Linear bandits with polylogarithmic minimax regret

Josep Lumbreras   JOSEP.LUMBRERAS@U.NUS.EDU
Centre for Quantum Technologies, National University of Singapore, Singapore

Marco Tomamichel   MARCO.TOMAMICHEL@NUS.EDU.SG
Department of Electrical and Computer Engineering and Centre for Quantum Technologies, National University of Singapore, Singapore

Abstract

We study a noise model for linear stochastic bandits for which the subgaussian noise parameter vanishes linearly as we select actions on the unit sphere closer and closer to the unknown vector. We introduce an algorithm for this problem that exhibits a minimax regret scaling as $\log^3(T)$ in the time horizon T, in stark contrast the square root scaling of this regret for typical bandit algorithms. Our strategy, based on weighted least-squares estimation, achieves the eigenvalue relation $\lambda_{\min}(V_t) = \Omega(\sqrt{\lambda_{\max}(V_t)})$ for the design matrix V_t at each time step t through geometrical arguments that are independent of the noise model and might be of independent interest. This allows us to tightly control the expected regret in each time step to be of the order $O(\frac{1}{t})$, leading to the logarithmic scaling of the cumulative regret.

Keywords: multi-armed bandit, linear stochastic bandit, adaptive noise, minimax regret

1. Introduction

Background. The linear bandit problem (see, e.g., Lattimore and Szepesvári (2020)) is one of the most versatile supervised learning frameworks with both applications and strong theoretical guarantees. In this setting, the learner interacts with an environment sequentially, and at each time step t selects an action from a set $\mathcal{A}_t \subset \mathbb{R}^d$ and receives a reward, r_t. The reward is a random variable that can be decomposed as $r_t = \mathbb{E}[r_t] + \epsilon_t$, where ϵ_t is statistical noise (with zero mean) and the expectation, $\mathbb{E}[r_t] = \langle \theta, a_t \rangle$, is given by the inner product of the action $a_t \in \mathcal{A}_t$ with an unknown vector θ from a set of environments $\mathcal{E} \subset \mathbb{R}^d$. The goal of the learner is to minimize the cumulative regret over T rounds, where the regret is the difference $\max_{a \in \mathcal{A}_t} \langle \theta, a \rangle - \mathbb{E}[r_t]$, for all possible vectors $\theta \in \mathcal{E}$, i.e., the minimax regret. Here, we focus on the case where both the sets $\mathcal{A}_t$ and $\mathcal{E}$ are continuous. The challenge posed by linear stochastic bandits lies in efficiently learning the optimal action for θ while simultaneously choosing actions that minimize the accumulated regret, an intriguing trade-off that our work aims to bring into a new regime.

Numerous algorithms have been proposed to address these challenges for the case of continuous action sets, most prominently LinUCB (see Dani et al. (2008); Rusmevichientong and Tsitsiklis (2010); Abbasi-Yadkori et al. (2011)) and Thompson sampling (see Abeille and Lazaric (2017)). Both of these algorithms share the feature that they achieve a minimax regret scaling as $\tilde{O}(\sqrt{T})$, where we are neglecting logarithmic terms and ignoring the dependence on d. In fact, under the prevalent assumption that ϵ_t follows a σ-subgaussian distribution for some finite $\sigma > 0$, this scaling can be shown to be optimal for various choices of action and environment sets (see Dani et al. (2008); Shamir (2015); Rusmevichientong and Tsitsiklis (2010)), where the latter work in particular covers the case of unit ball that is of interest in our work. Bandits with heteroscedastic noise have

also been studied (see Kirschner and Krause (2018)), where the learner has access to a link function that maps each action to the subgaussian parameter of the noise. The main difference with our model is that our noise also depends on the unknown vector.

Our model and main result. In this paper, we present the first non-trivial noise model and corresponding algorithm that achieves a polylogarithmic regret scaling in the time horizon T for linear bandits. We study a stochastic linear bandit with action and environment sets $\mathcal{A}_t = \mathcal{E} = \mathbb{S}^d = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$ and reward model given by $r_t = \langle \theta, a_t \rangle + \epsilon_t$ where ϵ_t is conditionally σ_t-subgaussian with

$$\sigma_t^2 \leq 1 - \langle \theta, a_t \rangle^2 \approx \left\| \theta - a_t \right\|_2^2, \tag{1}$$

where the approximation holds when $\langle \theta, a_t \rangle$ is close to 1.

The proposed model using a noise parameter growing with the distance between unknown vector and action is natural in some contexts. For recommendation systems, for example, it is natural to assume that the choice of a user becomes more certain when the recommendation (action) is close to the user's preference (unknown vector), and the reward may even become deterministic when the recommendation fits the preference perfectly. However our original motivation for the noise scaling comes from a very concrete problem in quantum information, specifically regret-optimal tomography of an unknown pure quantum state (Lumbreras et al. (2022)). In quantum mechanics measurement outcomes are random with probabilities determined by Born's rule. However, the variance of probabilistic measurement outcomes decreases quadratically for projections aligned with the unknown pure state.

Beyond specific applications, we believe that a model with decaying noise should be of independent interest as it allows to break the square root minimax regret barrier for continuous action sets, and we believe this is the first model that does this.

Analytically, this noise model is interesting because the usual methods for deriving lower bounds on the minimax regret fail. First, bounds where the unknown vector θ is taken from the unit ball as in Rusmevichientong and Tsitsiklis (2010) or (Lattimore and Szepesvári, 2020, Chapter 24, Theorem 24.2) do not apply since the worst-case unknown vectors they construct are far from the surface of the ball. However, for finite and constant σ-subgaussian noise we provide an adaptation of the lower bound for logistic bandits given in Abeille et al. (2021) and we obtain a minimax regret scaling of $\Omega(\sigma d \sqrt{T})$ in Appendix D. Nonetheless, these arguments fail for the noise model in (1) since they rely crucially on the fact that the distributions induced by two close unknown states have large overlap, which is no longer guaranteed when the noise vanishes. We discuss this in more detail in Appendix E. The lack of applicable lower bounds opens the possibility that we can find algorithms that break the $\Omega(\sqrt{T})$ barrier for the regret.

Indeed, our main result is an algorithm LinUCB-VN (see Algoirthm 1), based on LinUCB but with a weighted least square estimator that adapts to the vanishing noise. Our algorithm achieves a polylogarithmic scaling of the minimax regret. The main result can be stated as follows, and holds in the model prescribed above. We give the precise statement of the result in Section 6.

Theorem 1 (main result, informal version) *For any $T \in \mathbb{N}$ there exists an instance of* LinUCB-VN *such that, for any $\theta \in \mathcal{E}$, we have*

$$\mathbb{E}[\text{Regret}(T)] = O\left(d^4 \log^3(T)\right). \tag{2}$$
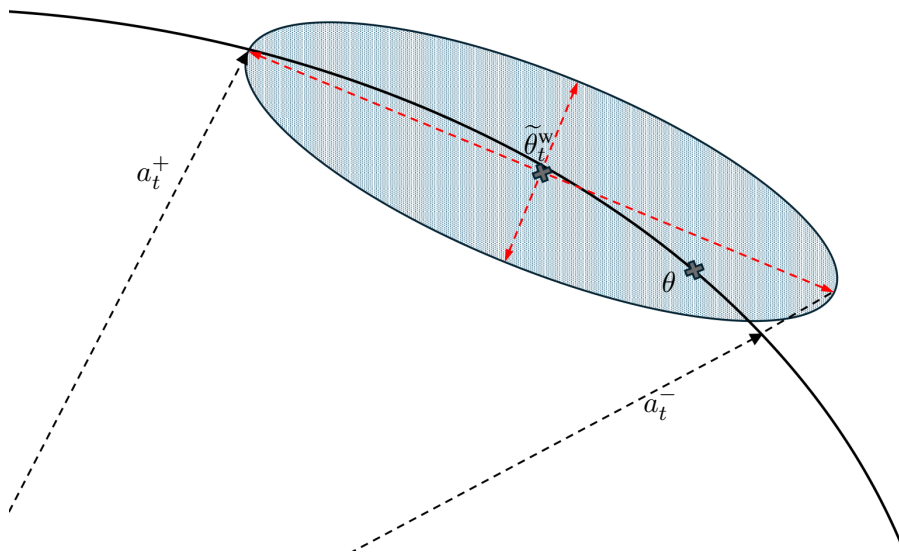
Figure 1: Scheme for the choice of actions a_t^+, a_t^- of LinUCB-VN. The actions are selected as the projections of the extremal points across the largest axis of the confidence region centered around a weighted least squares estimator $\widetilde{\theta}_t^{\mathrm{w}}$ of the unknown parameter θ. This choice is sufficient to increase the minimum eigenvalue of V_t such that the relation $\lambda_{\min}(V_t) = \Omega(\sqrt{\lambda_{\max}(V_t)})$ is satisfied. Moreover, the actions a_t^+ and a_t^- are sufficiently close to θ to keep the regret small.

The common linear bandit model describes a well-tuned trade-off between exploration and exploitation. Our breaking of the square root regret barrier, which we also observe numerically (see Figure 2), is due to the fact that exploitation can dominate in our model. This is due to the fact that playing actions near the unknown parameter reduces the statistical noise and we will see that we need to explore only the local neighborhood of our current estimate. One very rough way to think about our algorithm is that it tries to select actions as if it were trying to tune into an analog radio station. When we are tuning into a radio station we do not just randomly turn the knob but once we have locked to a signal (the noise starts to decrease) we only need to do small rotations (local exploration of the algorithm) to correct and find it exactly in a few steps.

Technical challenges. Our approach is based on an optimistic strategy like LinUCB, however, the standard regret analysis fails to give tight bounds on the regret. More precisely, if reg_t is the instantaneous regret, the usual technique for optimistic strategies used in the literature to bound the cumulative regret is to use the Cauchy–Schwarz inequality, i.e., $\mathrm{Regret}(T) = \sum_{t=1}^T \mathrm{reg}_t \leq (T \sum_{t=1}^T \mathrm{reg}_t^2)^{\frac{1}{2}}$ and then assert that $\sum_{t=1}^T \mathrm{reg}_t^2 = O(\log(T))$ using the elliptical potential lemma. This procedure always introduces a term $\sqrt{T}$ and for that reason, we need to develop a new technique to upper bound the cumulative regret based on a tight control of the instantaneous regret.

Our approach is thus to bound the instantaneous regret reg_t with high probability for all $t \in [T]$. The algorithm that we design satisfies $\mathrm{reg}_t \sim \|\theta - a_t\|_2^2 \sim 1/\lambda_{\min}(V_t)$ where V_t is the design matrix $V_t = \sum_{s=1}^t w_s a_s a_s^\mathsf{T}$ and w_s is a weight associated to a_s. Thus, our regret analysis will require to control the growth of $\lambda_{\min}(V_t)$. From Banerjee et al. (2023) we take the idea that for

smooth action sets (in particular the unit sphere) all strategies that minimize the regret must achieve $\lambda_{\min}(\mathbb{E}[V_t]) = \Omega(\sqrt{t})$ and generalize it to an equivalent condition in terms of eigenvalues of V_t that is noise independent. More precisely, we come up with a geometric construction that ensures that $\lambda_{\min}(V_t) = \Omega(\sqrt{\lambda_{\max}(V_t)})$ by choosing actions that explore the confidence region in the direction of the current estimate in a structured way (see Figure 1). We then prove that when applying this construction to our noise model we can achieve $\lambda_{\min}(V_t) = \Omega(t)$, which yields a polylogarithmic cumulative regret since $\text{reg}_t \sim \|\theta - a_t\|_2^2 \sim 1/t$.

2. Notation and model

We present the notation that we will use throughout the paper. For $k \in \mathbb{N}$ we use $[k] = \{1, ..., k\}$. The set of strictly positive numbers is denoted as $\mathbb{R}_{>0} = \{x \in \mathbb{R} : x > 0\}$ and the set of positive semi-definite matrices as $\mathsf{P}_+^d = \{X \in \mathbb{R}^{d \times d} : X \geq 0\}$. For two real vectors $u, v \in \mathbb{R}^d$ we denote its inner product as $\langle u, v \rangle = \sum_{i=1}^d u_i v_i$ and the Euclidean norm as $\|x\|_2^2 = \langle x, x \rangle$. The d-dimensional unit sphere is $\mathbb{S}^d = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$ and $\mathbb{B}_r^d(c) = \{x \in \mathbb{R}^d : \|x - c\|_2 \leq r\}$ is the ball of radius $r > 0$ centered around $c \in \mathbb{R}^d$. For a vector $x \in \mathbb{R}^d$ and semidefinite positive matrix $A \geq 0$ we denote the weighted norm of x by A as $\|x\|_A^2 = \langle x, Ax \rangle$. Given an Hermitian matrix $X \in \mathbb{R}^{d \times d}$ we denote $\lambda_{\max}(X)$ and $\lambda_{\min}(X)$ its maximum and minimum eigenvalues, respectively, and by $\lambda_i(X)$ its i-th smallest eigenvalue such that $\lambda_{\min}(X) = \lambda_1(X) \leq \lambda_2(X) \leq \cdots \leq \lambda_d(X) = \lambda_{\max}(X)$.

We consider a stochastic linear bandit with unknown parameter $\theta \in \mathbb{S}^d$ and action set $\mathcal{A}_t = \mathcal{A} = \mathbb{S}^d$. The learner interacts with the environment over a time horizon of length $[T]$. At time step $t \in [T]$ the learner selects an action $a_t \in \mathcal{A}$ and samples a reward

$$r_t = \langle \theta, a_t \rangle + \epsilon_t, \tag{3}$$

where ϵ_t is σ_t-subgaussian, i.e., $\mathbb{E}[\epsilon_t | \mathcal{F}_{t-1}] = 0$ and $\mathbb{E}[\exp(\lambda \epsilon_t) | \mathcal{F}_{t-1}] \leq \exp(\lambda^2 \sigma_t^2 / 2)$ where $\mathcal{F}_{t-1} = \sigma\{a_1, r_1, ..., a_{t-1}, r_{t-1}, a_t\}$ is the filtration of all information up to time step t before reward r_t is observed. We consider a noise model such that the subgaussian parameter satisfies

$$\sigma_t^2 \leq 1 - \langle \theta, a_t \rangle^2. \tag{4}$$

We call this bandit a *stochastic linear bandit with linear vanishing noise*. The learner uses a policy $\pi = (\pi_t)_{t=1}^\infty$ where $\pi_t(a_1, r_1, ..., a_{t-1}, r_{t-1}, a_t)$ maps the history up to time step t to a probability distribution over $\mathcal{A}$. The goal of the learner is to minimize the regret defined by

$$\text{Regret}(T) := \sum_{t=1}^T 1 - \langle \theta, a_t \rangle = \frac{1}{2} \sum_{t=1}^T \|\theta - a_t\|_2^2. \tag{5}$$

where we used that the maximum expected reward is $\max_{a \in \mathcal{A}} \langle a, \theta \rangle = 1$ since $a_t, \theta \in \mathbb{S}^d$ and the last equality follows from $\|x - y\|_2^2 = \langle x - y, x - y \rangle = 2 - 2\langle x, y \rangle$ for $x, y \in \mathbb{S}^d$.

3. Weighted regularized least squares estimator and confidence region

Our strategy is designed to select actions that are close to the unknown parameter and to quantify it we use the confidence region based on the regularized least squares estimator (RLSE) used in the LinUCB algorithm. Moreover, since our noise model is decreasing for actions near the unknown

parameter we use a weighted version. In this Section, we review the basic results that we will use later. The weights of our least squares estimator will be an estimator of the subgaussian parameter $\hat{\sigma}_t^2$ for the noise ϵ_t. This will allow us to bias our estimator to actions with less noise. We take the idea of using a weighted least squares estimator from Kirschner and Krause (2018) where they apply it to linear bandits with heteroscedastic noise where they can weight the actions with the exact subgaussian parameter. In our setting we need an estimator since the noise depends on the unknown parameter; however, our setting is more concrete and this allows us to design a simpler strategy with theoretical guarantees on the regret.

At time step $t \in [T]$ for action $a_t \in \mathcal{A}$ we define the estimator of the noise as

$$\hat{\sigma}_t^2 : \mathcal{H}_{t-1} \times \mathcal{A} \to \mathbb{R}_{>0}, \tag{6}$$

where $\mathcal{H}_{t-1} = (a_1, r_1, ..., a_{t-1}, r_{t-1})$ contains the past information of rewards and actions and it is independent of the unknown parameter θ. To simplify the notation we will often not mention the dependence on the past actions and rewards and simply write $\hat{\sigma}_t^2(a)$.

The weighted version of the linear least square estimator is defined with the following minimization problem

$$\widetilde{\theta}_t^{\mathrm{w}} := \operatorname*{argmin}_{\theta' \in \mathbb{R}^d} \sum_{s=1}^{t} \frac{1}{\hat{\sigma}_s^2(a)} \left(r_s - \langle \theta', a_s \rangle\right)^2 + \lambda \|\theta'\|_2, \tag{7}$$

where $\lambda \in \mathbb{R}_{>0}$ is a regularization term. The analytical solution is

$$\widetilde{\theta}_t^{\mathrm{w}} = V_t^{-1}(\lambda) \sum_{s=1}^{t} \frac{1}{\hat{\sigma}_s^2(a)} a_s r_s, \tag{8}$$

where

$$V_t(\lambda) = \lambda I + \sum_{s=1}^{t} \frac{1}{\hat{\sigma}_s^2(a)} a_s a_s^{\mathsf{T}}, \tag{9}$$

is the *design matrix*. Using the above estimator for θ we can define a confidence region under the condition that $\hat{\sigma}_t^2$ is a good estimator for σ_t^2 for any t. The Lemma below formalizes this result. The proof is an adaptation of the regularized least squares estimator confidence region in Lattimore and Szepesvári (2020)[Chapter 20] and we include it in the Appendix A for completeness.

