly more than 1 (they are in fact ∞) and we get back the results of Theorem 7.9 of $\Omega(\sqrt{n})$.

Remark 8.8. Note that the s defined in Theorem 8.2 is defined as $\frac{1}{2}\min(1, p, q, r)$, where each p, q, r represent the approximation error with respect to an LCH surface. Specifically with $\mathcal{E}$ being defined as the approximating Ellipsoid and $\mathcal{X}$ denoting the original action space, we have p defined as the rate at which $\left\|\bar{\mathbf{P}}_{\mathcal{X},\mathcal{E}}(x) - x\right\|$ goes to 0, that is $\left\|\bar{\mathbf{P}}_{\mathcal{X},\mathcal{E}}(x) - x\right\| = O(\varepsilon^p)$ for any $x \in \mathrm{OPT}_{\varepsilon,\mathcal{X}}(\theta)$. Similarly q is defined as $\left\|\bar{\mathbf{P}}_{\mathcal{X},\mathcal{E}}(x) - x\right\| = O(\varepsilon^q)$ for any $x \in \mathrm{OPT}_{\varepsilon,\mathcal{X}}(\theta')$ and r is defined as $\left\|\bar{\mathbf{P}}_{\mathcal{X},\mathcal{E}}(\mathrm{OPT}_{\mathcal{X}}(\theta')) - \mathrm{OPT}_{\mathcal{E}}(\theta')\right\| = O(\varepsilon^r)$.

8.1 Example of Locally Convex action space:

[5]In this subsection we demonstrate experimentally an action space which is *convex* and for which the minimum eigenvalue grows slower than $\Omega(\sqrt{n})$. Consider the action-space $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\|_{10} \leqslant 1\}$. Clearly, this is a convex set. We set the bandit parameter as $\theta = (1, \ldots, 1)$, as a vector of 1s of size d. We use Thompson Sampling as the representative algorithm and plot the growth of the minimum eigenvalue versus rounds n as in Section 4. In order to demonstrate the high probability phenomenon, we form a high confidence band of the mean observation of $\log \lambda_{\min}(V_n)/\log n$ with three standard deviations of width.We observe the the minimum eigenvalue grows less than $\Omega(\sqrt{n})$.

Justification : Let us first observe that the Hessian at $\mathrm{OPT}_{\mathcal{X}}(\theta)$ is positive definite and continuous for our choice of θ. We can also calculate that for our choice of the action space $\mathcal{X}$, any point $x \in \mathrm{OPT}_{\varepsilon,\mathcal{X}}(\theta)$,

$$\|x - \mathrm{OPT}_{\mathcal{X}}(\theta)\| = \varepsilon^{1/10} \,.$$

(This is easy to see for dimension 2.) Thus the $\left\|\bar{\mathbf{P}}_{\mathcal{X},\mathcal{E}}(x) - x\right\|$ as defined in the proof of Theorem 8.2 (see Section 8) is of the order of $\varepsilon^{1/10}$ and hence p as defined above is $1/10$. This means that $\lambda_{\min}(V_n) = \Omega(n^{0.1})$. This is true as observed from our experiments as observed in Figure 5. However we specifically also show that the minimum eigenvalue $\lambda_{\min}(V_n)$ is growing at a rate less than $\sqrt{n}$. This corroborates our results.

[5]Experiments can be found in `https://github.com/Debangshu93/analytic`

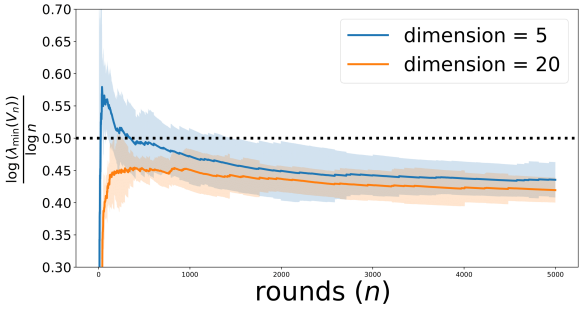


Figure 5: Scaling of the minimum eigenvalue of design matrix (generated by the Thompson sampling algorithm) with time for the action space $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\|_{10} \leqslant 1\}$. The plots represent averages over 20 independent trials. The X axis denotes the number of rounds n and the Y-axis denotes $\frac{\log \lambda_{\min}(V_n)}{\log n}$. We plot the mean trend of $\frac{\log \lambda_{\min}(V_n)}{\log n}$ along with three standard deviations. The dotted black line is the constant (exponent) $1/2$. Note that $\frac{\log \lambda_{\min}(V_n)}{\log n}$ settles at value below $1/2$. We plot this for two dimensions 5 and 20 to note the dependence on dimension d. We note that these corroborate with our theory.

9 MODEL SELECTION: PROOF OF RESULTS

Proof of Lemma 5.3. We consider doubling epochs, with lengths $n_i = 2^{i-1}n_1$, where n_1 is the initial epoch length and N is the total number of epochs. Then, from the doubling principle, we get

$$\sum_{i=1}^{N} 2^{i-1}n_1 = n \;\Rightarrow\; N = \log_2\left(1 + n/n_1\right) = \mathcal{O}\left(\log(n/n_1)\right).$$

Now, consider the i-th epoch. Let $\widehat{\theta}_{n_i}$ be the least square estimate of θ at the end of epoch i. The confidence interval at the end of epoch i, i.e., after OFUL is run with a norm estimate $b^{(i)}$ for n_i rounds with confidence level δ_i, is given by

$$\mathcal{B}_{n_i} = \left\{\theta \in \mathbb{R}^d : \|\theta - \widehat{\theta}_{n_i}\|_{(V_{n_i}+\lambda I)} \leqslant \sqrt{\beta_{n_i}(\delta_i)}\right\}.$$

Here $\beta_{n_i}(\delta_i)$ denotes the radius and Σ_{n_i} denotes the shape of the ellipsoid. Under Assumption 5.1, the ellipsoid takes the form

$$\mathcal{B}_{n_i} = \left\{\theta \in \mathbb{R}^d : \|\theta - \widehat{\theta}_{n_i}\| \leqslant \frac{\sqrt{\beta_{n_i}(\delta_i)}}{\sqrt{\lambda + \gamma_0 n_i}}\right\},$$

with probability at least $1 - \delta_i = 1 - \delta/2^{i-1}$. Here, we use the fact that $\gamma_{\min}(V_{n_i}) \geqslant \lambda + \gamma_0\sqrt{n_i}$ under Assumption 5.1, provided $n_i \geqslant n_0$. To ensure this, we choose $n_1 \geqslant n_0$. Now, note that $\theta \in \mathcal{B}_{n_i}$ with probability at least $1 - \delta_i$. Therefore, we have

$$\|\widehat{\theta}_{n_i}\| \leqslant \|\theta\| + \frac{\sqrt{\beta_{n_i}(\delta_i)}}{\sqrt{\lambda + \gamma_0\sqrt{n_i}}}$$

with probability at least $1 - 2\delta_i$. Recall that at the end of the i-th epoch, ALB set the estimate of $\|\theta\|$ to $b^{(i+1)} = \max_{\theta \in \mathcal{B}_{n_i}} \|\theta\|$. Then, from the definition of $\mathcal{B}_{n_i}$, we obtain

$$b^{(i+1)} = \|\widehat{\theta}_{n_i}\| + \frac{\sqrt{\beta_{n_i}(\delta_i)}}{\sqrt{\lambda + \gamma_0\sqrt{n_i}}} \leqslant \|\theta\| + 2\frac{\sqrt{\beta_{n_i}(\delta_i)}}{\sqrt{\lambda + \gamma_0\sqrt{n_i}}}$$

with probability higher than $1 - 2\delta_i$. If noise is distributed as $\mathcal{N}(0, 1)$, the confidence radius reads

$$\sqrt{\beta_{n_i}(\delta_i)} = b^{(i)}\sqrt{\lambda} + \sqrt{2\log(1/\delta_i) + d\log(1 + n_i/(\lambda d))},$$

We now substitute $n_i = 2^{i-1}n_1$ and $\delta_i = \frac{\delta}{2^{i-1}}$ to obtain $\sqrt{\lambda + \gamma_0\sqrt{n_i}} \geqslant 2^{\frac{i-1}{4}}\sqrt{\gamma_0\sqrt{n_1}}$, and

$$\frac{\sqrt{\beta_{n_i}(\delta_i)}}{\sqrt{\lambda+\gamma_0 n_i}} \leqslant C_1\frac{b^{(i)}\sqrt{\lambda}}{2^{\frac{i-1}{4}}\sqrt{\gamma_0\sqrt{n_1}}} + C_2\frac{i}{2^{\frac{i-1}{4}}\sqrt{\gamma_0\sqrt{n_1}}}\sqrt{2\log(1/\delta_1) + d\log(1+n_1/(\lambda d))}$$