Lemma 2 *Let $\delta \in (0, 1)$ and $(a_t)_{t=1}^{\infty}$ be the actions selected by some policy with corresponding rewards $(r_t)_{t=1}^{\infty}$ given by $r_t = \langle \theta, a_t \rangle + \epsilon_t$, where $\theta \in \mathbb{S}^d$ is the unknown parameter and ϵ_t is σ_t-subgaussian. Let $\hat{\sigma}_t^2$ be an estimator of the form (6) and define the following event,*

$$G_t := \left\{ \left((r_s, a_s)_{s=1}^{t-1}, a_t\right) : \sigma_s^2(a_s) \le \hat{\sigma}_s^2(a_1, r_1, \ldots, a_{s-1}, r_{s-1}, a_s) \ \forall s \in [t] \right\}. \tag{10}$$

Then we can define the following confidence region

$$\mathcal{C}_t := \{\theta' \in \mathbb{R}^d : \|\theta' - \widetilde{\theta}_t^{\mathrm{w}}\|_{V_t(\lambda)}^2 \le \beta_{t,\delta}\}, \tag{11}$$

where V_t and θ_t^{w} are defined with $\hat{\sigma}_s^2$ and

$$\beta_{t,\delta} = \left(\sqrt{2 \log \frac{1}{\delta} + \log \left(\frac{\det(V_t(\lambda))}{\det(V_0(\lambda))} \right)} + \sqrt{\lambda} \right)^2. \tag{12}$$

Then

$$\Pr \left[\forall s \in [t] : \theta \in \mathcal{C}_s^{\mathrm{wls}} \big| G_t \right] \geq 1 - \delta. \tag{13}$$

4. Algorithm for linear bandits with vanishing noise: LinUCB-VN

In this Section, we give the specific algorithm that minimizes the regret for the stochastic linear bandit with linear vanishing noise described in Section 2. The algorithm is based on the principle of "optimism in the face of uncertainty" (OFU) or upper confidence bounds (UCB). We name the algorithm LinUCB-VN, where VN stands for vanishing noise. The idea is to design a strategy that minimizes the regret (5) which means that we have to select actions close to θ keeping the trade-off of exploration and exploitation. The LinUCB-VN algorithm is designed to keep the relation $\lambda_{\min}(V_t) = \Omega(\sqrt{\lambda_{\max}(V_t)})$ at each time step t in line with the result of Banerjee et al. (2023) wich states that all strategies that minimize regret must achieve $\lambda_{\min}(V_t) = \Omega(\sqrt{t})$. This will allow us to bound the regret at each time step t without the need of using the Cauchy-Schwartz inequality in combination with the elliptical potential lemma that gives immediately a $O(\sqrt{T})$ regret (see (Lattimore and Szepesvári, 2020, Chapter 19)). Now we state the exact algorithm and in the subsequent sections we discuss the technical results that allow us to analyze the regret scaling.

Our algorithm works in batches of $2(d-1)$ rounds and with a slightly abuse of notation we will label each batch with t. At each batch $t \geq 1$ the algorithm selects the following actions, for $i \in [d-1]$:

$$a_{t,i}^{\pm} := \frac{\widetilde{a}_{t,i}^{\pm}}{\|\widetilde{a}_{t,i}^{\pm}\|_2}, \quad \text{where} \quad \widetilde{a}_{t,i}^{\pm} := \theta_{t-1}^w \pm \frac{1}{\sqrt{\lambda_{t-1,1}}} v_{t-1,i} \quad \text{and} \quad \theta_t^w := \frac{\widetilde{\theta}_t^w}{\|\widetilde{\theta}_t^w\|_2} \tag{14}$$

is the normalized least squares estimator of $\widetilde{\theta}_t^w$ defined in (7), $v_{t-1,i}$ is the normalized eigenvector with eigenvalue $\lambda_{t-1,i} = \lambda_i(V_{t-1}(\lambda))$. The design matrix V_t is updated at each batch t as

$$V_t(\lambda) := V_{t-1}(\lambda) + \omega(V_{t-1}(\lambda)) \sum_{i=1}^{d-1} \left(a_{t,i}^+(a_{t,i}^+)^{\mathsf{T}} + a_{t,i}^-(a_{t,i}^-)^{\mathsf{T}} \right), \tag{15}$$

with

$$\omega(V_{t-1}(\lambda)) = \frac{\sqrt{\lambda_{\max}(V_{t-1}(\lambda))}}{12\sqrt{d-1}\beta_{t-1,\delta}}, \quad \hat{\sigma}_t^2(a_{t,i}^{\pm}) = \frac{1}{\omega(V_{t-1}(\lambda))}, \tag{16}$$

and $\beta_{t,\delta}$ defined as in Lemma 2 with input V_t (15). When is clear from the context we will denote $V_t(\lambda)$ simply as V_t. We state the pseudo-code of LinUCB-VN below.
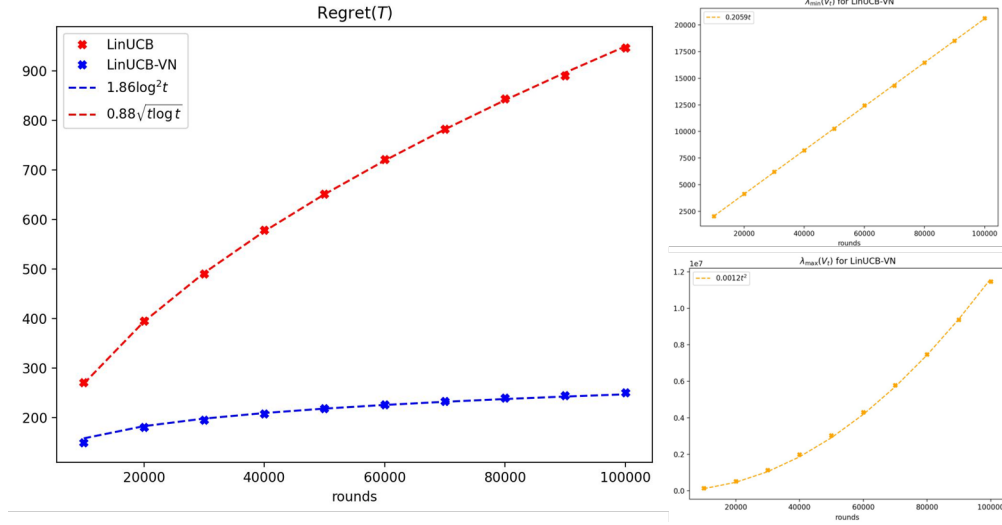
Figure 2: We numerically test LinUCB and LinUCB-VN in a linear bandit with action set $\mathcal{A} = \mathbb{S}^2$ and reward model $r_t = \mathcal{N}(\langle \theta, a_t \rangle, 1 - \langle \theta, a_t \rangle^2)$. Each point in the graphic is run independently and averaged over 100 instances for random environments $\theta \in \mathbb{S}^2$. *Left plot:* Scaling of the regret for LinUCB algorithm and LinUCB-VN. We fit the functions $R(t) = 1.86 \log^2 t$ for LinUCB-VN and $R(t) = 0.88\sqrt{t \log t}$ for LinUCB. *Right plots:* Scaling of the maximum and minimum eigenvalue of the matrix V_t for LinUCB-VN. The scaling shows the relation $\lambda_{\min} = \Omega(\sqrt{\lambda_{\max}})$. We fit the function $\lambda_{\min}(V_t) = 0.2059t$ for the minimum eigenvalue and $\lambda_{\max}(V_t) = 0.0012t^2$ for the maximum eigenvalue. The behavior $\lambda_{\min}(V_t) = \Theta(t)$ is the one that gives us the theoretical guarantee of polylogarithmic scaling of the regret.

Algorithm 1 LinUCB-VN

Require: $\lambda_0 \in \mathbb{R}_{>0}$, $\omega : \mathbf{P}_+^d \to \mathbb{R}_{\geq 0}$
Set initial design matrix $V_0 \leftarrow \lambda_0 \mathbb{I}_{d \times d}$
Choose initial estimator $\theta_0 \in \mathbb{S}^d$ for θ at random
for $t = 1, 2, \cdots$ **do**
  Optimistic action selection
  for $i = 1, 2, \cdots d - 1$ **do**
    Select actions $a_{t,i}^+$ and $a_{t,i}^-$ according to Eq. (14)
    Receive associated rewards $r_{t,i}^+$ and $r_{t,i}^-$
  end
  Update estimator of sub-gaussian noise for $a_{t,i}^+$
  $\hat{\sigma}_t^2 \leftarrow \frac{1}{\omega(V_{t-1}(\lambda_0))}$ for $t \geq 2$ or $\hat{\sigma}_t^2 \leftarrow 1$ for $t = 1$.
  Update design matrix and RLSE
  $V_t(\lambda_0) \leftarrow V_{t-1}(\lambda_0) + \frac{1}{\hat{\sigma}_t^2} \sum_{i=1}^{d-1} \left(a_{t,i}^+ (a_{t,i}^+)^{\mathsf{T}} + a_{t,i}^- (a_{t,i}^-)^{\mathsf{T}} \right)$
  $\widetilde{\theta}_t^{\mathrm{w}} \leftarrow V_t^{-1}(\lambda_0) \sum_{s=1}^{t} \frac{1}{\hat{\sigma}_t^2} \sum_{i=1}^{d-1} (a_{s,i}^+ r_{t,i}^+ + a_{s,i}^- r_{t,i}^-)$
end

---

5. Actions and eigenvalue analysis of design matrix

In this Section, we present our first main result which is an analysis that shows that the actions of the LinUCB-VN algorithm satisfy the relation $\lambda_{\min}(V_t) = \Omega(\sqrt{\lambda_{\max}(V_t)})$ at each batch t. Moreover, our result is very general since our analysis is independent of the randomness of the algorithm and also the noise model. We give the main ideas of the proof and give details in Appendix B.

Theorem 3 *Let $d \geq 2$, $\{c_t\}_{t=0}^{\infty} \subset \mathbb{S}^d$ be a sequence of normalized vectors and $\omega : \mathrm{P}_+^d \to \mathbb{R}_{\geq 0}$ a function such that*

$$\omega(X) \leq C\sqrt{\lambda_{\max}(X)} \quad \forall X \in \mathrm{P}_+^d, \tag{17}$$

for some constant $C > 0$. Let $\lambda_0 \geq \max\left\{2, \sqrt{\frac{2}{3(d-1)}} 2dC + \frac{2}{3(d-1)}\right\}$, and define a sequence of matrices $\{V_t\}_{t=0}^{\infty} \subset \mathbb{R}^{d\times d}$ as

$$V_0 := \lambda_0 \mathbb{I}_{d\times d}, \quad V_{t+1} := V_t + \omega(V_t) \sum_{i=1}^{d-1} P_{t,i}, \tag{18}$$

where

$$P_{t,i} := a_{t+1,i}^+(a_{t+1,i}^+)^{\mathsf{T}} + a_{t+1,i}^-(a_{t+1,i}^-)^{\mathsf{T}}, \quad a_{t+1,i}^{\pm} := \frac{\tilde{a}_{t+1,i}^{\pm}}{\|\tilde{a}_{t+1,i}^{\pm}\|_2}, \quad \tilde{a}_{t+1,i}^{\pm} := c_t \pm \frac{1}{\sqrt{\lambda_{t,1}}} v_{t,i}, \tag{19}$$

with $\lambda_{t,i} = \lambda_i(V_t)$ the eigenvalues of V_t with corresponding normalized eigenvectors $v_{t,1}, ..., v_{t,d} \in \mathbb{S}^d$. Then we have

$$\lambda_{\min}(V_t) \geq \sqrt{\frac{2}{3(d-1)} \lambda_{\max}(V_t)} \quad \text{for all} \quad t \geq 0. \tag{20}$$

Proof sketch. The idea to prove this result is to use induction and analyse separately the cases when $\lambda_{t,1} < \lambda_{t,d}$ and $\lambda_{t,1} \approx \lambda_{t,d}$ for each $t \geq 1$. The first case is the most delicate and we must quantify the minimum (maximum) non-trivial perturbation that the terms $P_{t,i}$ can contribute to the smallest (biggest) eigenvalue when adding to the matrix V_t. For the second case, the idea is that if $\lambda_{t,1} \approx \lambda_{t,d}$ then if the inequality (20) holds at time step t, the term $\sum_{i=1}^{d-1} P_{t,i}$ is not able to break the relation even if it does not contribute to increase the minimum eigenvalue.

Case 1: $\lambda_{t,1} < \lambda_{t,d}$. For this case, we want to establish a matrix inequality between V_{t+1} and V_t that is independent of the sequence $\{c_t\}_{t=0}^{\infty}$. A simple calculation shows $\|\tilde{a}_{t+1,i}^{\pm}\|_2 \leq 1 + 1/\sqrt{\lambda_{t,1}}$ and using that $\tilde{a}_{t+1,i}^+(\tilde{a}_{t+1,i}^+)^{\mathsf{T}} + \tilde{a}_{t+1,i}^-(\tilde{a}_{t+1,i}^-)^{\mathsf{T}}$ does not depend on the cross terms $c_t v_{t,i}^{\mathsf{T}}$ we can show

$$P_{t,i} \geq \frac{2}{(1 + \sqrt{\lambda_{t,1}})^2} v_{t,i} v_{t,i}^{\mathsf{T}}, \tag{21}$$

which quantifies how much $P_{t,i}$ contributes to increase the eigenvalue on the direction $v_{t,i}$. This immediately leads to

$$V_{t+1} \geq V_t + p_t \sum_{i=1}^{d-1} v_{t,i} v_{t,i}^{\mathsf{T}}, \quad p_t := \frac{2\omega(V_t)}{(1 + \sqrt{\lambda_{t,1}})^2}, \tag{22}$$

which is the matrix inequality that we will use to relate the eigenvalues of V_{t+1} and V_t. In our proof we distinguish the cases $\lambda_{t,d-1} + p_t \leq \lambda_{t,d}$ and $\lambda_{t,d-1} + p_t < \lambda_{t,d} < \lambda_{t,1} + p_t$. However, since both cases use the same technique (with some subtleties) we focus on the first case here and leave the details of the second in the Appendix B. Using the inequality (22) in combination with the minimax principle for eigenvalues in (Bhatia, 1997, Corollary III.1.2) we have

$$\lambda_{t+1,i} \geq \lambda_{t,i} + p_t \quad \text{for} \quad i = 1, ..., d-1. \tag{23}$$

Then we can control how much the maximum eigenvalue grows using the above bound and $\text{Tr}(V_{t+1}) = \lambda_{t,d} + \sum_{i=1}^{d-1} \lambda_{t,i} + 2(d-1)\omega(V_t)$. This allows us to establish that

$$\lambda_{t+1,d} \leq \lambda_{t,d} + (2\omega(V_t) - p_t)(d-1). \tag{24}$$

We finalize this case showing that

$$(\lambda_{t,1} + p_t)^2 \geq \frac{2}{3(d-1)}(\lambda_{t,d} + (2\omega(V_t) - p_t)(d-1)), \tag{25}$$

under the assumption that induction relation (20) holds at time step t. Then the inequality at time step $t+1$ $\lambda_{t+1,1} \geq \sqrt{2/(3(d-1)\lambda_{t+1,d}}$ follows from (23) and (24).

  Case 2: $\lambda_{t,1} \approx \lambda_{t,d}$. For this case, we do not need a tight control on the eigenvalues of V_{t+1}. It suffices to use the following bounds

$$\lambda_{t+1,1} \geq \lambda_{t,1}, \tag{26}$$
$$\lambda_{t+1,d} \leq \lambda_{t,d} + 2\omega(V_t)(d-1), \tag{27}$$

which follows from the inequalities $\lambda_{\min}(A + B) \geq \lambda_{\min}(A) + \lambda_{\min}(B)$ and $\lambda_{\max}(A + B) \leq \lambda_{\max}(A) + \lambda_{\max}(B)$ for $A, B \in \mathrm{P}_+^d$. Then combining the induction hypothesis at time step t, (20) the condition $\omega(V_t) \leq C\sqrt{\lambda_{t,d}}$ and $\lambda_{t,1} \geq \lambda_0$, we can establish a series of inequalities that lead to

$$\lambda_{t+1,1} \geq \lambda_{t,1} \geq \sqrt{\frac{2}{3(d-1)}(2(d-1)\omega(V_t) + \lambda_{t,d})} \geq \sqrt{\frac{2}{3(d-1)}\lambda_{t+1,d}}, \tag{28}$$

where the first and last inequality follows from the trivial perturbation bounds (26). This concludes the induction. ∎

  Note. In the Appendix B.2 we provide an alternative proof for the special case of $d = 2$ that slightly improves the constant in the relation $\lambda_{\min}(V_t) = \Omega(\sqrt{\lambda_{\max}(V_t)})$. In this proof, we compute the worst-case scenario depending on the centers c_t and we provide an exact computation of the eigenvalues of the matrix V_t.

6. Regret analysis

In this Section, we present the regret analysis of LinUCB-VN for a linear bandit with linear vanishing noise. We use the result of the previous section to show a polylogarithmic scaling of the regret with high probability. We develop a new technique to analyze an algorithm based on upper confidence bounds and we rely on the relation that our algorithm keeps the relation $\lambda_{\min}(V_t) = \Omega(\sqrt{\lambda_{\max}(V_t)})$.