for some universal constants C_1, C_2. Using this, we further obtain

$$\begin{aligned} b^{(i+1)} &\leqslant \|\theta\| + 2C_1\frac{b^{(i)}\sqrt{\lambda}}{2^{\frac{i-1}{4}}\sqrt{\gamma_0\sqrt{n_1}}} + 2C_2\frac{i}{2^{\frac{i-1}{4}}\sqrt{\gamma_0\sqrt{n_1}}}\sqrt{2\log(1/\delta_1) + d\log(1+n_1/(\lambda d))} \\ &\leqslant \|\theta\| + b^{(i)}\frac{p}{2^{\frac{i-1}{4}}}\sqrt{\frac{1}{\sqrt{n_1}}} + \frac{iq}{2^{\frac{i-1}{4}}}\sqrt{\frac{d}{\sqrt{n_1}}}, \end{aligned}$$

with probability at least $1-2\delta_i$, where we introduce the terms

$$p = \frac{2C_1\sqrt{\lambda}}{\sqrt{\gamma_0}} \quad \text{and} \quad q = \frac{2C_2}{\sqrt{\gamma_0}}\sqrt{2\log(1/\delta_1) + \log(1+n_1/\lambda)}.$$

Therefore, with probability at least $1-2\delta_i$, we obtain

$$b^{(i+1)} - b^{(i)} \leqslant \|\theta\| - \left(1 - \frac{p}{2^{\frac{i-1}{4}}}\frac{1}{n_1^{1/4}}\right)b^{(i)} + \frac{iq}{2^{\frac{i-1}{4}}}\frac{\sqrt{d}}{n_1^{1/4}}.$$

Note that by construction, $b^{(i)} \geqslant \|\theta\|$. Hence, provided $n_1 > \frac{2p^4}{2^i}$, we have

$$b^{(i+1)} - b^{(i)} \leqslant \frac{p}{2^{\frac{i-1}{4}}}\frac{\|\theta\|}{n_1^{1/4}} + \frac{iq}{2^{\frac{i-1}{4}}}\frac{\sqrt{d}}{n_1^{1/4}},$$

with probability at least $1-2\delta_i$. From the above expression, we have $\sup_i b^{(i)} < \infty$ with probability greater than or equal to $1 - \sum_i 2\delta_i = 1 - \sum_i 2\delta/2^{i-1} = 1 - 4\delta$. From the expression of $b^{(i+1)}$ and using the above fact, we get $\lim_{i\to\infty} b^{(i)} \leqslant \|\theta\|$. However, by construction $b^{(i)} \geqslant \|\theta\|$. Using this, along with the above observation, we obtain

$$\lim_{i\to\infty} b^{(i)} = \|\theta\|.$$

with probability exceeding $1-4\delta$. Therefore, we deduce that the sequence $\{b^{(1)}, b^{(2)}, ...\}$ converges to $\|\theta\|$ with probability at least $1-4\delta$, and hence our successive refinement algorithm is consistent.

Rate of Convergence. Since $b^{(i+1)} - b^{(i)} = \tilde{\mathcal{O}}\left(\frac{i}{2^{i/4}}\right)$ with high probability, the rate of convergence of the sequence $\{b^{(i)}\}_{i=1}^{\infty}$ is exponential in the number of epochs.

A uniform upper bound on $b^{(i)}$. Consider the sequences $\left\{\frac{i}{2^{\frac{i-1}{4}}}\right\}_{i=1}^{\infty}$ and $\left\{\frac{1}{2^{\frac{i-1}{4}}}\right\}_{i=1}^{\infty}$. Let t_i and u_i denote the i-th term of the sequences respectively. It is easy to see that $\sup_i t_i < \infty$ and $\sup_i u_i < \infty$, and that the sequences $\{t_i\}_{i=1}^{\infty}$ and $\{u_i\}_{i=1}^{\infty}$ are convergent. Now, we have

$$b^{(2)} \leqslant \|\theta\| + u_1\frac{pb^{(1)}}{n_1^{1/4}} + t_1\frac{q\sqrt{d}}{n_1^{1/4}}$$

with probability at least $1-2\delta$. Similarly, we write $b^{(3)}$ as

$$b^{(3)} \leqslant \|\theta\| + u_2\frac{pb^{(2)}}{n_1^{1/4}} + t_2\frac{q\sqrt{d}}{n_1^{1/4}} \leqslant \left(1 + u_2\frac{p}{n_1^{1/4}}\right)\|\theta\| + \left(u_1u_2\frac{p}{n_1^{1/4}}\frac{p}{n_1^{1/4}}b^{(1)}\right) + \left(t_1u_2\frac{p}{n_1^{1/4}}\frac{q\sqrt{d}}{n_1^{1/4}} + t_2\frac{q\sqrt{d}}{n_1^{1/4}}\right)$$

with probability at least $1-2\delta-\delta = 1-3\delta$. Similarly, we write expressions for $b^{(4)}, b^{(5)}, ...$ Now, provided $n_1 \geqslant C\,d^2\left(\max\{p,q\}\,b^{(1)}\right)^4$, where C is a sufficiently large constant, $b^{(i)}$ can be upper-bounded, with with probability at least $1 - \sum_i 2\delta_i = 1 - 4\delta$, as

$$b^{(i)} \leqslant c_1\|\theta\| + c_2$$

for all i, where $c_1, c_2 > 0$ are some universal constants, which are obtained from summing an infinite geometric series with decaying step size. □

Proof of Corollary 5.4. The cumulative regret of ALB is given by

$$R(n) \leqslant \sum_{i=1}^{N} R^{\texttt{OFUL}}(n_i, \delta_i, b^{(i)}),$$

where N denotes the total number of epochs and $R^{\texttt{OFUL}}(n_i, \delta_i, b^{(i)})$ denotes the cumulative regret of OFUL, when it is run with confidence level δ_i and norm upper bound $b^{(i)}$ for n_i episodes. Using the result of Abbasi-yadkori et al. (2011), we have

$$R^{\texttt{OFUL}}(n_i, \delta_i, b^{(i)}) = \mathcal{O}\left(b_i\sqrt{dn_i \log n_i} + d\sqrt{n_i \log n_i \log(n_i/\delta)}\right)$$

with probability at least $1 - \delta_i$. Now, using Lemma 5.3, we obtain

$$R(n) \leqslant (c_1\|\theta\| + c_2)\sum_{i=1}^{N} \mathcal{O}\left(\sqrt{dn_i \log n_i}\right) + \sum_{i=1}^{N} \mathcal{O}\left(d\sqrt{n_i \log n_i \log(n_i/\delta)}\right)$$

with probability at least $1 - 4\delta - \sum_i \delta_i$. Substituting $n_i = 2^{i-1}n_1$ and $\delta_i = \frac{\delta}{2^{i-1}}$, we get

$$R(n) \leqslant (c_1\|\theta\| + c_2)\sum_{i=1}^{N} \mathcal{O}\left(\sqrt{i\, dn_i \log n_1}\right) + \sum_{i=1}^{N} \mathcal{O}\left(\mathsf{poly}(i)\, d\sqrt{n_i \log n_1 \log(n_1/\delta)}\right)$$

with probability at least $1 - 4\delta - 2\delta = 1 - 6\delta$. Using the above expression, we get the regret bound

$$\begin{aligned}
R(n) &\leqslant \mathcal{O}\left((c_1\|\theta\| + c_2)\sqrt{d \log n_1} + d\sqrt{\log n_1 \log(n_1/\delta)}\right) \sum_{i=1}^{N} \mathsf{poly}(i)\sqrt{n_i} \\
&\leqslant \mathcal{O}\left((c_1\|\theta\| + c_2)\sqrt{d \log n_1} + d\sqrt{\log n_1 \log(n_1/\delta)}\right) \mathsf{poly}(N) \sum_{i=1}^{N} \sqrt{n_i} \\
&\leqslant \mathcal{O}\left((c_1\|\theta\| + c_2)\sqrt{d \log n_1} + d\sqrt{\log n_1 \log(n_1/\delta)}\right) \mathsf{polylog}(n/n_1) \sum_{i=1}^{N} \sqrt{n_i} \\
&\leqslant \mathcal{O}\left((c_1\|\theta\| + c_2)\sqrt{d \log n_1} + d\sqrt{\log n_1 \log(n_1/\delta)}\right) \mathsf{polylog}(n/n_1)\sqrt{n} \\
&= \mathcal{O}\left(\left(\|\theta\|\sqrt{dn \log n_1} + d\sqrt{n \log n_1 \log(n_1/\delta)}\right) \mathsf{polylog}(n/n_1)\right),
\end{aligned}$$

where we have used that $N = \mathcal{O}\left(\log(n/n_1)\right)$, and $\sum_{i=1}^{N} \sqrt{n_i} = \mathcal{O}(\sqrt{n})$. The above regret bound holds with probability greater than or equal to $1 - 6\delta$, which completes the proof. □