Theorem 4 *Let $d \geq 2$, $\delta \in (0,1)$ and $T = 2(d-1)\widetilde{T}$ for some $\widetilde{T} \in \mathbb{N}$. Let $\omega(X)$ defined as in* (16) *using δ and λ_0 satisfy the constraints in Theorem* 3. *Then if we apply* LinUCB-VN *1*(λ_0, ω) *to a d dimensional stochastic linear bandit with linear vanishing noise* (4) *with probability at least $1 - \delta$ the regret satisfies*

$$\text{Regret}(T) \leq 4(d-1) + \left(144d^2\beta_{T,\max}^2 + 24(d-1)^{\frac{3}{2}}\beta_{T,\max}\right)\log\left(\frac{T}{2(d-1)}\right), \quad (29)$$

where

$$\beta_{T,\max} := \left(\lambda_0 + \sqrt{2\log\frac{T}{\delta} + d\log\left(\frac{1}{144\lambda_0}T^2 + \frac{1}{6\sqrt{\lambda_0}}T + 1\right)}\right)^2. \quad (30)$$

From the above Theorem we have that the scaling of the regret on d and T is

$$\text{Regret}(T) = O(d^4\log^3(T)), \quad (31)$$

and in Corollary 11 we prove that choosing $\delta = \frac{1}{T}$ gives

$$\mathbb{E}[\text{Regret}(T)] = O(d^4\log^3(T)). \quad (32)$$

Proof sketch. Using that the algorithm works in $\widetilde{T}$ batches of $2(d-1)$ actions we can express the regret as

$$\text{Regret}(T) = \frac{1}{2}\sum_{t=1}^{\widetilde{T}}\sum_{i=1}^{d-1}\left(\|\theta - a_{t,i}^+\|_2^2 + \|\theta - a_{t,i}^-\|_2^2\right). \quad (33)$$

Thus, to upper bound the regret, we have to give a high probability bound on $\|\theta - a_{t,i}^{\pm}\|_2^2$ for each batch $t \in [\widetilde{T}]$. For that, we can use triangle inequality to obtain

$$\|\theta - a_{t,i}^{\pm}\|_2 \leq \|\theta - \widetilde{\theta}_{t-1}^{\text{w}}\|_2 + \|\theta_{t-1}^{\text{w}} - \widetilde{\theta}_{t-1}^{\text{w}}\|_2 + \|\theta_{t-1}^{\text{w}} - a_{t,i}^{\pm}\|_2. \quad (34)$$

Under the assumption that $\theta \in \mathcal{C}_{t-1}$ with $\mathcal{C}_{t-1}$ defined as in Lemma 2 we can use that $\mathcal{C}_{t-1} \subseteq \mathbb{B}_r^d(\widetilde{\theta}_{t-1}^{\text{w}})$ with $r = \sqrt{\frac{\beta_{t-1,\delta}}{\lambda_{\min}(V_t)}}$ and upper bound the three terms by r. The bound on $\|\theta - \widetilde{\theta}_{t-1}^{\text{w}}\|_2$ follows from the definition of $\mathcal{C}_{t-1}$ and the remaining terms we provide geometrical proofs in the Appendix C. Now the main idea is to use our previous result Theorem 3 (under a careful check of the assumptions and fixing λ_0) and use that our particular choice of actions allows us to keep the inequality $\lambda_{\min}(V_t) \geq \sqrt{\frac{2}{3(d-1)}}\lambda_{\max}(V_t)$. Then we can upper bound $\|\theta - a_{t,i}^{\pm}\|_2^2$ as

$$\|\theta - a_{t,i}^{\pm}\|_2^2 \leq \frac{9\beta_{t-1,\delta}}{\lambda_{\min}(V_{t-1})} \leq \frac{12\sqrt{d-1}\beta_{t-1,\delta}}{\sqrt{\lambda_{\max}(V_{t-1})}}. \quad (35)$$

Thus, if we can control the growth of $\lambda_{\max}(V_t)$ then we can bound the regret under the assumption that $\theta \in \mathcal{C}_t$ for all $t \in [\widetilde{T}]$. In the Appendix C, we rigorously prove that in fact, we can apply Lemma 2 and bound the probability that the following event holds

$$E_t := \{\left(r_{s,1}^+, a_{s,1}^+, r_{s,1}^-, a_{s,1}^-, ..., r_{s,d-1}^+, a_{s,d-1}^+, r_{s,d-1}^-, a_{s,d-1}^-\right)\right)_{s=1}^t : \forall s \in [t], \theta \in \mathcal{C}_s\}. \quad (36)$$

To apply Lemma 2 we need that the event G_t (10) holds which means essentially that $\hat{\sigma}_t^2$ (16) is a good estimator of the subgaussian parameter of the noise. Using the definition of the subgaussian parameter we have $\sigma_t^2(a_{t,i}^{\pm}) \leq \|\theta - a_{t,i}^{\pm}\|_2^2$. and from the above argument we see that we can define a "good" estimator of the noise $\hat{\sigma}_t^2(a_{t,i}^{\pm})$ in the sense that

$$\sigma_t^2(a_{t,i}^{\pm}) \leq \hat{\sigma}_t^2(a_{t,i}^{\pm}) = \frac{1}{w(V_{t-1})} \quad \text{if} \quad \theta \in \mathcal{C}_{t-1}. \tag{37}$$

Thus, after bounding the probabilities that the events G_t and E_t hold what remains to do is to provide bounds for $\lambda_{\max}(V_t)$. Our result will show $\lambda_{\max}(V_t) = \Theta(t^2)$ and we provide a careful computation in the Appendix C. To do that, we just need to use the rule update of V_t and our computation will be independent of the algorithm. Here we provide the idea for a lower bound on $\lambda_{\max}(V_t)$ that is essentially what we need to upper bound the regret. A similar idea works for an upper bound and we provide the full computation in the Appendix C. The upper bound is necessary to prove $\beta_t \leq \beta_{T,\max}$. Let's provide the main ideas for the lower bound.

From the definition of V_t and the choice of $\omega(V_t)$ (16) we have

$$\text{Tr}(V_t) \geq \sum_{s=2}^t 2(d-1)\omega(V_{s-1}) \geq \frac{\sqrt{d-1}}{6\beta_{\tilde{T},\delta}} \sum_{s=1}^{t-1} \sqrt{\lambda_{\max}(V_s)}. \tag{38}$$

Then we can use $\text{Tr}(V_s) \leq d\lambda_{\max}(V_s)$ and some algebra to establish the following inequality

$$\lambda_{\max}(V_t) \geq \frac{1}{1 + 6\frac{d}{\sqrt{d-1}}\beta_{\tilde{T},\delta}} \sum_{s=1}^t \sqrt{\lambda_{\max}(V_s)}. \tag{39}$$

This inequality already gives the intuition that $\lambda_{\max}(V_t) = \Omega(t^2)$ because if we propose $\lambda_{\max}(V_t) \sim t^\alpha$ and we substitute in the above inequality, a comparison of exponents shows that $\alpha \geq 1 + \alpha/2$. Solving this gives $\alpha \geq 2$. However, we have to formalize this idea. To do that we are going to extend the function $\lambda_{\max}(V_t)$ to the continuous on t and this will allow us to establish a differential inequality that we can exactly solve. First, we define the following linear interpolation

$$g_2(x) := (t+1-x)\lambda_{\max}(V_t) + (x-t)\lambda_{\max}(V_{t+1}) \quad \text{for} \quad x \in [t, t+1) \quad \text{and} \quad t \in \{0, ..., \tilde{T}\}, \tag{40}$$

which satisfies $g_2(t) = \lambda_{\max}(V_t)$. Combining the above definition with the inequality (39) and some algebra we can show that

$$g_2(x) \geq \frac{1}{1 + 6\frac{d}{\sqrt{d-1}}\beta_{\tilde{T},\delta}} \int_0^x \sqrt{g_2(s)}ds. \tag{41}$$

This leads to the following differential inequality

$$\frac{dG_2(x)}{dx} \geq \sqrt{\frac{1}{1 + 6\frac{d}{\sqrt{d-1}}\beta_{\tilde{T},\delta}}} \sqrt{G_2(x)}. \tag{42}$$

11

where $\frac{dG_2(x)}{dx} = \sqrt{g_2(x)}$. Using a result of differential equations (more details in the Appendix C) we can solve just the equality and show that

$$G_2(x) \geq \frac{1}{4 + 24\frac{d}{\sqrt{d-1}}\beta_{\tilde{T},\delta}}x^2. \tag{43}$$

This leads to the result $\lambda_{\max}(V_t) = \Omega(t^2)$. Thus, from the expression of the regret, $\beta_{t,\delta} = O(\log(T))$ and (35) we have

$$\text{Regret}(T) = O\left(\text{polylog}(T)\sum_{t=1}^{\tilde{T}}\frac{1}{\sqrt{\lambda_{\max}(V_t)}}\right) = O\left(\text{polylog}(T)\sum_{t=1}^{\tilde{T}}\frac{1}{t}\right) = O\left(\text{polylog}(T)\right). \tag{44}$$

■

Note. Setting $\omega(V_t) = 1$ allows the algorithm to deal with the usual 1-subgaussian noise. For this case, our proof simplifies and leads to the usual $\text{Regret}(T) = \widetilde{O}(\sqrt{T})$.

7. Open problems

- Our model can be extended in different ways, but we felt a focus on a relatively simple geometry would be more suitable to introduce new techniques. A generalisation to Locally Constant Hessian (LCH) surfaces and locally convex action sets Banerjee et al. (2023) would require us to replace the normalization of the actions by a projection to the corresponding surface. We expect that this would still yield a relation $\lambda_{\min} = \Omega(\lambda_{\max}^s)$ where $s \in (0, 1/2]$ depends on how well the action set approximates a LCH, and from there the analysis would be similar. A relatively straight-forward generalisation of our analysis also works for noise decaying as $\|\theta - a_t\|_2^{2\alpha}$ for $\alpha \in [0, 1]$. This eventually leads to a regret scaling as $T^\beta \text{polylog}(T)$ with $\beta = \frac{1-\alpha}{2-\alpha}$. This behaviour interpolates between the $\sqrt{T}$ seen in the constant noise case and polylog(T) seen in our work.

- In Banerjee et al. (2023) the authors prove (under some mild assumptions) that any strategy that minimizes the regret for linear bandits with continuous smooth action sets must satisfy $\lambda_{\min}(V_t) = \Omega(\sqrt{t})$ for constant noise models. This result does not apply to our model since the noise is not constant. Our strategy achieves the relation $\lambda_{\min}(V_t) = \Omega(\sqrt{\lambda_{\max}(V_t)})$ that is a more general condition independent of the noise. Thus, we propose the following conjecture:

  Conjecture 5 (informal) *Consider a linear stochastic bandit with $\mathcal{A} \subset \mathbb{R}^d$ a smooth continuous action set and reward model with arbitrary bounded noise. Then, any strategy that minimizes the regret must achieve the relation $\lambda_{\min}(V_t) = \Omega(\sqrt{\lambda_{\max}(V_t)})$.*

- How can we extend the linear stochastic bandit techniques for minimax lower bound of the regret to the non-constant noise setting? Our model misses a matching minimax lower bound.

- Our strategy plays batches of $2(d-1)$ actions. This is done to simplify the calculations of our main Theorems. However, a technique that adapts at each time step could potentially reduce the dimensional dependence of the regret.

Acknowledgments

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

Marc Abeille and Alessandro Lazaric. Linear Thompson Sampling Revisited. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54, pages 176–184. PMLR, 2017.

Marc Abeille, Louis Faury, and Clément Calauzènes. Instance-wise minimax-optimal algorithms for logistic bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 3691–3699. PMLR, 2021.

Debangshu Banerjee, Avishek Ghosh, Sayak Ray Chowdhury, and Aditya Gopalan. Exploration in linear bandits with rich action sets and its implications for inference. In *International Conference on Artificial Intelligence and Statistics*, pages 8233–8262. PMLR, 2023.

Rajendra Bhatia. *Matrix Analysis*, volume 169. Springer, 1997.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. In *Proceedings of the 21st Conference on Learning Theory*, volume 2, page 3, 2008.

Johannes Kirschner and Andreas Krause. Information directed sampling and bandits with heteroscedastic noise. In *Conference On Learning Theory*, pages 358–384. PMLR, 2018.

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

Josep Lumbreras, Erkka Haapasalo, and Marco Tomamichel. Multi-armed quantum bandits: Exploration versus exploitation when learning properties of quantum states. *Quantum*, 6:749, 2022.

Michel Petrovitch. Sur une manière d'étendre le théorème de la moyenne aux équations différentielles du premier ordre. *Mathematische Annalen*, 54(3):417–436, 1901.

Paat Rusmevichientong and John N Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

Ohad Shamir. On the complexity of bandit linear optimization. In *Conference on Learning Theory*, pages 1523–1551. PMLR, 2015.

Appendix A. Proofs of Section 3

The proof of the confidence region for the regularized least squares estimator relies on the following result about supermatingales.

Theorem 6 (Theorem 3.9 in Lattimore and Szepesvári (2020)) *Let $(X_t)_{t=0}^\infty$ be a supermartingale with $X_t \geq 0$ almost surely for all t. Then for any $\epsilon > 0$,*

$$\Pr\left(\sup_t X_t \geq \epsilon\right) \leq \frac{\mathbb{E}\left[X_0\right]}{\epsilon}. \tag{45}$$

A.1. Proof of Lemma 2

The proof below is an adaptation from the one presented in Lattimore and Szepesvári (2020)[Chapter 20] to our setting. We refer to the reference for detailed computations.

Proof To simplify notation we use $\hat{\sigma}_s^2 = \hat{\sigma}_s^2(a_s)$. From now we will condition all our calculations on the events G_t since we want to prove (13).

First let $\widetilde{S}_t = \sum_{s=1}^t \frac{1}{\hat{\sigma}_s^2} \epsilon_s a_s$ and define the following process

$$\widetilde{M}_t(x) := \exp\left(\langle x, \widetilde{S}_t \rangle - \frac{1}{2}\|x\|_{\widetilde{V}_t}^2\right), \tag{46}$$

for all $x \in \mathbb{R}^d$ and $V_t = V_t(0)$. We want to check that $\widetilde{M}_t(x)$ is a supermartingale, i.e $\mathbb{E}\left[\widetilde{M}_t(x)|\mathcal{F}_{t-1}\right] \leq \widetilde{M}_{t-1}(x)$. From a direct calculation, we have

$$\mathbb{E}\left[\widetilde{M}_t(x)|\mathcal{F}_{t-1}\right] = \widetilde{M}_{t-1}(x)\mathbb{E}\left[\exp\left(\frac{\epsilon_t}{\hat{\sigma}_t}\left\langle x, \frac{a_t}{\hat{\sigma}_t}\right\rangle - \frac{1}{2}\|x\|_{\frac{1}{\hat{\sigma}_t^2}a_t a_t^\top}\right)|\mathcal{F}_{t-1}\right]. \tag{47}$$

Then, using that in the definition of $\hat{\sigma}_t^2$ (6) is defined only using the information up to time step $t-1$, the subgaussian property, and that the event G_t holds we have

$$\mathbb{E}\left[\exp\left(\frac{\epsilon_t}{\hat{\sigma}_t}\left\langle x, \frac{a_t}{\hat{\sigma}_t}\right\rangle\right)|\mathcal{F}_{t-1}\right] \leq \exp\left(\frac{1}{2}\left\langle x, \frac{a_t}{\hat{\sigma}_t}\right\rangle^2\right) = \exp\left(\frac{1}{2}\|x\|_{\frac{1}{\hat{\sigma}_t^2}a_t a_t^\top}\right). \tag{48}$$

Inserting the above expression into (47) we immediately get $\mathbb{E}\left[\widetilde{M}_t(x)|\mathcal{F}_{t-1}\right] \leq \widetilde{M}_{t-1}(x)$. Using Lemma 20.3 in Lattimore and Szepesvári (2020)[Chapter 20] we have that

$$\bar{M}_t := \int_{\mathbb{R}^d} \widetilde{M}_t(x)dh(x) \tag{49}$$

is a supermartingale where h is a probability measure on $\mathbb{R}^d$. In particular we choose $h = \mathcal{N}(0, H^{-1})$ with $H = \lambda \mathbb{I}_{d\times d} \in \mathbb{R}^{d\times d}$ and we get

$$\bar{M}_t = \frac{1}{\sqrt{(2\pi)^d \det(H^{-1})}}\int_{\mathbb{R}^d} \exp\left(\langle x, \widetilde{S}_t \rangle - \frac{1}{2}\|x\|_{V_t}^2 - \frac{1}{2}\|x\|_H\right)dx. \tag{50}$$

The above quantity can be exactly computed as

$$\bar{M}_t = \left(\frac{\det(V_0)}{\det(V_t(\lambda))} \right)^{\frac{1}{2}} \exp \left(\frac{1}{2} \|\widetilde{S}_t\|^2_{V_t^{-1}(\lambda)} \right), \tag{51}$$

where we used the solution for a Gaussian integral, we have completed the square inside the exponential term and we used that $V_t(\lambda) = H + V_t$. Then we can use Theorem 6 in order to get

$$\Pr \left(\sup_t \log \left(\bar{M}_t \right) \geq \log \left(\frac{1}{\delta} \right) \right) = \Pr \left(\sup_t \bar{M}_t \geq \frac{1}{\delta} \right) \leq \delta, \tag{52}$$

where we used that by definition of $\widetilde{M}_{t-1}(x)$ we have $\bar{M}_t \geq 0$ almost surely and $\bar{M}_0 = 1$. Inserting (51) into the above equation we have

$$\Pr \left(\sup_t \|\widetilde{S}_t\|^2_{V_t^{-1}(\lambda)} \geq 2 \log \left(\frac{1}{\delta} \right) + \log \frac{\det(V_t(\lambda))}{\det(V_0(\lambda))} \right) \leq \delta. \tag{53}$$

Finally using the expression for the weighted least squares estimator (8) we have

$$\|\tilde{\theta}_t^{\text{wls}} - \theta\|_{V_t(\lambda)} \leq \|\widetilde{S}_t\|_{V_t^{-1}(\lambda)} + \sqrt{\lambda}, \tag{54}$$

where we used triangle inequality and $\|\theta\|_2^2 = 1$. And the result follows by combining the above expression with (53) and conditioning under the event G_t. ∎

Appendix B. Proofs of Section 5

We will need the following properties for positive semidefinite matrices $A, B \in \mathrm{P}_+^d$:

$$\lambda_{\min}(A + B) \geq \lambda_{\min}(A) + \lambda_{\min}(B), \tag{55}$$
$$\lambda_{\max}(A + B) \leq \lambda_{\max}(A) + \lambda_{\max}(B).$$

And the following mini-max characterization of eigenvalues for Hermitian matrices.