10 CLUSTERING IN MULTI AGENT BANDITS: PROOFS OF RESULT

Proof of Lemma 5.6. Let us look at the parameter estimate of agent i, and without loss of generality, assume that agent i belongs to cluster j. Since we let the agents play OFUL for n time steps, from Abbasi-yadkori et al. (2011), for agent i, we obtain

$$\|\widehat{\theta}^{(i)} - \theta_j^*\|_{\bar{V}_n} \leqslant 2\sqrt{d \log(n/\delta)},$$

where $\bar{V}_n = \sum_{t=1}^{n} x_{A_t,t} x_{A_t,t}^\top + \lambda I$, where A_t is the action of agent i at time t. The above holds with probability at least $1 - \delta$. Furthermore, we assume that OFUL is run with regularization parameter λ chosen as $\mathcal{O}(1)$. Continuing, we obtain

$$\sqrt{\lambda_{\min}(\bar{V}_n)}\, \|\widehat{\theta}^{(i)} - \theta_j^*\| \leqslant 2\sqrt{d \log(n/\delta)}.$$

We now use Assumption 5.1, with $\lambda = \mathcal{O}(1)$. Using Weyl's inequality, we obtain, $\lambda_{\min}(\bar{V}_n) > \frac{\gamma}{2}\sqrt{n}\sqrt{\log(d/\delta)}$, with probability at least $1 - \delta$. With this we have,

$$\|\widehat{\theta}^{(i)} - \theta_j^*\| \leqslant \frac{2\sqrt{2}}{\sqrt{\gamma} n^{1/4}} \frac{1}{\log(d/\delta)^{1/4}} \sqrt{d \log(n/\delta)} = \frac{1}{n^{1/4}} \frac{2\sqrt{2}\sqrt{d}}{\sqrt{\gamma}} \sqrt{\frac{\log(n/\delta)}{\log^{1/2}(d/\delta)}}.$$

From the separation condition on Δ and the choice of threshold η, we obtain

$$\|\widehat{\theta}^{(i)} - \theta_j^*\| \leqslant \frac{\eta}{2},$$

with probability at least $1 - 2\delta$. We now consider 2 cases:

Case I: Agents i and i' belong to same cluster j: In this setup we have

$$\begin{aligned}\|\widehat{\theta}^{(i)} - \widehat{\theta}^{(i')}\| &\leqslant \|\widehat{\theta}^{(i)} - \theta_j^*\| + \|\widehat{\theta}^{(i')} - \theta_j^*\| \\ &\leqslant \frac{\eta}{2} + \frac{\eta}{2} = \eta,\end{aligned}$$

with probability at least $1 - 4\delta$.

Case II: Agents i and i' belong to different cluster j and j' respectively: In this case we have

$$\begin{aligned}\|\widehat{\theta}^{(i)} - \widehat{\theta}^{(i')}\| &= \|(\widehat{\theta}^{(i)} - \theta_j^*) + (\theta_{j'}^* - \widehat{\theta}^{(i')}) - (\theta_{j'}^* - \theta_j^*)\| \\ &\geqslant \|\theta_j^* - \theta_{j'}^*\| - \|(\widehat{\theta}^{(i)} - \theta_j^*)\| - \|(\widehat{\theta}^{(i')} - \theta_{j'}^*)\| \\ &\geqslant \Delta - \frac{\eta}{2} - \frac{\eta}{2} = \eta,\end{aligned}$$

with probability at least $1 - 4\delta$, where we use the condition that $\Delta > 2\eta$.

From the above 2 cases, if we select the threshold to be $\eta = \frac{\Delta}{2}$, every pair of machines are correctly clustered with probability at least $1 - 4\delta$. Taking the union bound over all $\binom{N}{2}$ pairs, we obtain the result.

Regret of agent i: We now characterize the regret of agent i. Since agent i played `OFUL` for n steps, from Abbasi-yadkori et al. (2011), we obtain

$$R_i \leqslant \tilde{\mathcal{O}}\left(d\sqrt{n}\right)\log(1/\delta_1)$$

with probability at least $1 - \delta_1$.

□

11 HIGH PROBABILITY LOWER BOUND

We note that any optimistic algorithm chooses actions based on a high probability confidence set of the true θ, namely

$$\|\theta - \widehat{\theta}\|_{V_n} \leqslant c\sqrt{\ln n}$$

Naturally it would be very useful if we could get an estimate on the lower bound for $\lambda_{\min}(V_n)$ instead of what we have on $\lambda_{\min}(\mathbb{E}[V_n])$. More precisely we would like a theorem which suggests $\lambda_{\min}(V_n) \geqslant \sqrt{n}$ for all n more than some n_0, where n_0 depends upon the specific algorithm chosen and the geometry of the action space with high probability given that the algorithmic regret is $O(\sqrt{n})$.

For the time being this seems difficult in the present setup. However what we would like to emphasize is what such a result could establish. We illustrate two practical problems in the Applications sections one in Model Selection and the other in Clustering. For the time being we should emphasize that a direct corollary of this would be that a good regret algorithm would result in a best arm identification, given that the arm set is diverse enough. In this section we aim to discuss Assumption 5.1. We show that under a technical assumption, we can prove a high probability variant of Theorem 2.2 as well as supplement the high probability claim by experimental observations added in section 4.

Assumption 11.1 (Stability Assumption)**.** Let any algorithm π, satisfy, for all $k \in [n]$, the following:

$$\lambda_{\max}\Big(\mathbb{E}\left[\sum_{i=k+1}^{n} A_i A_i^T | \mathcal{F}_k\right] - \mathbb{E}\left[\sum_{i=k+1}^{n} A_i A_i^T | \mathcal{F}_{k-1}\right]\Big) \leqslant C \textit{ almost surely},$$

where $\{A_i\}_{i=1}^k$ are the actions selected by π, and $\mathcal{F}_k$ is a filtration such that $\{A_i\}_{i=1}^k$ are adapted to $\mathcal{F}_k$ and C is some positive constant.

We agree that bridging the gap between in-expectation and in-high-probability results is an important open direction. Stability is one such technical tool.

Example of stability assumption being satisfied Consider K armed bandit problem with arm means $\{\mu_i\}_{i=1}^K$. In this setting, standard algorithms like UCB or TS suffers instance dependent regret of $O(1)$ [6]. Thus the number of times any sub-optimal arm is played is at most a constant number of times in the entire horizon. This implies

$$\left\| \mathbb{E}\left[\sum_{i=k+1}^{n} A_i A_i^T | \mathcal{F}_k\right] - \mathbb{E}\left[\sum_{i=k+1}^{n} A_i A_i^T | \mathcal{F}_{k-1}\right] \right\| \leqslant \sum_{i=k+1}^{O(1)} 2 \left\|A_i\right\|^2 ,$$

for any k, and hence the stability assumption is trivially satisfied.

The assumption essentially implies that playing a random action in the middle of an algorithmic run will not affect the overall trajectory of the actions-played drastically. Under this assumption, we can show that $\lambda_{\min}(V_n) \geqslant \Omega(\sqrt{n})$ with high probability. To do so, we shall use matrix version of the Azuma-Hoeffding Inequality Tropp (2012, 2015). For completeness we present the result in the appendix (see Lemma 12.5). We shall first prove a minimum eigenvalue analogue of the matrix Azuma-Hoeffding Inequality.

Corollary 11.2. *Consider a finitely adapted sequence $\{\Delta_k\}$ of self adjoint matrices in dimension d and a fixed sequence of matrices $\{A_k\}$ of self-adjoint matrices such that $\mathbf{E}[\Delta_k|\mathcal{F}_{k-1}] = 0$ and $\Delta_k^2 \preccurlyeq A_k^2$ almost surely. Then for all $t \geqslant 0$, we have*

$$\mathbf{P}\{\lambda_{\min}(\sum_{k=1}^{n} \Delta_k) \leqslant -t\} \leqslant d e^{-\frac{t^2}{8\sigma^2}}$$

Proof. Note that

$$\mathbf{P}\{\lambda_{\min}(\sum_{k=1}^{n} \Delta_k) \leqslant -t\} = \mathbf{P}\{-\lambda_{\min}(\sum_{k=1}^{n} \Delta_k) \geqslant t\} = \mathbf{P}\{\lambda_{\max}(-\sum_{k=1}^{n} \Delta_k) \geqslant t\},$$

as $\lambda_{\max}(-A) = -\lambda_{\min}(A)$. Now, the proof follows by applying Theorem 12.5 to the sequence $\{-\Delta_k\}$. □

Now we are ready to state and prove the high probability version of Theorem 2.2.

Theorem 11.3. *Fix a $\delta \in (0, 1]$ and $n \geqslant n_0$,. Then, under the hypothesis of Theorem 2.2 and Assumption 11.1, with probability at least $1 - \delta$,*

$$\lambda_{\min}(V_n) \geqslant \gamma\sqrt{n} - (2 + C)\sqrt{8n \ln \frac{d}{\delta}}.$$

Furthermore, if $2 + C < \frac{\gamma}{\ln d}$, then for any $\delta \in (de^{-\frac{\gamma^2}{8(2+C)^2}}, 1)$, there exists a positive constant γ_2 such that

$$\lambda_{\min}(V_n) \geqslant \gamma_2\sqrt{n}$$

with probability more than $1 - \delta$. The constant γ_2 can be calculated as $\gamma_2 = \gamma - (2 + C)\sqrt{8 \ln \frac{d}{\delta}}$.