Corollary 7 (Corollary III.1.2 in Bhatia (1997)) *Let $A \in \mathbb{C}^{d \times d}$ be a Hermitian matrix, then*

$$\lambda_k(A) = \max_{\substack{\mathcal{M} \subset \mathbb{C}^d \\ \dim(\mathcal{M}) = d-k+1}} \min_{\substack{x \in \mathcal{M} \\ \|x\|_2 = 1}} \langle x, Ax \rangle = \min_{\substack{\mathcal{M} \subset \mathbb{C}^d \\ \dim(\mathcal{M}) = k}} \max_{\substack{x \in \mathcal{M} \\ \|x\|_2 = 1}} \langle x, Ax \rangle. \tag{56}$$

In particular, if $A \geq B$ then $\lambda_k(A) \geq \lambda_k(B)$.

B.1. Proof of Theorem 3

Proof We start giving an upper bound of the Euclidean norm of $\tilde{a}_{t+1,i}^{\pm}$ (19) with the following calculation

$$\|\tilde{a}_{t+1,i}^{\pm}\|_2^2 = 1 \pm \frac{2}{\sqrt{\lambda_{1,t}}} \langle c_t, v_{t,i} \rangle + \frac{1}{\lambda_{t,1}} \tag{57}$$

$$\leq 1 + \frac{2}{\sqrt{\lambda_{t,1}}} + \frac{1}{\lambda_{t,1}} = \left(1 + \frac{1}{\sqrt{\lambda_{t,1}}} \right)^2, \tag{58}$$

where we used that $c_t, v_{t,i} \in \mathbb{S}^d$. Thus using the definition of $P_{t,i}$ (19),

$$P_{t,i} \geq \left(1 + \frac{1}{\sqrt{\lambda_{t,1}}}\right)^{-2} \left(\tilde{a}_{t+1,i}^+(\tilde{a}_{t+1,i}^+)^\mathsf{T} + \tilde{a}_{t+1,i}^-(\tilde{a}_{t+1,i}^-)^\mathsf{T}\right). \tag{59}$$

From the definition of $\tilde{a}_{t+1,i}^\pm$ (19) we have

$$\tilde{a}_{t+1,i}^\pm(\tilde{a}_{t+1,i}^\pm)^\mathsf{T} = c_t c_t^\mathsf{T} + \frac{1}{\lambda_{t,1}} v_{t,i} v_{t,i}^\mathsf{T} \pm \frac{1}{\sqrt{\lambda_{t,1}}} \left(c_t v_{t,i}^\mathsf{T} + v_{t,i} c_t^\mathsf{T}\right). \tag{60}$$

The above allow us to bound $P_{t,i}$ as

$$P_{t,i} \geq 2 \left(1 + \frac{1}{\sqrt{\lambda_{t,1}}}\right)^{-2} \left(c_t c_t^\mathsf{T} + \frac{1}{\lambda_{t,1}} v_{t,i} v_{t,i}^\mathsf{T}\right) \tag{61}$$

$$\geq \frac{2}{(1 + \sqrt{\lambda_{t,1}})^2} v_{t,i} v_{t,i}^\mathsf{T}, \tag{62}$$

where we used $c_t c_t^\mathsf{T} \geq 0$ and $\frac{1}{\lambda_{t,1}}\left(1 + \frac{1}{\sqrt{\lambda_{t,1}}}\right)^{-2} = (1 + \sqrt{\lambda_{t,i}})^{-2}$. Thus, from the abound bound and the definition of V_t (18) we obtain the following matrix inequality

$$V_{t+1} \geq V_t + \frac{2w(V_t)}{(1 + \sqrt{\lambda_{t,1}})^2} \sum_{i=1}^{d-1} v_{t,i} v_{t,i}^\mathsf{T}. \tag{63}$$

We want to prove the result using induction. The case $t = 0$ is immediately satisfied since $\lambda_{0,d} = \lambda_{0,1} \geq \frac{2}{3(d-1)}$. Now we will assume that

$$\lambda_{t,1} \geq \sqrt{\frac{2}{3(d-1)}\lambda_{t,d}}, \tag{64}$$

is satisfied and we want to prove the same inequality for $t + 1$. We distinguish cases depending on the growth of the maximum eigenvalue of V_t.

Case 1: $\lambda_{t,d} \geq \lambda_{t,d-1} + \frac{2w(V_t)}{(1+\sqrt{\lambda_{t,1}})^2}$

Using the hypothesis $\lambda_{t,d} \geq \lambda_{t,d-1} + \frac{2w(V_t)}{(1+\sqrt{\lambda_{t,1}})^2}$, the fact that V_t diagonalizes in the $\{v_{t,i}\}_{i=1}^d$ basis and the ordering $\lambda_{t,1} \leq .... \leq \lambda_{t,d-1} \leq \lambda_{t,d}$ we have

$$\lambda_i \left(V_t + \frac{2w(V_t)}{(1 + \sqrt{\lambda_{t,1}})^2} \sum_{i=1}^{d-1} v_{t,i} v_{t,i}^\mathsf{T}\right) = \lambda_{t,i} + \frac{2w(V_t)}{(1 + \sqrt{\lambda_{t,1}})^2} \quad \text{for} \quad i = 1, ..., d-1 \tag{65}$$

$$\lambda_d \left(V_t + \frac{2w(V_t)}{(1 + \sqrt{\lambda_{t,1}})^2} \sum_{i=1}^{d-1} v_{t,i} v_{t,i}^\mathsf{T}\right) = \lambda_{t,d}. \tag{66}$$

16

Then combining with the mini-max principle for eigenvalues 7 and using that both sides of (63) are real and symmetric, we arrive at

$$\lambda_{i,t+1} \geq \lambda_{i,t} + \frac{2w(V_t)}{(1 + \sqrt{\lambda_{t,1}})^2} \quad \text{for} \quad i = 1, ..., d-1. \tag{67}$$

From the expression for V_{t+1} (18), we deduce that

$$\mathrm{Tr}(V_{t+1}) = \left(\lambda_{t+1,d} + \sum_{i=1}^{d-1} \lambda_{t+1,i}\right) = \lambda_{t,d} + \sum_{i=1}^{d-1} \lambda_{t,i} + 2(d-1)w(V_t). \tag{68}$$

Also from (67)

$$\mathrm{Tr}(V_{t+1}) \geq \lambda_{t+1,d} + \sum_{i=1}^{d-1} \lambda_{t,i} + \frac{2(d-1)w(V_t)}{(1 + \sqrt{\lambda_{t,1}})^2}. \tag{69}$$

Combining the above we can bound the maximum eigenvalue as

$$\lambda_{t+1,d} \leq \lambda_{t,d} + \sum_{i=1}^{d-1} \lambda_{t,i} + 2(d-1)w(V_t) - \left(\sum_{i=1}^{d-1} \lambda_{t,i} + \frac{2(d-1)w(V_t)}{(1 + \sqrt{\lambda_{t,1}})^2}\right) \tag{70}$$

$$= \lambda_{d,t} + 2(d-1)w(V_t)\frac{\lambda_{t,1} + 2\sqrt{\lambda_{t,1}}}{(1 + \sqrt{\lambda_{t,1}})^2}. \tag{71}$$

Recall that we want to check

$$\lambda_{t+1,1} \geq \sqrt{\frac{2}{3(d-1)}\lambda_{t+1,d}}. \tag{72}$$

Using (67) and (70) we can square the above and see that it suffices to check

$$\left(\lambda_{1,t} + \frac{2w(V_t)}{(1 + \sqrt{\lambda_{t,1}})^2}\right)^2 \geq \frac{2}{3(d-1)}\lambda_{d,t} + \frac{4w(V_t)(\lambda_{t,1} + 2\sqrt{\lambda_{t,1}})}{3(1 + \sqrt{\lambda_{t,1}})^2}. \tag{73}$$

Multiplying out the terms, this is equivalent to the condition

$$\underbrace{\lambda_{1,t}^2 - \frac{2}{3(d-1)}\lambda_{d,t}}_{(i)} + \underbrace{\frac{4w^2(V_t)}{(1 + \sqrt{\lambda_{t,1}})^4}}_{(ii)} + \underbrace{\frac{4w(V_t)}{(1 + \sqrt{\lambda_{t,1}})^2}(\lambda_{t,1} - \frac{1}{3}(\lambda_{t,1} + 2\sqrt{\lambda_{t,1}}))}_{(iii)} \geq 0. \tag{74}$$

It remains to observe that (i) is positive by induction at time step t (cf. (64)), (ii) is always positive and (iii) is positive for $\lambda_{t,1} \geq 1$ and this is true since $\lambda_{t,1} \geq \lambda_0 \geq 2$, $\lambda_{t,1}$ is non-decreasing in t and $f(x) = x - \sqrt{x}$ is positive for $x \geq 1$.

Case 2.1: $\lambda_{t,1} + \frac{2w(V_t)}{(1+\sqrt{\lambda_{t,1}})^2} < \lambda_{t,d} < \lambda_{t,d-1} + \frac{2w(V_t)}{(1+\sqrt{\lambda_{t,1}})^2}$

First we find $k \in \mathbb{N}$, $1 < k \leq d-1$ such that

$$\lambda_{t,k-1} + \frac{2w(V_t)}{(1+\sqrt{\lambda_{t,1}})^2} \leq \lambda_{t,d} \leq \lambda_{t,k} + \frac{2w(V_t)}{(1+\sqrt{\lambda_{t,1}})^2}. \tag{75}$$

Then using that $\{v_{t,i}\}_{i=1}^d$ are the eigenvectors of V_t we have

$$\lambda_i\left(V_t + \frac{2w(V_t)}{(1+\sqrt{\lambda_{t,1}})^2}\sum_{i=1}^{d-1}v_{t,i}v_{t,i}^\mathsf{T}\right) = \lambda_{t,i} + \frac{2w(V_t)}{(1+\sqrt{\lambda_{t,1}})^2} \quad \text{for} \quad i = 1, ..., k-1 \tag{76}$$

$$\lambda_k\left(V_t + \frac{2w(V_t)}{(1+\sqrt{\lambda_{t,1}})^2}\sum_{i=1}^{d-1}v_{t,i}v_{t,i}^\mathsf{T}\right) = \lambda_{t,d} \tag{77}$$

$$\lambda_i\left(V_t + \frac{2w(V_t)}{(1+\sqrt{\lambda_{t,1}})^2}\sum_{i=1}^{d-1}v_{t,i}v_{t,i}^\mathsf{T}\right) = \lambda_{t,i-1} + \frac{2w(V_t)}{(1+\sqrt{\lambda_{t,1}})^2} \quad \text{for} \quad i = k+1, ..., d. \tag{78}$$

Again using the mini-max principle for eigenvalues 7 and that both sides of (63) are real and symmetric

$$\lambda_{t+1,i} \geq \lambda_{t,i} + \frac{2w(V_t)}{(1+\sqrt{\lambda_{t,1}})^2} \quad \text{if} \quad i \in \{1, ..., k-1\}, \tag{79}$$

$$\lambda_{t+1,i} \geq \lambda_{t,d} \quad \text{if} \quad i = k, \tag{80}$$

$$\lambda_{t+1,i} \geq \lambda_{t,i-1} + \frac{2w(V_t)}{(1+\sqrt{\lambda_{t,1}})^2} \quad \text{if} \quad i \in \{k+1, ..., d\}. \tag{81}$$

Thus, using the above inequalities we can bound the trace of V_{t+1} as

$$\mathrm{Tr}(V_{t+1}) = \sum_{i=1}^{k-1}\lambda_{t+1,i} + \lambda_{t+1,k} + \sum_{i=k+1}^{d-1}\lambda_{t+1,i} + \lambda_{t+1,d} \tag{82}$$

$$\geq \left(\sum_{i=1}^{k-1}\lambda_{t,i}\right) + \frac{2(k-1)w(V_t)}{(1+\sqrt{\lambda_{t,1}})^2} + \lambda_{t,d} + \left(\sum_{i=k}^{d-2}\lambda_{t,i}\right) + \frac{2(d-k-1)w(V_t)}{(1+\sqrt{\lambda_{t,1}})^2} + \lambda_{t+1,d} \tag{83}$$

$$= \lambda_{t+1,d} + \lambda_{t,d} + \sum_{i=1}^{d-2}\lambda_{t,i} + \frac{2(d-2)w(V_t)}{(1+\sqrt{\lambda_{t,1}})^2} \tag{84}$$

$$\geq \lambda_{t+1,d} + \sum_{i=1}^{d-1}\lambda_{t,i} + \frac{2(d-2)w(V_t)}{(1+\sqrt{\lambda_{t,1}})^2}, \tag{85}$$

where in the last bound we used simply $\lambda_{t,d} \geq \lambda_{t,d-1}$. From the expression of V_{t+1} (18)

$$\mathrm{Tr}(V_{t+1}) = \lambda_{t,d} + \sum_{i=1}^{d-1}\lambda_{t,i} + 2(d-1)w(V_t), \tag{86}$$

18

and combining with the previous bound we obtain

$$\lambda_{t+1,d} \leq \lambda_{d,t} + 2(d-1)w(V_t) - \frac{2(d-2)w(V_t)}{(1+\sqrt{\lambda_{t,1}})^2} \tag{87}$$

$$= \lambda_{d,t} + \frac{2w(V_t)}{(1+\sqrt{\lambda_{t,1}})^2} \left(\lambda_{t,1}(d-1) + 2\sqrt{\lambda_{t,1}}(d-1) + 1 \right) \tag{88}$$

$$\leq \lambda_{d,t} + \frac{2(d-1)w(V_t)}{(1+\sqrt{\lambda_{t,1}})^2} \left(\lambda_{t,1} + 2\sqrt{\lambda_{t,1}} + 1 \right). \tag{89}$$

Again to check

$$\lambda_{t+1,1} \geq \sqrt{\frac{2}{3(d-1)}} \lambda_{t+1,d}, \tag{90}$$

we can square both sides and using the above bounds on $\lambda_{t+1,1}$ and $\lambda_{t+1,d}$ it suffices to check

$$\left(\lambda_{1,t} + \frac{2w(V_t)}{(1+\sqrt{\lambda_{t,1}})^2} \right)^2 \geq \frac{2}{3(d-1)} \lambda_{d,t} + \frac{4w(V_t)(\lambda_{t,1} + 2\sqrt{\lambda_{t,1}} + 1)}{3(1+\sqrt{\lambda_{t,1}})^2}. \tag{91}$$

Multiplying out the terms, this is equivalent to the condition

$$\underbrace{\lambda_{1,t}^2 - \frac{2}{3(d-1)}\lambda_{d,t}}_{(i)} + \underbrace{\frac{4w^2(V_t)}{(1+\sqrt{\lambda_{t,1}})^4}}_{(ii)} + \underbrace{\frac{4w(V_t)}{(1+\sqrt{\lambda_{t,1}})^2}\left(\lambda_{t,1} - \frac{1}{3}(\lambda_{t,1} + 2\sqrt{\lambda_{t,1}} + 1)\right)}_{(iii)} \geq 0. \tag{92}$$

Finally, we observe that (i) is positive by induction at time step t, (ii) is always positive and (iii) is positive since $\lambda_{t,1} \geq 2$, $\lambda_{t,1} \geq 2$ is non-decreasing and $f(x) = 2x - 2\sqrt{x} - 1$ is positive for $x \geq 2$.

Case 2.2: $\lambda_{t,d} \leq \lambda_{t,1} + \frac{2w(V_t)}{(1+\sqrt{\lambda_{t,1}})^2}$

From the statement of the theorem we have

$$\lambda_{t,1} \geq \lambda_0 \geq \sqrt{\frac{2}{3(d-1)}2dC} + \frac{2}{3(d-1)}. \tag{93}$$

Multiplying both sides by $\lambda_{t,1}$ we have

$$\lambda_{t,1}^2 \geq \sqrt{\frac{2}{3(d-1)}2dC}\lambda_{t,1} + \frac{2}{3(d-1)}\lambda_{t,1} \tag{94}$$

$$\geq \frac{2}{3(d-1)}\left(2dC\sqrt{\lambda_{t,d}} + \lambda_{t,1} \right) \tag{95}$$

$$\geq \frac{2}{3(d-1)}\left(2dw(V_t) + \lambda_{t,1} \right) \tag{96}$$

$$= \frac{2}{3(d-1)}\left(2(d-1)w(V_t) + 2w(V_t) + \lambda_{t,1} \right) \tag{97}$$

$$\geq \frac{2}{3(d-1)}\left(2(d-1)w(V_t) + \lambda_{t,d} \right), \tag{98}$$

where the second inequality follows from the induction hypothesis $\lambda_{t,1} \geq \sqrt{\frac{2}{3(d-1)}\lambda_{t,d}}$, the third inequality from $w(V_t) \leq C\sqrt{\lambda_{t,d}}$ and the fourth inequality from the assumption $\lambda_{t,d} \leq \lambda_{t,1} + \frac{2w(V_t)}{(1+\sqrt{\lambda_{t,1}})^2} \leq \lambda_{t,1} + 2w(V_t)$. Thus taking the square root in both sides we have

$$\lambda_{t,1} \geq \sqrt{\frac{2}{3(d-1)}\left(\lambda_{t,d} + 2(d-1)w(V_t)\right)}, \tag{99}$$

and the induction at time step $t+1$, $\lambda_{t+1,1} \geq \sqrt{\frac{2}{3(d-1)}\lambda_{t+1,d}}$ follows from the bounds

$$\lambda_{t+1,1} \geq \lambda_{t,1}, \tag{100}$$
$$\lambda_{t+1,d} \leq \lambda_{t,d} + 2w(V_t)(d-1), \tag{101}$$

where we used the inequalities (55) and the definition of V_t (18). ∎

B.2. Alternative proof for special case $d = 2$

We present an alternative proof of the previous result for the particular case of $d = 2$ where the main difference is that we provide an exact calculation of the eigenvalues of the matrix V_{t+1}. This technique can slightly improve the constant on the lower bound $\lambda_{\min}(V_t) = \Omega(\sqrt{\lambda_{\max}(V_t)})$.