Proof. Let us choose the filtration $\mathcal{F}_k = \sigma(A_1, \cdots, A_k)$ for $k = 1, \cdots, n$, as the natural filtration associated with the action-sequence $\{A_i\}_{i=1}^k$ and define the martingale difference sequence $\{\Delta_k\}_{k=1}^n$ as

$$\Delta_k = \mathbb{E}[\sum_{i=1}^{n} A_i A_i^\top | \mathcal{F}_k] - \mathbb{E}[\sum_{i=1}^{n} A_i A_i^\top | \mathcal{F}_{k-1}]$$

Note that, by the stability property of conditional expectation, we obtain

$$\begin{aligned}
\Delta_k &= \mathbb{E}[\sum_{i=1}^{n} A_i A_i^\top | \mathcal{F}_k] - \mathbb{E}[\sum_{i=1}^{n} A_i A_i^\top | \mathcal{F}_{k-1}] \\
&= A_k A_k^\top - \mathbb{E}[A_k A_k^\top | \mathcal{F}_{k-1}] + \mathbb{E}[\sum_{i=k+1}^{n} A_i A_i^\top | \mathcal{F}_k] - \mathbb{E}[\sum_{i=k+1}^{n} A_i A_i^\top | \mathcal{F}_{k-1}],
\end{aligned}$$

[6] In our notation of Big O, we suppress the poly-logarithmic factors.

Therefore, we get by triangle inequality

$$\|\Delta_k\| \leqslant \left\|A_k A_k^\top - \mathbb{E}[A_k A_k^\top | \mathcal{F}_{k-1}]\right\| + \left\|\mathbb{E}[\sum_{i=k+1}^{n} A_i A_i^\top | \mathcal{F}_k] - \mathbb{E}[\sum_{i=k+1}^{n} A_i A_i^\top | \mathcal{F}_{k-1}]\right\| \leqslant 2 + C$$

where the last inequality follows by by assumption on the norm of the arms and the stability assumption 11.1. Therefore, by orthogonality of the martingale difference sequences, we have

$$\|\Delta_k^2\| = \|\Delta_k\|^2 \leqslant (2+C)^2 \triangleq D^2,$$

where we define the quantity $(2+C)$ as D. Thus we have $\Delta_k^2 \preccurlyeq D^2 \mathbb{I}_{d\times d}$ almost surely, where $\mathbf{I}_{d\times d}$ is the d-dimensional identity element and from the definition of σ in Theorem 12.5, $\sigma^2 = \|\sum_{i=1}^n D^2 \mathbf{I}_{d\times d}\| = nD^2$.

Furthermore, note that $\sum_{k=1}^n \Delta_k = \sum_{i=1}^n A_i A_i^\top - \mathbb{E}[\sum_{i=1}^n A_i A_i^\top] \triangleq V_n - \mathbb{E}[V_n]$. Then, using Corollary 11.2, we have for any $t \geqslant 0$, that

$$\mathbf{P}\{\lambda_{\min}(\sum_{k=1}^{n} \Delta_k) \leqslant -t\} = \mathbf{P}\{\lambda_{\min}(V_n - \mathbb{E}[V_n]) \leqslant -t\} \leqslant d e^{-\frac{t^2}{8nD^2}} .$$

Let us choose $t \geqslant \sqrt{8nD^2 \ln \frac{d}{\delta}}$ for any $\delta \in (0,1)$. Then, we have

$$\lambda_{\min}(V_n - \mathbb{E}[V_n]) \geqslant -\sqrt{8nD^2 \ln \frac{d}{\delta}}$$

with probability more than $1-\delta$. From Weyl's inequality (Lemma 12.4), we now obtain

$$\lambda_{\min}(V_n) - \lambda_{\min}(\mathbb{E}[V_n]) = \lambda_{\min}(V_n) + \lambda_{\max}(-\mathbb{E}[V_n]) \geqslant \lambda_{\min}(V_n - \mathbf{E}[V_n]) \geqslant -\sqrt{8nD^2 \ln \frac{d}{\delta}}$$

with probability more than $1-\delta$. Thus we get,

$$\lambda_{\min}(V_n) \geqslant \lambda_{\min}(\mathbf{E}[V_n]) - \sqrt{8nD^2 \ln \frac{d}{\delta}}$$

with probability more than $1-\delta$. Now, From Theorem 2.2 we have $\lambda_{\min}(\mathbb{E}[V_n]) \geqslant \gamma\sqrt{n}$ for some constant $\gamma > 0$ and for all $n \geqslant n_0$ for some n_0 which depends on γ and the algorithmic constant c. This gives us

$$\lambda_{\min}(V_n) \geqslant \gamma\sqrt{n} - \sqrt{8nD^2 \ln \frac{d}{\delta}}$$

with probability more than $1-\delta$ for all $n \geqslant n_0$ hence proving the first part of the theorem.