Theorem 8 *Let $\{c_t\}_{t=0}^{\infty} \subset \mathbb{S}^2$ be a sequence of normalized vectors and $\omega : P_+^2 \to \mathbb{R}_{\geq 0}$ a function such that*

$$\omega(X) \leq C\sqrt{\lambda_{\max}(X)}, \tag{102}$$

for a constant $C > 0$ and any $X \in P_+^2$. Let $\lambda_0 \geq 4\sqrt{2}C + 2$, and define a sequence of matrices $\{V_t\}_{t=0}^{\infty} \subset \mathbb{R}^{2\times 2}$ as

$$V_0 := \lambda_0 \mathbb{I}_{2\times 2}, \quad V_{t+1} := V_t + \omega(V_t)\left(a_{t+1}^+(a_{t+1}^+)^{\mathsf{T}} + a_{t+1}^-(a_{t+1}^-)^{\mathsf{T}}\right), \tag{103}$$

where $a_{t+1}^+, a_{t+1}^- \in \mathbb{S}^2$ are defined as

$$a_{t+1}^+ := \frac{c_t + \frac{1}{\sqrt{\lambda_{\min}(V_t)}}v_{t,\min}}{\sqrt{1 + \frac{2\langle c_t, v_{t,\min}\rangle}{\sqrt{\lambda_{\min}(V_t)}} + \frac{1}{\lambda_{\min}(V_t)}}}, \quad a_{t+1}^- := \frac{c_t - \frac{1}{\sqrt{\lambda_{\min}(V_t)}}v_{t,\min}}{\sqrt{1 - \frac{2\langle c_t, v_{t,\min}\rangle}{\sqrt{\lambda_{\min}(V_t)}} + \frac{1}{\lambda_{\min}(V_t)}}}, \tag{104}$$

and $v_{\min,t}$ is the normalized eigenvector corresponding to the minimum eigenvalue . Then we have

$$\lambda_{\min}(V_t) \geq \sqrt{2\lambda_{\max}(V_t)} \quad \text{for all} \quad t \geq 0. \tag{105}$$

Proof To simplify the notation in the proof we define

$$\lambda_{\min,t} := \lambda_{\min}(V_t), \quad \lambda_{\max,t} := \lambda_{\max}(V_t), \tag{106}$$

with corresponding normalized eigenvectors $v_{t,\min}, v_{t,\max} \in \mathbb{S}^2$.

We are going to prove the result by induction. At time step $t = 0$ we have $\lambda_{\min,t} = \lambda_{\max,t} = \lambda_0$ and the inequality $\lambda_0 \geq \sqrt{2\lambda_0}$ holds since $\lambda_0 \geq 2$. Then at time step $t \geq 1$ we can express V_t as

$$V_t = \begin{pmatrix} \lambda_{t,\min} & 0 \\ 0 & \lambda_{t,\max} \end{pmatrix}, \tag{107}$$

using the basis $\{v_{t,\min}, v_{t,\max}\}$. Then we assume that the inequality $\lambda_{t,\min} \geq \sqrt{2\lambda_{\max,t}}$ is satisfied and we want to check the same inequality at time step $t + 1$. We express the vectors that update V_{t+1} (103) on the same basis as

$$a_{t+1}^+ = \begin{pmatrix} \langle a_{t+1}^+, v_{t,\min} \rangle \\ \langle a_{t+1}^+, v_{t,\max} \rangle \end{pmatrix} \quad a_{t+1}^- = \begin{pmatrix} \langle a_{t+1}^-, v_{t,\min} \rangle \\ \langle a_{t+1}^-, v_{t,\max} \rangle \end{pmatrix}, \tag{108}$$

and define the following quantities,

$$x_t := \langle a_{t+1}^+, v_{t,\min} \rangle^2 + \langle a_{t+1}^-, v_{t,\min} \rangle^2, \tag{109}$$

$$y_t := \langle a_{t+1}^+, v_{t,\max} \rangle^2 + \langle a_{t+1}^-, v_{t,\max} \rangle^2, \tag{110}$$

$$z_t := \langle a_{t+1}^+, v_{t,\min} \rangle \langle a_{t+1}^+, v_{t,\max} \rangle + \langle a_{t+1}^-, v_{t,\min} \rangle \langle a_{t+1}^-, v_{t,\max} \rangle. \tag{111}$$

Using that $a_{t+1}^+, a_{t+1}^- \in \mathbb{S}^2$ we get the following relation

$$y_t = 2 - x_t. \tag{112}$$

Thus the perturbation at time step $t + 1$ can be written as

$$a_{t+1}^+(a_{t+1}^+)^\mathsf{T} + a_{t+1}^-(a_{t+1}^-)^\mathsf{T} = \begin{pmatrix} x_t & z_t \\ z_t & y_t \end{pmatrix}, \tag{113}$$

and V_{t+1} as

$$V_{t+1} = \begin{pmatrix} \lambda_{t,\min} + \omega(V_t)x_t & \omega(V_t)z_t \\ \omega(V_t)z_t & \lambda_{t,\max} + \omega(V_t)(2 - x_t) \end{pmatrix}. \tag{114}$$

In order to analyze the eigenvalue of V_{t+1} we have to control the overlap $\langle v_{t,\min}, c_t \rangle$, so we define the following variable

$$\alpha_t := \langle v_{t,\min}, c_t \rangle \in [-1, 1], \tag{115}$$

and using that

$$\langle v_{t,\min}, a_{t+1}^+ \rangle = \frac{\alpha_t + \frac{1}{\sqrt{\lambda_{\min,t}}}}{\sqrt{1 + \frac{2}{\sqrt{\lambda_{\min,t}}}\alpha_t + \frac{1}{\lambda_{\min,t}}}}, \quad \langle v_{t,\min}, a_{t+1}^- \rangle = \frac{\alpha_t - \frac{1}{\sqrt{\lambda_{\min,t}}}}{\sqrt{1 - \frac{2}{\sqrt{\lambda_{\min,t}}}\alpha_t + \frac{1}{\lambda_{\min,t}}}} \tag{116}$$

we can express x_t, z_t in terms of α_t as

$$x_t(\alpha_t) = \frac{\left(\alpha_t + \frac{1}{\sqrt{\lambda_{\min,t}}}\right)^2}{1 + \frac{2}{\sqrt{\lambda_{\min,t}}}\alpha_t + \frac{1}{\lambda_{\min,t}}} + \frac{\left(\alpha_t - \frac{1}{\sqrt{\lambda_{\min,t}}}\right)^2}{1 - \frac{2}{\sqrt{\lambda_{\min,t}}}\alpha_t + \frac{1}{\lambda_{\min,t}}}, \tag{117}$$

$$z_t(\alpha_t) = \left(\frac{\alpha_t + \frac{1}{\sqrt{\lambda_{\min,t}}}}{1 + \frac{2}{\sqrt{\lambda_{\min,t}}}\alpha_t + \frac{1}{\lambda_{\min,t}}} + \frac{\alpha_t - \frac{1}{\sqrt{\lambda_{\min,t}}}}{1 - \frac{2}{\sqrt{\lambda_{\min,t}}}\alpha_t + \frac{1}{\lambda_{\min,t}}}\right)\sqrt{1 - \alpha_t^2}. \tag{118}$$

21

Then we can directly compute the minimum eigenvalue of V_{t+1} as

$$\lambda_{\min,t+1}(\alpha_t) = \frac{\lambda_{\max,t} + \lambda_{\min,t}}{2} + \omega(V_t) - \frac{1}{2}\sqrt{(\lambda_{\max,t} - \lambda_{\min,t} + 2\omega(V_t)(1 - x_t(\alpha_t)))^2 + 4\omega^2(V_t)z_t^2(\alpha_t)}.$$
(119)

Now we define the difference of eigenvalues at time step t,

$$\phi_t := \lambda_{\max,t} - \lambda_{\min,t},$$
(120)

and analyze two different regimes for the induction to hold at time step $t + 1$.

Case 1: $\phi_t \geq 2\omega(V_t)$

For this case we want to justify that $\lambda_{\min,t+1}(\alpha_t) \geq \lambda_{\min,t+1}(0)$ for $\alpha_t \in [-1, 1]$, and later prove the induction using this lower bound. Defining the following

$$f(\alpha_t) := (\phi_t + 2\omega(V_t)(1 - x(\alpha_t)))^2 + 4\omega^2(V_t)z^2(\alpha_t)$$
(121)

$$= \phi_t^2 + 4\omega(V_t)\big(\phi_t(1 - x(\alpha_t)) + \omega(V_t)(1 - x(\alpha_t))^2 + \omega(V_t)z^2(\alpha_t)\big),$$
(122)

and comparing with the exact minimum eigenvalue (119) we see that it suffices to prove that $f(\alpha_t)$ achieves a maximum at $\alpha_t = 0$ in the range $\alpha_t \in [-1, 1]$ in order to have $\lambda_{\min,t+1}(\alpha_t) \geq \lambda_{\min,t+1}(0)$ for all $\alpha_t \in [-1, 1]$. A direct computation shows that

$$f(\alpha_t) = \phi_t^2 + 4\omega(V_t)g(\alpha_t), \quad g(\alpha_t) := \frac{2\big(\lambda_{\min,t} - 1\big)\lambda_{\min,t}\phi_t\alpha_t^2 + (1 - \lambda_{\min,t}^2)\phi_t - (1 - \lambda_{\min,t})^2\omega(V_t)}{4\lambda_{\min,t}\alpha_t^2 - \big(1 + \lambda_{\min,t}\big)^2}.$$
(123)

We have that $g(\alpha_t)$ is of the form $g(x) = \frac{ax^2 + b}{cx^2 + d}$ which has an unique maximum at $x = 0$ if $ad - bc < 0$. Then identifying the coefficients we have to check that

$$p := 2(1 - \lambda_{\min,t})\lambda_{\min,t}(1 + \lambda_{\min,t})^2\phi_t - 4\lambda_{\min,t}(1 - \lambda_{\min,t}^2)\phi_t + 4\lambda_{\min,t}(1 - \lambda_{\min,t})^2\omega(V_t) < 0.$$
(124)

Using that $\lambda_{\min,t} \geq 2$ and $\omega(V_t) \leq \frac{\phi_t}{2}$ we can bound the above as

$$p \leq \phi_t \big(2(1 - \lambda_{\min,t})\lambda_{\min,t}(1 + \lambda_{\min,t})^2 - 4\lambda_{\min,t}(1 - \lambda_{\min,t}^2) + 2\lambda_{\min,t}(1 - \lambda_{\min,t})^2\big). \quad (125)$$

Finally summing all the terms we get,

$$p \leq -2\phi_t \big(\lambda_{\min,t}^2(1 - \lambda_{\min,t})^2\big) \leq 0,$$
(126)

where the inequality follows from $\phi_t \geq 0$. Thus, we conclude that

$$\lambda_{\min,t+1}(\alpha_t) \geq \lambda_{\min,t+1}(0).$$
(127)

Using that

$$x_t(0) = \frac{2}{1 + \lambda_{\min,t}}, \quad z_t(0) = 0,$$
(128)

and $\phi_t \geq 2\omega(V_t)$ we have

$$\lambda_{\min,t+1}(0) = \lambda_{\min,t} + \frac{2\omega(V_t)}{1 + \lambda_{\min,t}}. \tag{129}$$

Thus, using (127)(129) and

$$\text{Tr}(V_{t+1}) = \lambda_{\min,t+1}(\alpha_t) + \lambda_{\max,t+1}(\alpha_t) = \lambda_{\max,t} + \lambda_{\min,t} + 2\omega(V_t), \tag{130}$$

we can upper bound the maximum eigenvalue as

$$\lambda_{\max,t+1}(\alpha_t) = \lambda_{\max,t} + \lambda_{\min,t} - \lambda_{\min,t+1}(\alpha_t) + 2\omega(V_t) \tag{131}$$

$$\leq \lambda_{\max,t} + \lambda_{\min,t} - \lambda_{\min,t+1}(0) + 2\omega(V_t) \tag{132}$$

$$= \lambda_{\max,t} + 2\omega(V_t)\frac{\lambda_{\min,t}}{1 + \lambda_{\min,t}}. \tag{133}$$

Finally in order to check the induction step at $t + 1$, $\lambda_{\min,t+1} \geq \sqrt{2\lambda_{\max,t+1}}$ we can use the bounds (127)(131) and it suffices to check

$$\lambda_{\min,t} + \frac{2\omega(V_t)}{1 + \lambda_{\min,t}} \geq \sqrt{2\left(\lambda_{\max,t} + 2\omega(V_t)\frac{\lambda_{\min,t}}{1 + \lambda_{\min,t}}\right)}. \tag{134}$$

Squaring both sides and rearranging we obtain

$$\underbrace{\lambda_{\min,t}^2 - 2\lambda_{\max,t}}_{(i)} + \underbrace{\frac{4\omega^2(V_t)}{(1 + \lambda_{\min,t})^2}}_{(ii)} \geq 0, \tag{135}$$

where (i) is positive because we assume $\lambda_{\min,t} \geq \sqrt{2\lambda_{\max,t}}$ to hold at time step t and (ii) is always positive. This concludes the case $\phi_t \geq 2\omega(V_t)$ induction.

Case 2: $\phi_t \leq 2\omega(V_t)$

From the definition of λ_0 we have

$$\lambda_{\min,t} \geq \lambda_0 \geq 4\sqrt{2}C + 2. \tag{136}$$

Multiplying both sides by $\lambda_{\min,t}$,

$$\lambda_{\min,t}^2 \geq 4\sqrt{2}C\lambda_{\min,t} + 2\lambda_{\min,t} \tag{137}$$

$$\geq 2\left(4C\sqrt{\lambda_{\max,t}} + \lambda_{\min,t}\right) \tag{138}$$

$$\geq 2\left(\lambda_{\min,t} + 4\omega(V_t)\right) \tag{139}$$

$$\geq 2\left(\lambda_{\min,t} + \phi_t + 2\omega(V_t)\right) \tag{140}$$

$$= 2\left(\lambda_{\max,t} + 2\omega(V_t)\right), \tag{141}$$

23

where in the second inequality we used the induction hypothesis $\lambda_{\min,t} \geq \sqrt{2\lambda_{\max,t}}$, the third inequality $\omega(V_t) \leq C\sqrt{\lambda_{\max,t}}$, the fourth inequality $\phi_t \leq 2\omega(V_t)$ and the last equality the definition $\phi_t = \lambda_{\max,t} - \lambda_{\min,t}$. Thus, we have

$$\lambda_{\min,t} \geq \sqrt{2(\lambda_{\max,t} + 2\omega(V_t))}, \tag{142}$$

and the inequality at time step $t + 1$, $\lambda_{\min,t} \geq \sqrt{2\lambda_{\max,t+1}}$ follows from the bounds

$$\lambda_{\min,t+1} \geq \lambda_{\min,t}, \tag{143}$$

$$\lambda_{\max,t+1} \leq \lambda_{\max,t} + 2\omega(V_t), \tag{144}$$

where we used (55) with $\lambda_{\min}(a_t a_t^{\mathsf{T}}) = \lambda_{\min}(b_t b_t^{\mathsf{T}}) = 0$ and $\lambda_{\max}(a_t a_t^{\mathsf{T}}) = \lambda_{\max}(b_t b_t^{\mathsf{T}}) = 1$. ∎

Appendix C. Proofs of Section 6

In our regret analysis, we need to control the distance between the regularized least squares estimator and the normalized version. The Lemma below is the technical result that we use.

Lemma 9 *Given two normalized vectors $c, v \in \mathbb{S}^d$, a positive constant $\lambda > 1$ and the following vectors*

$$\widetilde{a}^{\pm} = c \pm \frac{1}{\sqrt{\lambda}}v, \quad a^{\pm} = \frac{\widetilde{a}^{\pm}}{\|\widetilde{a}^{\pm}\|_2}. \tag{145}$$

Then we have

$$\|a^{\pm} - c\|_2^2 \leq \frac{2}{\lambda}. \tag{146}$$

Proof We are going to give the proof for a^-. The one for a^+ follows from an identical calculation.