For the second part of the theorem, assume $2 + C < \frac{\gamma}{\ln d}$ then we have $d e^{\frac{\gamma^2}{8(1+C)^2}} < 1$. Thus we can find a δ in the open interval $(d e^{\frac{\gamma^2}{8(1+C)^2}}, 1)$. Choosing such a δ, we see $\gamma_2 \triangleq \gamma - \sqrt{8D^2 \ln \frac{d}{\delta}} > 0$. Thus we have

$$\lambda_{\min}(V_n) \geqslant \gamma_2 \sqrt{n}$$

with probability more than $1-\delta$ for all $n \geqslant n_0$

□

12 TECHNICAL LEMMAS

Lemma 12.1 (Information Inequality)**.** *Kaufmann et al. (2016) Let θ and θ' be two bandit parameters with policy induced measures $\mathbb{P}$ and $\mathbb{P}'$. Then for any measurable $Z \in (0,1)$, we have*

$$\mathrm{KL}(\mathbb{P}||\mathbb{P}') \geqslant \mathrm{KL}(\mathrm{Ber}(\mathbb{E}_\theta[Z])||\mathrm{Ber}(\mathbb{E}_{\theta'}[Z]))$$

Lemma 12.2 (Divergence Decomposition). *Lattimore and Szepesvari (2017); Lattimore and Szepesvári (2020) Let $\mathbb{P}$ and $\mathbb{P}'$ be the action-observation sequence for a fixed bandit policy π interacting with a linear bandit with standard Gaussian noise and parameters θ and θ' respectively.We have*

$$\mathrm{KL}(\mathbb{P}||\mathbb{P}') = \frac{1}{2}\mathbb{E}_\theta\left[\sum_{t=1}^{n}\langle A_t, \theta - \theta'\rangle^2\right] = \frac{1}{2}\|\theta - \theta'\|^2_{\bar{G}_n}$$

Lemma 12.3 (Davis-Kahan $\sin\theta$ theorem). *Stewart (1990) Let A and H be two symmetric $d \times d$ matrices. Define $\tilde{A} = A + H$ Let the spectral decomposition of A and $\tilde{A}$ be $A = \sum_{i=1}^{d}\lambda_i u_i u_i^T$ and $\tilde{A} = \sum_{i=1}^{d}\tilde{\lambda}_i\tilde{u}_i\tilde{u}_i^T$, respectively, where $\lambda_1 \gg \lambda_2 \geqslant \cdots \geqslant \lambda_d$ and $\tilde{\lambda}_1 \gg \tilde{\lambda}_2 \geqslant \cdots \geqslant \tilde{\lambda}_d$. Then*

$$\|u_1^T\begin{bmatrix}\tilde{u}_2 & \tilde{u}_3 & \cdots & \tilde{u}_d\end{bmatrix}\|_2 \leqslant \frac{\|H\|}{\delta},$$

where δ is the eigenvalue separation between λ_1 and $\tilde{\lambda}_2, \cdots, \tilde{\lambda}_d$.

Lemma 12.4 (Weyl's Inequality). *Stewart (1990) For symmetric A and H we have*

$$\lambda_i(A + H) \leqslant \lambda_i(A) + \lambda_{\max}(H)$$

Theorem 12.5 (Matrix Azuma). *Tropp (2012, 2015) Consider a finitely adapted sequence $\{\Delta_k\}$ of self adjoint matrices in dimension d and a fixed sequence of matrices $\{A_k\}$ of self-adjoint matrices such that $\mathbf{E}[\Delta_k|\mathcal{F}_{k-1}] = 0$ and $\Delta_k^2 \preccurlyeq A_k^2$ almost surely. Then for all $t \geqslant 0$, we have*

$$\mathbf{P}\{\lambda_{\max}(\sum_{k=1}^{n}\Delta_k) \geqslant t\} \leqslant de^{-\frac{t^2}{8\sigma^2}},$$

where $\sigma^2 = \|\sum_{k=1}^{n}A_k^2\|$.