First, we can relate the distance to the inner product of the vectors as

$$\|a^- - c\|_2^2 = \langle a^- - c, a^- - c \rangle = 2 - 2\langle a^-, c \rangle. \tag{147}$$

Using the normalization factor is

$$\|\widetilde{a}^-\|_2^2 = 1 - \frac{2}{\sqrt{\lambda}}\langle c, v \rangle + \frac{1}{\lambda}, \tag{148}$$

then

$$\langle a^-, c \rangle = \frac{1 - \frac{1}{\sqrt{\lambda}}\alpha}{\sqrt{1 + \frac{1}{\lambda} - \frac{2}{\sqrt{\lambda}}\alpha}}, \quad \text{where} \quad \alpha := \langle c, v \rangle \in [-1, 1]. \tag{149}$$

In order to study the behavior of (147) in terms of the overlap $\langle c, v \rangle$ we define

$$f(\alpha, \lambda) := 2\left(1 - \frac{1 - \frac{1}{\sqrt{\lambda}}\alpha}{\sqrt{1 + \frac{1}{\lambda} - \frac{2}{\sqrt{\lambda}}\alpha}}\right) \tag{150}$$

and we are going to check the maximum in the range $\alpha \in [-1, 1]$. Note that $f(\alpha, \lambda) = \|a^- - c\|_2^2$. Taking the derivative respect to α,

$$\frac{\partial f(\alpha, \lambda)}{\partial \alpha} = 2 \frac{\frac{1}{\sqrt{\lambda}}\sqrt{1 + \frac{1}{\lambda} - \frac{2}{\sqrt{\lambda}}\alpha} - \left(1 - \frac{1}{\sqrt{\lambda}}\alpha\right)\frac{1}{\sqrt{\lambda}\sqrt{1 + \frac{1}{\lambda} - \frac{2}{\sqrt{\lambda}}\alpha}}}{1 + \frac{1}{\lambda} - \frac{2}{\sqrt{\lambda}}\alpha} \tag{151}$$

Then we can find the extremal points of $f(\alpha)$ as

$$\frac{\partial f(\alpha, \lambda)}{\partial \alpha} = 0 \Leftrightarrow 1 + \frac{1}{\lambda} - \frac{2}{\sqrt{\lambda}}\alpha = 1 - \frac{1}{\sqrt{\lambda}}\alpha \tag{152}$$

$$\Leftrightarrow \alpha = \frac{1}{\sqrt{\lambda}} \tag{153}$$

Using that $\lambda > 1$ we have the following inequalites

$$\left.\frac{\partial f(\alpha, \lambda)}{\partial \alpha}\right|_{\alpha=1} = -\frac{2}{\lambda\left(1 - \frac{1}{\sqrt{\lambda}}\right)^2} < 0, \quad \left.\frac{\partial f(\alpha, \lambda)}{\partial \alpha}\right|_{\alpha=0} = \frac{2}{\sqrt{\lambda}\sqrt{1 + \frac{1}{\lambda}}}\left(1 - \frac{1}{1 + \frac{1}{\lambda}}\right) > 0. \tag{154}$$

Thus, $\alpha = \frac{1}{\sqrt{\lambda}}$ (150) is a maximum and it is achieved at

$$f\left(\frac{1}{\sqrt{\lambda}}, \lambda\right) = 2\left(1 - \sqrt{1 - \frac{1}{\lambda}}\right). \tag{155}$$

Using the original definition of $f(\alpha, \lambda)$ we can conclude that

$$\|a^- - c\|_2^2 \leq 2\left(1 - \sqrt{1 - \frac{1}{\lambda}}\right) \leq \frac{2}{\lambda}, \tag{156}$$

where the last inequality follows from $\lambda > 1$, and $1 - x \leq \sqrt{1 - x}$ for $0 \leq x < 1$. ∎

In our regret analysis, we will check the scaling of $\lambda_{\max}(V_t)$ and to do it we will prove a differential inequality involving $\lambda_{\max}(V_t)$. The Lemma below will help us to solve that inequality.

Theorem 10 (Petrovitch (1901)) *If u satisfies the differential inequality $\frac{du(t)}{dt} \lesseqgtr f(u(t), t)$, and y is the solution to the ordinary differential equation (ODE) $\frac{dy(t)}{dt} = f(y(t), t)$ under the boundary condition $u(t_0) = y(t_0)$, then:*

$$\forall t > t_0, u(t) \lesseqgtr y(t). \tag{157}$$

C.1. Proof of Theorem 4

Proof To simplify notation we use

$$\lambda_{\min,t} := \lambda_{\min}(V_t), \quad \lambda_{\max,t} := \lambda_{\max}(V_t) \tag{158}$$

First, we fix

$$\lambda_0 = \max\left\{ 2, \sqrt{\frac{2}{3(d-1)}} \frac{d}{6\sqrt{d-1}} + \frac{2}{3(d-1)} \right\}, \tag{159}$$

and later we will justify this choice. Note that the regret can be written as

$$\text{Regret}(T) = \frac{1}{2} \sum_{t=1}^{\widetilde{T}} \sum_{i=1}^{d-1} \left(\|\theta - a_{t,i}^+\|_2^2 + \|\theta - a_{t,i}^-\|_2^2 \right), \tag{160}$$

so to upper bound the regret, we need to quantify the distance $\|\theta - a_{t,i}^+\|_2^2$ between the unknown parameter and the actions we select at each batch t. The history up to batch $t \le \widetilde{T}$ is defined as

$$\mathcal{H}_t := \left(r_{s,1}^+, a_{s,1}^+, r_{s,1}^-, a_{s,1}^-, ..., r_{s,d-1}^+, a_{s,d-1}^+, r_{s,d-1}^-, a_{s,d-1}^- \right)_{s=1}^t. \tag{161}$$

Lemma 2 gives us this distance with a certain probability under the assumption that the event

$$G_t := \{ \left(\mathcal{H}_{t-1}, a_{s,i}^\pm \right) : \sigma_s^2(a_{s,i}^\pm) \le \hat{\sigma}_s^2(\mathcal{H}_{s-1}, a_{s,i}^\pm) \, \forall s \in [t] \}, \tag{162}$$

holds. Using the definition of the subgaussian parameter for the noise ϵ_t (4) we can upper bound it for our choice of actions as

$$\sigma_t^2(a_{t,i}^\pm) \le 1 - \langle a_{t,i}^\pm, \theta \rangle^2 = 1 - (1 - \frac{1}{2}\|\theta - a_{t,i}^\pm\|_2^2)^2$$

$$= \|\theta - a_{t,i}^\pm\|_2^2 - \frac{1}{4}\|\theta - a_{t,i}^\pm\|_2^4 \le \|\theta - a_{t,i}^\pm\|_2^2. \tag{163}$$

Thus, we need to use an estimator of the form (6) that upper bounds the distance $\|\theta - a_{t,i}^\pm\|_2^2$. This quantity depends on the unknown parameter θ, thus we can not guarantee that G_t holds with probability one at each time step t. The other event that we will need to hold to quantify the distance is the following

$$E_t := \{ \mathcal{H}_t : \forall s \in [t], \theta \in \mathcal{C}_s \}, \tag{164}$$

which is guaranteed to hold with probability at least $1 - \delta$ if G_t holds by Lemma 2. We leave the probability computation that both events hold for the end of the proof.

Choosing $w(V_T)$ and instantaneous regret bound under E_t

For now, we will assume that E_t always holds to justify our choice of weights through a bound on $\|\theta - a_{t,i}^\pm\|_2$. This also will allow us to bound the instantaneous regret. From triangle inequality we have

$$\|\theta - a_{t,i}^\pm\|_2 \le \|\theta - \tilde{\theta}_{t-1}^w\|_2 + \|\tilde{\theta}_{t-1}^w - \theta_{t-1}^w\|_2 + \|\theta_{t-1}^w - a_{t,i}^\pm\|_2$$

$$\le \underbrace{\sqrt{\frac{\beta_{t-1,\delta}}{\lambda_{\min,t-1}}}}_{(i)} + \underbrace{\sqrt{\frac{\beta_{t-1,\delta}}{\lambda_{\min,t-1}}}}_{(ii)} + \underbrace{\sqrt{\frac{2}{\lambda_{\min,t-1}}}}_{(iii)}$$

$$\le 3\sqrt{\frac{\beta_{t-1,\delta}}{\lambda_{\min,t-1}}}, \tag{165}$$
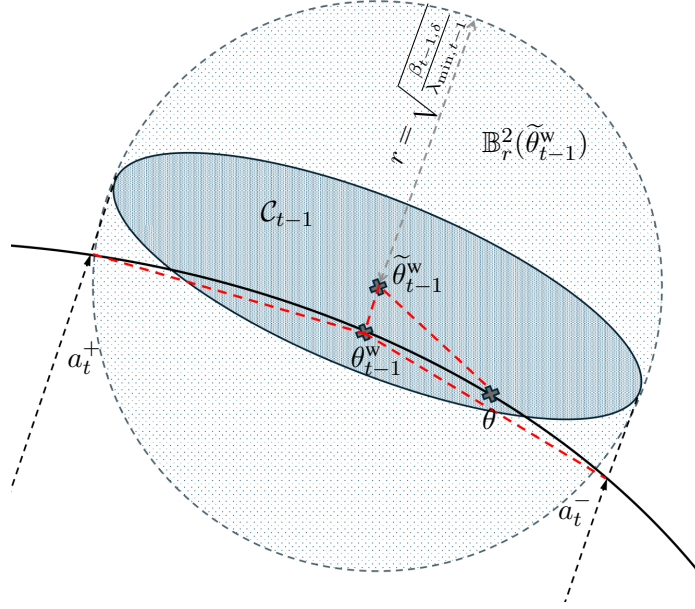
where for the above bounds we use:

Figure 3: Sketch for the triangle inequality used to bound $\|\theta - a_{t,i}^{\pm}\|_2$ in (165). The red lines represent the distances $(i), (ii)$ and (iii). Under the event $\theta \in \mathcal{C}_{t-1}$ we can use $\mathcal{C}_{t-1} \subseteq \mathbb{B}_r^d(\tilde{\theta}_{t-1}^w)$ with r being the longest axis of the ellipsoid and bound all distances by the diameter $2r$.

- (i). Since $\theta \in \mathcal{C}_{t-1}$, by definition of $\mathcal{C}_{t-1}$ (11) we have $\|\theta - \tilde{\theta}_{t-1}^w\|_{V_{t-1}} \leq \sqrt{\beta_{t-1,\delta}}$ and then the inequality follows from $\lambda_{\min,t-1} \mathbb{I}_{d \times d} \leq V_{t-1}$.

- (ii). Since $\theta \in \mathcal{C}_{t-1}$ and $\theta \in \mathbb{S}^d$ this implies that $\mathbb{B}_r^d(\tilde{\theta}_{t-1}^w) \cap \mathbb{S}^d$ is non-empty with $r = \sqrt{\frac{\beta_{t-1,\delta}}{\lambda_{\min,t-1}}}$ (the longest axis of the ellipsoid). Then using that $\mathcal{C}_{t-1} \subseteq \mathbb{B}_r^d(\tilde{\theta}_{t-1}^w)$ and θ_{t-1}^w is the normalized vector of $\tilde{\theta}_{t-1}^w$ which is the center of $\mathcal{C}_{t-1}$ we get $\theta_{t-1}^w \in \mathbb{B}_r^d(\tilde{\theta}_{t-1}^w)$.

- For (iii). We apply Lemma 9 using the expression of $a_t^{\pm}$ (14) with $c = \theta_{t-1}^w$ and $\lambda = \lambda_{\min,t-1}$.

The last inequality follows from $\beta_t \geq 2$. For $t = 0$,

$$\beta_{0,\delta} = \left(\sqrt{\lambda_0} + \sqrt{2 \log \left(\frac{1}{\delta} \right)} \right). \tag{166}$$

Note that at batch t we use that E_{t-1} holds instead of E_t since we want to define $\hat{\sigma}_t$ such that only depends on past information up to time step $t - 1$. Thus, a choice that will guarantee the event G_t to hold under the assumption that E_{t-1} holds is $\hat{\sigma}_t^2 := \frac{9\beta_{t-1,\delta}}{\lambda_{\min,t-1}}$. However our particular update of V_t with the choices of actions $a_{t,i}^+, a_{t,i}^-$ guarantees that

$$\lambda_{\min,t} \geq \sqrt{\frac{2}{3(d-1)}} \lambda_{\max,t}, \tag{167}$$

27

if we use Theorem 3 (later we will check that we are under the right assumptions to use it). It is important to note that the above bound is independent of the events G_t, E_t and it is a consequence only of our particular choice of actions (14). Combining (167) and (165)

$$\|\theta - a_{t,i}^{\pm}\|_2^2 \leq 9\frac{\beta_{t-1,\delta}}{\lambda_{\min,t-1}} \leq \frac{9\sqrt{3}\sqrt{d-1}\beta_{t-1,\delta}}{\sqrt{2}\sqrt{\lambda_{\min,t-1}}}. \tag{168}$$

Thus, using that $\frac{9\sqrt{3}}{\sqrt{2}} \leq 12$ we see that with our definitions of estimator (8) and weights

$$\hat{\sigma}_t^2(a_{t,i}^{\pm}) = \frac{12\sqrt{d-1}\beta_{t-1,\delta}}{\sqrt{\lambda_{\max,t-1}}}, \quad \omega(V_{t-1}) = \frac{1}{\hat{\sigma}_t^2}, \tag{169}$$

are well defined since only depends on the history $\mathcal{H}_{t-1}$, and

$$\sigma_t^2(a_{t,i}^{\pm}) \leq \hat{\sigma}_t^2(a_{t,i}^{\pm}) \quad \text{if} \quad \theta \in \mathcal{C}_{t-1}. \tag{170}$$

To bound the regret our technique uses upper and lower bounds on the scaling of $\text{Tr}(V_t)$ as a function of the number of rounds. In the standard LinUCB technique if the actions are bounded ($\|a_t\|_2 = \Theta(1)$) and this immediately gives the linear scaling $\text{Tr}(V_t) = \Theta(t)$. Since we are updating V_t using $\omega(V_t)$ we need to do some extra work, and we will see that $\text{Tr}(V_t) = \tilde{\Theta}(t^2)$.

Upper bound for $\text{Tr}(V_t)$

The reason why we need an upper bound for $\text{Tr}(V_t)$ is because we will need an upper bound for $\beta_{t,\delta}$ that depends on $\text{Tr}(V_t)$. A direct calculation shows

$$\text{Tr}(V_t) = \lambda_0 + \sum_{s=1}^{t} 2(d-1)\omega(V_{s-1}) \tag{171}$$

$$= \lambda_0 + \sum_{s=1}^{t} \frac{2(d-1)}{12\sqrt{d-1}\beta_{s-1,\delta}}\sqrt{\lambda_{\max,s-1}} \tag{172}$$

$$\leq \lambda_0 + \frac{\sqrt{d-1}}{6\beta_{0.\delta}} \sum_{s=0}^{t-1} \sqrt{\lambda_{\max,s}}, \tag{173}$$

where we used $\beta_{0,\delta} \leq \beta_{t,\delta}$. Then using $\lambda_{\max,t} \leq \text{Tr}(V_t)$ and $\beta_{0,\delta} \geq 1$ we have

$$\lambda_{\max,t} \leq \lambda_0 + \frac{\sqrt{d}}{6} \sum_{s=0}^{t-1} \sqrt{\lambda_{\max,s}}. \tag{174}$$

Now we want to extend $\lambda_{\max,t}$ to a monotonic function on the interval $[0, \widetilde{T}]$. In order to do that we define the linear interpolation $g_1 : [0, \widetilde{T}] \to \mathbb{R}_{\geq 0}$ as

$$g_1(x) := (t+1-x)\lambda_{\max,t} + (x-t)\lambda_{\max,t+1} \quad \text{for} \quad x \in [t, t+1) \quad \text{and} \quad t \in \{0, ..., \widetilde{T}\}. \tag{175}$$

Then using (174) and $g_1(t) = \lambda_{\max,t}$ we have

$$g_1(t) \leq \lambda_0 + \frac{\sqrt{d}}{6} \sum_{s=0}^{t-1} \sqrt{g_1(s)} \leq \lambda_0 + \frac{\sqrt{d}}{6} \int_0^t \sqrt{g_1(x)} \mathrm{d}x \quad \text{for} \quad t \in \{0, ..., \widetilde{T}\}. \tag{176}$$

Now we want to prove the above inequality but for $g_1(x)$ with $x \in [0, \widetilde{T}]$. From the above inequality and the definition of $g_1(x)$ (175) we have that for $x \in [t, t+1)$

$$g_1(x) \leq \lambda_0 + \frac{\sqrt{d}}{6} \sum_{s=0}^{t-1} \sqrt{g_1(s)} + \frac{\sqrt{d}}{6}(x - t)\sqrt{g_1(t)} \tag{177}$$

$$\leq \lambda_0 + \frac{\sqrt{d}}{6} \int_0^t \sqrt{g_1(x)} \mathrm{d}x + \frac{\sqrt{d}}{6}(x - t)\sqrt{g_1(t)}. \tag{178}$$

Then we can use the above relation and

$$\int_t^x \sqrt{g_1(s)} \mathrm{d}s = \int_0^{x-t} \sqrt{g_1(s+t)} \mathrm{d}s = \int_0^{x-t} \sqrt{(1-s)g_1(t) + sg_1(t+1)} \mathrm{d}s \tag{179}$$

$$\geq \int_0^{x-t} \sqrt{g_1(t)} \geq (x-t)\sqrt{g_1(t)}, \tag{180}$$

where the second equality follows from the definition of $g_1(x)$ (175) and the first inequality $g_1(t) = \lambda_{\max,t} \leq \lambda_{\max,t+1} = g_1(t+1)$. We conclude that

$$g_1(x) \leq \lambda_0 + \frac{\sqrt{d}}{6} \int_0^x \sqrt{g_1(s)} \mathrm{d}s. \tag{181}$$

Using the fundamental theorem of calculus

$$g_1(x) \leq \lambda_0 + \frac{\sqrt{d}}{6}(G_1(x) - G_1(0)) \quad \text{where} \quad \frac{\mathrm{d}G_1}{\mathrm{d}x} = \sqrt{g_1(x)}. \tag{182}$$

Fixing $G(0) = 0$ and rearranging

$$\frac{\mathrm{d}G_1}{\mathrm{d}x} \leq \sqrt{\lambda_0 + \frac{\sqrt{d}}{6}G_1(x)}. \tag{183}$$

Now solving for the equality with $G(0) = 0$ and using Theorem 10 we arrive at

$$G_1(x) \leq \frac{\sqrt{d}}{24}x^2 + \sqrt{\lambda_0}x. \tag{184}$$

Finally inserting the above into (182) using that $g_1(t) = \lambda_{\max,t}$

$$\lambda_{\max,t} = g_1(t) \leq \frac{d}{144}t^2 + \frac{\sqrt{d\lambda_0}}{6}t + \lambda_0, \tag{185}$$

and this gives the following bound on $\mathrm{Tr}(V_t)$

$$\mathrm{Tr}(V_t) \leq d\lambda_{\max,t} \leq d\left(\frac{d}{144}t^2 + \frac{\sqrt{d\lambda_0}}{6}t + \lambda_0\right). \tag{186}$$

29

Upper bound for $\beta_{t,\delta}$

In order to give an upper bound for $\beta_{t,\delta}$ (12) we have to give an upper bound for $\det(V_t)$. The inequality of arithmetic and geometric means gives

$$\det(V_t) \leq \left(\frac{\mathrm{Tr}(V_t)}{d}\right)^d, \tag{187}$$

which combined with (186) and (12) gives

$$\beta_{t,\delta} \leq \left(\lambda_0 + \sqrt{2\log\frac{1}{\delta} + d\log\left(\frac{d}{144\lambda_0}t^2 + \frac{\sqrt{d}}{\sqrt{\lambda_0}6}t + 1\right)}\right)^2. \tag{188}$$

Lower bound for $\lambda_{\max,t}$

In order to give an upper bound for regret our main technique is to control the growth of $\lambda_{\max,t}$ through a lower bound for $\lambda_{\max,t}$. We will employ the same technique as in the previous part. For the lower bound we are only interested in the leading terms so we start by lower bounding $\mathrm{Tr}(V_t)$ as

$$\mathrm{Tr}(V_t) \geq \sum_{s=2}^{t} 2(d-1)\omega(V_{s-1}) \tag{189}$$

$$= \sum_{s=2}^{t} \frac{\sqrt{d-1}\sqrt{\lambda_{\max,s-1}}}{6\beta_{s-1,\delta}} \tag{190}$$

$$\geq \frac{\sqrt{d-1}}{6\beta_{\widetilde{T},\delta}} \sum_{s=1}^{t-1} \sqrt{\lambda_{\max,s}}. \tag{191}$$

where we used $\beta_{t,\delta} \leq \beta_{\widetilde{T},\delta}$. Then since our problem is restricted to $\mathbb{R}^d$ we have that for the maximum eigenvalue of V_t,

$$\lambda_{\max,t} \geq \frac{\mathrm{Tr}(V_t)}{d}. \tag{192}$$

Thus combining with the above expressions we have

$$\lambda_{\max,t} \geq \frac{\sqrt{d-1}}{6d\beta_{\widetilde{T},\delta}} \sum_{s=1}^{t-1} \sqrt{\lambda_{\max,s}}. \tag{193}$$

Addding each side by $\frac{\sqrt{d-1}}{6d\beta_{\widetilde{T},\delta}}\sqrt{\lambda_{\max,t}}$ and using the fact that $\lambda_{\max,t} \geq \sqrt{\lambda_{\max,t}}$ since $\lambda_{\max,t} \geq 1$ we have

$$\lambda_{\max,t} + \frac{\sqrt{d-1}}{6d\beta_{\widetilde{T},\delta}}\sqrt{\lambda_{\max,t}} \geq \frac{\sqrt{d-1}}{6d\beta_{\widetilde{T},\delta}} \sum_{s=1}^{t-1} \sqrt{\lambda_{\max,s}} + \frac{\sqrt{d-1}}{6d\beta_{\widetilde{T},\delta}}\sqrt{\lambda_{\max,t}}, \tag{194}$$

$$\left(1 + \frac{\sqrt{d-1}}{6d\beta_{\widetilde{T},\delta}}\right)\lambda_{\max,t} \geq \lambda_{\max,t} + \frac{\sqrt{d-1}}{6d\beta_{\widetilde{T},\delta}}\sqrt{\lambda_{\max,t}} \geq \frac{\sqrt{d-1}}{6d\beta_{\widetilde{T},\delta}} \sum_{s=1}^{t} \sqrt{\lambda_{\max,s}}, \tag{195}$$

$$\lambda_{\max,t} \geq \frac{1}{1 + 6\frac{d}{\sqrt{d-1}}\beta_{\widetilde{T},\delta}} \sum_{s=1}^{t} \sqrt{\lambda_{\max,s}}. \tag{196}$$

As before we want to extend $\lambda_{\max,t}$ to a monotonic function on the interval $[0, \widetilde{T}]$ and define the lower linear interpolation $g_2 : [0, \widetilde{T}] \to \mathbb{R}_{\geq 0}$ as

$$g_2(x) := (t+1-x)\lambda_{\max,t} + (x-t)\lambda_{\max,t+1} \quad \text{for} \quad x \in [t, t+1) \quad \text{and} \quad t \in \{0, ..., \widetilde{T}\}. \tag{197}$$

Then we have that

$$g_2(t) \geq \frac{1}{1 + 6\frac{d}{\sqrt{d-1}}\beta_{\widetilde{T},\delta}} \sum_{s=1}^{t} \sqrt{g_2(s)} \tag{198}$$

$$\geq \frac{1}{1 + 6\frac{d}{\sqrt{d-1}}\beta_{\widetilde{T},\delta}} \int_0^t \sqrt{g_2(x)}\mathrm{d}x \quad \text{for} \quad t \in \{0, ..., \widetilde{T}\}, \tag{199}$$

since $g_2(t) = \lambda_{\max,t}$ for $t \in \{0, ..., \widetilde{T}\}$. Then we can check that the above inequality also holds for $x \in [0, \widetilde{T}]$. If $x \in [t, t+1)$ we have

$$g_2(x) \geq \frac{1}{1 + 6\frac{d}{\sqrt{d-1}}\beta_{\widetilde{T},\delta}} \left(\sum_{s=1}^{t} \sqrt{g_2(s)} + (x-t)\sqrt{g_2(t+1)} \right) \tag{200}$$

$$\geq \frac{1}{1 + 6\frac{d}{\sqrt{d-1}}\beta_{\widetilde{T},\delta}} \left(\int_0^t \sqrt{g_2(x)}\mathrm{d}x + (x-t)\sqrt{g_2(t+1)} \right), \tag{201}$$

where the first inequality we applied the definition of $g_2(x)$ (197) and the inequality for $\lambda_{\max,t}$ (193). Then,

$$\int_t^x \sqrt{g_2(s)}\mathrm{d}s = \int_0^{x-t} \sqrt{g_2(s+t)}\mathrm{d}s = \int_0^{x-t} \sqrt{(1-s)g_2(t) + sg_2(t+1)}\mathrm{d}s \tag{202}$$

$$\leq \int_0^{x-t} \sqrt{g_2(t+1)}\mathrm{d}s \leq (x-t)\sqrt{g_2(t+1)}, \tag{203}$$

where the second equality follows from the definition of $g_2(x)$ and the first inequality from $g_2(t) = \lambda_{\max,t} \leq \lambda_{\max,t+1} \leq g_2(t+1)$. Combining with (200) we have

$$g_2(x) \geq \frac{1}{1 + 6\frac{d}{\sqrt{d-1}}\beta_{\widetilde{T},\delta}} \int_0^x \sqrt{g_2(s)}\mathrm{d}s. \tag{204}$$

Using the fundamental theorem of calculus

$$g_2(x) \geq \frac{1}{1 + 6\frac{d}{\sqrt{d-1}}\beta_{\widetilde{T},\delta}} (G_2(x) - G_2(0)) \quad \text{where} \quad \frac{\mathrm{d}G_2(x)}{\mathrm{d}x} = \sqrt{g_2(x)}. \tag{205}$$

Fixing $G_2(0) = 0$ and rearranging

$$\frac{\mathrm{d}G_2(x)}{\mathrm{d}x} \geq \sqrt{\frac{1}{1 + 6\frac{d}{\sqrt{d-1}}\beta_{\widetilde{T},\delta}}} \sqrt{G_2(x)}. \tag{206}$$

Now solving the equality with $G_2(0) = 0$ and using Theorem 10

$$G_2(x) \geq \frac{1}{4 + 24\frac{d}{\sqrt{d-1}}\beta_{\widetilde{T},\delta}}x^2. \tag{207}$$

Using that $g_2(t) = \lambda_{\max,t}$ and inserting the above result into (205)

$$\lambda_{\max,t} = g_2(t) \geq \frac{1}{4(1 + 6\frac{d}{\sqrt{d-1}}\beta_{\widetilde{T},\delta})^2}t^2. \tag{208}$$

Bounding the regret

From the weight update (16) we have

$$\omega(V_t) = \frac{\sqrt{\lambda_{\max,t-1}}}{12\sqrt{d-1}\beta_{t,\delta}} \leq \frac{1}{12\sqrt{d-1}}\sqrt{\lambda_{\max,t-1}}, \tag{209}$$

so to use Therorem 3 we can choose $C = \frac{1}{12\sqrt{d-1}}$, our choice of λ_0 (159), actions (14) and the sequence $\{\theta_t^{\mathrm{w}}\}_{t=0}^{\infty} \subset \mathbb{S}^d$. Then our update given by (15) satisfies all conditions of Theorem 3 and our choice of actions guarantees

$$\lambda_{\min,t} \geq \sqrt{\frac{2}{3(d-1)}}\lambda_{\max,t}. \tag{210}$$

Then from (188) we can upper bound $\beta_{t,\delta}$ as

$$\beta_{t,\delta} \leq \beta_{T,\max} \quad \text{where} \quad \beta_{T,\max} := \left(\lambda_0 + \sqrt{2\log\frac{1}{\delta} + d\log\left(\frac{d}{144\lambda_0}T^2 + \frac{\sqrt{d}}{\sqrt{\lambda_0}6}T + 1\right)}\right)^2. \tag{211}$$

Finally, we can use the bound from (163)(165) (under the assumptions that E_t and G_t hold) combined with (210) to arrive at

$$\|\theta - a_{t,i}^{\pm}\|_2^2 \leq \frac{9\beta_{T,\max}}{\lambda_{\min,t-1}} \tag{212}$$

$$\leq \frac{12\sqrt{d-1}\beta_{T,\max}}{\sqrt{\lambda_{\max,t-1}}} \tag{213}$$

$$\leq \frac{24\sqrt{d-1}\beta_{T,\max}(1 + 6\frac{d}{\sqrt{d-1}}\beta_{T,\max})}{t-1} \tag{214}$$

$$= \frac{144d\beta_{T,\max}^2 + 24\sqrt{d-1}\beta_{T,\max}}{t-1} \tag{215}$$

where we used the lower bound for $\lambda_{\max,t}$ (208) and $\frac{9\sqrt{3}}{\sqrt{2}} \leq 12$. Then using the above result we can bound the regret as

$$\text{Regret}(T) = \frac{1}{2}\sum_{t=1}^{\widetilde{T}}\sum_{i=1}^{d-1}\left(\|\theta - a_{t,i}^+\|_2^2 + \|\theta - a_{t,i}^-\|_2^2\right) \tag{216}$$

$$\leq 4(d-1) + \frac{1}{2}\sum_{t=2}^{\widetilde{T}}\sum_{i=1}^{d-1}\left(\|\theta - a_{t,i}^+\|_2^2 + \|\theta - a_{t,i}^-\|_2^2\right) \tag{217}$$

$$\leq 4(d-1) + \left(144d^2\beta_{T,\max}^2 + 24(d-1)^{\frac{3}{2}}\beta_{T,\max}\right)\sum_{t=2}^{\tilde{t}}\frac{1}{t-1} \tag{218}$$

$$\leq 4(d-1) + \left(144d^2\beta_{T,\max}^2 + 24(d-1)^{\frac{3}{2}}\beta_{T,\max}\right)\log\left(\frac{T}{2(d-1)}\right). \tag{219}$$

Using that $\beta_{T,\max} = O(d\log(T))$ we have that

$$\text{Regret}(T) = \widetilde{O}(d^4\log^3(T)). \tag{220}$$

Success probability analysis

From the computations in (163)(165)(169)(170) and our choice of $\sigma_t^2(a_t)$ (16) we have that

$$\text{if}\quad \theta \in \mathcal{C}_{s-1} \Rightarrow \sigma_s^2(a_{s,i}^\pm) \leq \frac{12\sqrt{d-1}\beta_{t-1,\delta}}{\sqrt{\lambda_{\max,t-1}}} = \hat{\sigma}_s^2(a_{s,i}^\pm), \tag{221}$$

thus,

$$\Pr(G_t) \geq \Pr(E_{t-1}) \tag{222}$$

where we have used the definitions of the events G_t (162),E_t (164).

The initial choice $\hat{\sigma}_1^2 = 1$ implies $\sigma^2 \leq \hat{\sigma}_1^2$ and using Lemma 2 we have

$$\Pr(G_1) = 1, \tag{223}$$

$$\Pr(E_1) \geq 1 - \delta. \tag{224}$$

Using Bayes theorem

$$\Pr(E_t) = \frac{\Pr(E_t|G_t)\Pr(G_t)}{\Pr(G_t|E_t)} = \Pr(E_t|G_t)\Pr(G_t), \tag{225}$$

where in the last equality we used

$$\Pr(G_t|E_t) = 1. \tag{226}$$

From Lemma 2 we have

$$\Pr(E_t|G_t) \geq 1 - \delta. \tag{227}$$

Thus, applying recursively the above inequalities

$$\Pr(G_2) \geq \Pr(E_1) \geq 1 - \delta, \tag{228}$$

$$\Pr(E_2) = \Pr(E_2|G_2)\Pr(G_2) \geq (1 - \delta)^2 \tag{229}$$

$$\vdots$$

$$\Pr(G_t) \geq \Pr(E_{t-1}) \geq (1 - \delta)^{t-1} \tag{230}$$

$$\Pr(E_t) = \Pr(E_t|G_t)\Pr(G_t) \geq (1 - \delta)^t \tag{231}$$

Finally to bound the regret we used the assumption that $\theta \in \mathcal{C}_t$ for any $t \in \{1, ..., \widetilde{T} - 1\}$. Thus we can bound the probability that the obtained bound holds as

$$\Pr\left(\text{Regret}(T) \leq 4(d-1) + \left(144d^2\beta_{T,\max}^2 + 24(d-1)^{\frac{3}{2}}\beta_{T,\max}\right)\log\left(\frac{T}{2(d-1)}\right)\right) \tag{232}$$

$$\geq \Pr(E_{\widetilde{T}} \cap G_{\widetilde{T}}) = \Pr(G_{\widetilde{T}})\Pr(E_{\widetilde{T}}|G_{\widetilde{T}}). \tag{233}$$

$$\geq (1 - \delta)^{\widetilde{T}-1}(1 - \delta) = (1 - \delta)^{\widetilde{T}}. \tag{234}$$

The result follows choosing $\delta = \frac{\delta'}{T}$ for some $0 < \delta' < 1$, the inequality $\left(1 - \frac{\delta'}{T}\right)^{\widetilde{T}} \geq 1 - \delta'$ and $\frac{\sqrt{d}}{d-1} \leq 2$. ∎

Corollary 11 *Under the same assumptions of Theorem 4, we can choose $\delta = \frac{1}{\widetilde{T}}$ and Algorithm 1 achieves*

$$\mathbb{E}[\text{Regret}(T)] \leq 8(d-1) + \left(144d^2\beta_{T,\max}^2 + 24(d-1)^{\frac{3}{2}}\beta_{T,\max}\right)\log\left(\frac{T}{2(d-1)}\right), \tag{235}$$

or

$$\mathbb{E}[\text{Regret}(T)] = O(d^4 \log^3(T)). \tag{236}$$

Proof Defining the event

$$R_T = \left\{\mathcal{H}_{\widetilde{T}} : \text{Regret}(T) \leq 4(d-1) + \left(144d^2\beta_{T,\max}^2 + 24(d-1)^{\frac{3}{2}}\beta_{T,\max}\right)\log\left(\frac{T}{2(d-1)}\right)\right\} \tag{237}$$

we have by Theorem 4 that

$$\Pr(R_T) \geq 1 - \delta = 1 - \frac{1}{\widetilde{T}}, \tag{238}$$

$$\Pr(R_T^C) \leq \frac{1}{\widetilde{T}}. \tag{239}$$

Then

$$
\mathbb{E}[\text{Regret}(T)] = \mathbb{E}\left[\text{Regret}(T)\mathbb{I}\{R_T\}\right] + \mathbb{E}\left[\text{Regret}(T)\mathbb{I}\{R_T^C\}\right]
$$

$$
\leq 4(d-1) + \left(144d^2\beta_{T,\max}^2 + 24(d-1)^{\frac{3}{2}}\beta_{T,\max}\right)\log\left(\frac{T}{2(d-1)}\right) + 4(d-1)\widetilde{T}\Pr\left(R_T^C\right)
$$

(240)

$$
\leq 8(d-1) + \left(144d^2\beta_{T,\max}^2 + 24(d-1)^{\frac{3}{2}}\beta_{T,\max}\right)\log\left(\frac{T}{2(d-1)}\right),
$$

(241)

and the results follows using that for $\delta = \frac{1}{T}$, $\beta_{T,\max} = O(d\log(T))$. ∎

Appendix D. Minimax lower bound for linear bandit $\mathcal{A} = \mathbb{S}^d, \theta \in \mathbb{S}^d$ and constant noise

In this section, we prove that any small perturbation in our noise model leads to $\Omega(\sqrt{T})$ regret. We study a minimax lower bound for the following reward model

$$
r_t = \langle a_t, \theta \rangle + \mathcal{N}\left(0, \sigma_{t,\theta}^2\right) + \mathcal{N}\left(\tilde{\mu}, \tilde{\sigma}^2\right) = \mathcal{N}(\tilde{\mu}_{t,\theta}, \tilde{\sigma}_{t,\theta}^2),
$$

(242)

where $\theta \in \mathbb{S}^d$, $\tilde{\mu}, \tilde{\sigma}^2 \in \mathbb{R}$, $\tilde{\sigma}^2 > 0$, and

$$
\sigma_{t,\theta}^2 := 1 - \langle a_t, \theta \rangle^2, \quad \tilde{\mu}_{t,\theta} := \langle a_t, \theta \rangle + \tilde{\mu}_{t,\theta}, \quad \tilde{\sigma}_{t,\theta}^2 := \sigma_{t,\theta}^2 + \tilde{\sigma}^2.
$$

(243)

Recall that the regret is given by

$$
\text{Regret}(T) = \sum_{t=1}^{T} 1 - \langle a_t, \theta \rangle.
$$

(244)

Our lower bound proof is an adaptation of the lower bound given in Abeille et al. (2021) that was introduced for logistic bandits and this provides a lower bound for linear bandits with $\mathcal{A} = \mathbb{S}^d, \theta \in \mathbb{S}^d$. For completeness, in Lemma 12 we reproduce the main steps and we note that in our setting some parts simplify. In order to state the result we define $\{e_i\}_{i=1}^d$ the standard basis in $\mathbb{R}^d$ and the flip operator. Given $\theta \in \mathbb{R}$, $i \in [d]$ the flip operator is defined as

$$
\text{Flip}_i(\theta) := (\theta_1, \theta_2, ..., -\theta_i, ..., \theta_d).
$$

(245)

Lemma 12 *Given a stochastic linear bandit with action set $\mathcal{A} = \mathbb{S}^d = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$, a policy π, a reference parameter $\theta_* = \|\theta_*\|_2 e_1 \in \mathbb{R}^d$ and the set of parameters*

$$
\Xi = \left\{\theta_* + \epsilon\sum_{i=2}^{d} v_i e_i, \quad v_i \in \{-1, 1\}\right\},
$$

(246)

where $0 < \epsilon \leq \frac{1}{d-1}$. Then the average of the expected regret over the set can be lower bounded as

$$\frac{1}{|\Xi|} \sum_{\theta \in \Xi} \mathbb{E}_\theta \left[\text{Regret}(T) \right] \geq T\epsilon^2 \left(\frac{d}{8} - \frac{\sqrt{d}}{4} \sqrt{\frac{1}{|\Xi|} \sum_{\theta \in \Xi} \sum_{i=2}^{d} D_{KL} \left(\mathbb{P}_\theta, \mathbb{P}_{\text{Flip}_i(\theta)} \right)} \right), \tag{247}$$

where $\mathbb{P}_\theta, \mathbb{P}_{\text{Flip}_i(\theta)}$ are the probability measures of actions and rewards obtained by the interaction of the policy π with the environments θ,$\text{Flip}_i(\theta)$ respectively. Moreover, ϵ can be fixed such that for all $\theta \in \Xi$ then $\theta \in \mathcal{B}_2^d$.

Proof First, we note that for any $\theta \in \Xi$ it has a constant norm, i.e

$$\|\theta\| = \sqrt{\|\theta_*\|^2 + (d-1)\epsilon^2}, \quad \text{for } \theta \in \Xi. \tag{248}$$

Thus, we fix

$$\|\theta_*\| = \sqrt{1 - (d-1)\epsilon^2} \tag{249}$$

and we have that

$$\text{if} \quad \theta \in \Xi \Rightarrow \theta \in \mathcal{S}_{d-1}. \tag{250}$$

Note that when choosing ϵ we will have the following restriction

$$\epsilon^2 \leq \frac{1}{d-1}. \tag{251}$$

From the expression of the regret we have

$$\mathbb{E}_\theta \left[\text{Regret}_\theta(T) \right] = \frac{1}{2} \mathbb{E}_\theta \left[\sum_{t=1}^{T} \|\theta - a_t\|^2 \right] = \frac{1}{2} \mathbb{E}_\theta \left[\sum_{i=1}^{d} \sum_{t=1}^{T} [\theta - a_t]_i^2 \right] \tag{252}$$

$$\geq \frac{1}{2} \mathbb{E}_\theta \left[\sum_{i=1}^{d} \sum_{t=1}^{T} [\theta - a_t]_i^2 \, \mathbb{1}\{A_i(\theta)\} \right] = \text{(a)} \tag{253}$$

where the event $A_i(\theta)$ is defined as

$$A_i(\theta) := \left\{ \left[\theta - \frac{\theta_*}{\|\theta_*\|} \right]_i \left[\sum_{t=1}^{T} \frac{\theta_*}{\|\theta_*\|} - a_t \right]_i \geq 0 \right\}, \tag{254}$$

for $i = 1, 2, ..., d$. Then we want to compare with the reference environment θ_*, and we can introduce it as

$$\text{(a)} = \frac{1}{2} \mathbb{E}_\theta \left[\sum_{i=1}^{d} \sum_{t=1}^{T} \left[\theta - \frac{\theta_*}{\|\theta_*\|} + \frac{\theta_*}{\|\theta_*\|} - a_t \right]_i^2 \mathbb{1}\{A_i(\theta)\} \right] \tag{255}$$

$$\geq \frac{1}{2} \mathbb{E}_\theta \left[\sum_{i=1}^{d} \sum_{t=1}^{T} \left[\theta - \frac{\theta_*}{\|\theta_*\|} \right]_i^2 \mathbb{1}\{A_i(\theta)\} \right] = \frac{T}{2} \mathbb{E}_\theta \left[\sum_{i=1}^{d} \left[\theta - \frac{\theta_*}{\|\theta_*\|} \right]_i^2 \mathbb{1}\{A_i(\theta)\} \right] \tag{256}$$

where the last inequality follows from the identity $(x + y)^2 = x^2 + 2xy + y^2$ with $x = \theta - \frac{\theta_*}{\|\theta_*\|}$, $y = \frac{\theta_*}{\|\theta_*\|} - a_t$ and the definition of the event $A_i(\theta)$. Using the above computation and that $\theta \in \Xi$,

$$\mathbb{E}_\theta\left[\text{Regret}_\theta(T)\right] \geq \frac{T}{2}\mathbb{E}_\theta\left[\sum_{i=2}^d \left[\theta - \frac{\theta_*}{\|\theta_*\|}\right]_i^2 \mathbb{1}\{A_i(\theta)\}\right] \tag{257}$$

$$\geq \frac{T\epsilon^2}{2}\sum_{i=2}^d \mathbb{E}_\theta\left[\mathbb{1}\{A_i(\theta)\}\right] = \frac{T\epsilon^2}{2}\sum_{i=2}^d \mathbb{P}_\theta\left(A_i(\theta)\right). \tag{258}$$

Now we want to apply the averaging hammer technique that was introduced in (Lattimore and Szepesvári, 2020, Chapter 24, Theorem 24.1). Let $i \in \{2, ..., d\}$, fix $\theta \in \Xi$ then using that $[\theta_*]_i = 0$ it is easy to check that

$$A_i(\text{Flip}_i(\theta)) = A_i^C(\theta). \tag{259}$$

Then using the definition of the total variation distance and Pinsker inequality we have

$$\mathbb{P}_{\text{Flip}_i(\theta)}(A_i(\text{Flip}_i(\theta)) \geq \mathbb{P}_\theta(A_i(\text{Flip}_i(\theta))) - D_{TV}(\mathbb{P}_\theta, \mathbb{P}_{\text{Flip}_i(\theta)}) \tag{260}$$

$$\geq \mathbb{P}_\theta(A_i^C(\theta)) - \sqrt{\frac{1}{2}D_{KL}(\mathbb{P}_\theta, \mathbb{P}_{\text{Flip}_i(\theta)})}. \tag{261}$$

Then applying the above results we have

$$\frac{1}{|\Xi|}\sum_{\theta \in \Xi}\sum_{i=2}^d \mathbb{P}_\theta\left(A_i(\theta)\right) \geq \frac{1}{2|\Xi|}\sum_{i=2}^d\sum_{\theta \in \Xi}\left(\mathbb{P}_\theta(A_i(\theta)) + \mathbb{P}_{\text{Flip}_i(\theta)}(A_i(\text{Flip}_i(\theta)))\right) \tag{262}$$

$$\geq \frac{1}{2|\Xi|}\sum_{i=2}^d\sum_{\theta \in \Xi}1 - \sqrt{\frac{1}{2}D_{KL}(\mathbb{P}_\theta, \mathbb{P}_{\text{Flip}_i(\theta)})}. \tag{263}$$

Using Jensen inequality, Cauchy-Schwartz, and the fact that $d \geq 1$ we have

$$\frac{1}{|\Xi|}\sum_{\theta \in \Xi}\sum_{i=2}^d \mathbb{P}_\theta\left(A_i(\theta)\right) \geq \frac{d}{4} - \frac{\sqrt{d}}{2}\sqrt{\frac{1}{|\Xi|}\sum_{\theta \in \Xi}\sum_{i=2}^d D_{KL}\left(\mathbb{P}_\theta, \mathbb{P}_{\text{Flip}_i(\theta)}\right)} \tag{264}$$

And the result follows from combining (257) and (264). ∎

Before proving the main theorem we need a formula for the KL divergence between two normal distributions. Given $\mathcal{N}(\mu_1, \sigma_1^2)$ and $\mathcal{N}(\mu_2, \sigma_2^2)$ it follows from a direct calculation that

$$D_{KL}\left(\mathcal{N}(\mu_1, \sigma_1^2), \mathcal{N}(\mu_2, \sigma_2^2)\right) = \frac{1}{2}\left(\log\left(\frac{\sigma_2^2}{\sigma_1^2}\right) + \frac{\sigma_1^2}{\sigma_2^2} - 1\right) + \frac{(\mu_1 - \mu_2)^2}{\sigma_2^2}. \tag{265}$$

With these results, we are ready to prove the main theorem.

Theorem 13 *Given a stochastic linear bandit with action set $\mathcal{A} = \mathbb{S}^d = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$ and reward model given by (242), then there exists an unknown parameter $\theta \in \mathbb{R}^d$ such that $\|\theta\|_2 = 1$ and*

$$\mathbb{E}_\theta[\text{Regret}(T)] \geq \frac{1}{100}\hat{\sigma}d\sqrt{T}, \tag{266}$$

for $T \geq \frac{1}{6400}d^2\tilde{\sigma}^2$.

37

Proof First we suppose that for any $\theta \in \Xi$,

$$\mathbb{E}_\theta[\text{Regret}(T)] \leq Cd\tilde{\sigma}\sqrt{T}, \tag{267}$$

for some constant $C > 0$ that without loss of generality we set to $C = 1$. We see from Lemma 12 that we need to compute $D_{KL}\left(\mathbb{P}_\theta, \mathbb{P}_{\text{Flip}_i(\theta)}\right)$. From Lattimore and Szepesvári (2020)[Chapter 24, Theorem 24.1] we have

$$D_{KL}\left(\mathbb{P}_\theta, \mathbb{P}_{\text{Flip}_i(\theta)}\right) = \mathbb{E}_\theta\left[\sum_{t=1}^T D_{KL}\left(\mathcal{N}(\tilde{\mu}_{t,\theta}, \tilde{\sigma}_{t,\theta}^2), \mathcal{N}(\tilde{\mu}_{t,\text{Flip}_i(\theta)}, \tilde{\sigma}_{t,\text{Flip}_i(\theta)}^2)\right)\right]. \tag{268}$$

Using (243) and (265) we have

$$D_{KL}\left(\mathcal{N}(\tilde{\mu}_{t,\theta}, \tilde{\sigma}_{t,\theta}^2), \mathcal{N}(\tilde{\mu}_{t,\text{Flip}_i(\theta)}, \tilde{\sigma}_{t,\theta'}^2)\right) = \tag{269}$$

$$\frac{1}{2}\left(\underbrace{\log\left(\frac{1 - \langle a_t, \text{Flip}_i(\theta)\rangle^2 + \tilde{\sigma}^2}{1 - \langle a_t, \theta\rangle^2 + \tilde{\sigma}^2}\right)}_{(a)} + \underbrace{\frac{1 - \langle a_t, \theta\rangle^2 + \tilde{\sigma}^2}{1 - \langle a_t, \text{Flip}_i(\theta)\rangle^2 + \tilde{\sigma}^2} - 1}_{(b)}\right) \tag{270}$$

$$+ \frac{\langle a_t, \theta - \text{Flip}_i(\theta)\rangle^2}{1 - \langle a_t, \text{Flip}_i(\theta)\rangle^2 + \tilde{\sigma}^2}. \tag{271}$$

Now we are going to upper bound the terms (a) and (b).

$$(a) = \log\left(1 + \frac{\langle a_t, \theta\rangle^2 - \langle a_t, \text{Flip}_i(\theta)\rangle^2}{1 - \langle a_t, \theta\rangle^2 + \tilde{\sigma}^2}\right) \leq \frac{\langle a_t, \theta\rangle^2 - \langle a_t, \text{Flip}_i(\theta)\rangle^2}{1 - \langle a_t, \theta\rangle^2 + \tilde{\sigma}^2} \tag{272}$$

$$\leq \frac{\langle a_t, \theta\rangle^2 - \langle a_t, \text{Flip}_i(\theta)\rangle^2}{\tilde{\sigma}^2} \tag{273}$$

$$(b) = 1 + \frac{\langle a_t, \text{Flip}_i(\theta)\rangle^2 - \langle a_t, \theta\rangle^2}{1 - \langle a_t, \text{Flip}_i(\theta)\rangle^2 + \tilde{\sigma}^2} \leq 1 + \frac{\langle a_t, \text{Flip}_i(\theta)\rangle^2 - \langle a_t, \theta\rangle^2}{\tilde{\sigma}^2} \tag{274}$$

Thus, we have that $(a) + (b) - 1 \leq 0$ and

$$D_{KL}\left(\mathcal{N}(\tilde{\mu}_{t,\theta}, \tilde{\sigma}_{t,\theta}^2), \mathcal{N}(\tilde{\mu}_{t,\text{Flip}_i(\theta)}, \tilde{\sigma}_{t,\theta'}^2)\right) \leq \frac{4\epsilon^2}{\tilde{\sigma}^2}[a_t]_i^2. \tag{275}$$

Inserting the above into $D_{KL}\left(\mathbb{P}_\theta, \mathbb{P}_{\text{Flip}_i(\theta)}\right)$ we have

$$\sum_{i=2}^d D_{KL}\left(\mathbb{P}_\theta, \mathbb{P}_{\text{Flip}_i(\theta)}\right) \leq \frac{4\epsilon^2}{\tilde{\sigma}^2}\sum_{i=2}^d \mathbb{E}_\theta\left[\sum_{t=1}^T [\theta - a_t + \theta]_i^2\right] \tag{276}$$

$$\leq \frac{8\epsilon^2}{\tilde{\sigma}^2}\mathbb{E}_\theta\left[\sum_{t=1}^T\sum_{i=2}^d [\theta - a_t]_i^2 + [\theta]_i^2\right] \leq \frac{8\epsilon^2}{\tilde{\sigma}^2}\mathbb{E}_\theta\left[\sum_{t=1}^T\sum_{i=1}^d [\theta - a_t]_i^2 + \sum_{t=1}^T\sum_{i=2}^d [\theta]_i^2\right] \tag{277}$$

$$= \frac{8\epsilon^2}{\tilde{\sigma}^2}\left(2\mathbb{E}_\theta[\text{Regret}(T)] + T(d-1)\epsilon^2\right) \leq \frac{16\epsilon^2\tilde{\sigma}d\sqrt{T}}{\tilde{\sigma}^2} + \frac{8d\epsilon^4 T}{\tilde{\sigma}^2} \tag{278}$$

where we have used that $d \geq 1$, $\theta \in \Xi$, and (267). Thus, inserting the above into the result of Lemma 12,

$$\frac{1}{|\Xi|} \sum_{\theta \in \Xi} \mathbb{E}_\theta \left[\text{Regret}(T) \right] \geq dT\epsilon^2 \left(\frac{1}{8} - \frac{1}{4} \sqrt{\frac{16\tilde{\sigma}\epsilon^2 \sqrt{T} + 8\epsilon^4 T}{\tilde{\sigma}^2}} \right) \tag{279}$$

Finally, choosing $\epsilon^2 = \frac{\tilde{\sigma}}{80\sqrt{T}}$ we have

$$\frac{1}{|\Xi|} \sum_{\theta \in \Xi} \mathbb{E}_\theta \left[\text{Regret}(T) \right] \geq \tilde{\sigma} d\sqrt{T} \left(\frac{1}{8} - \frac{1}{4} \sqrt{\frac{16}{80} + \frac{8}{6400}} \right) \tag{280}$$

$$\geq \frac{1}{100} \tilde{\sigma} d\sqrt{T}. \tag{281}$$

We note that in order to have $\epsilon^2 \leq \frac{1}{d-1}$ we need $T \geq \frac{1}{6400} d^2 \tilde{\sigma}^2$. Note that we proved the result under the hypothesis that (267) holds. If (267) does not hold the result follows trivially. ∎

Appendix E. Failure of lower-bound methods for vanishing noise

Specifically, we encounter the problem that these calculations rely on the computation of the KL divergences $D_{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta'})$ for two "close" unknown parameters $\theta, \theta' \in \mathbb{S}^d$ and it is not possible to give a non-trivial bound since their variance is arbitrarily close to 0 when we select actions near the unknown parameters. We can easily see this fact for example if we consider Gaussian distributed rewards and use the divergence decomposition lemma that leads to

$$D_{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = \mathbb{E}_\theta \left[\sum_{t=1}^T D_{KL} \left(\mathcal{N}(\langle \theta, a_t \rangle, 1 - \langle \theta, a_t \rangle^2), \mathcal{N}(\langle \theta', a_t \rangle, 1 - \langle \theta', a_t \rangle^2) \right) \right], \tag{282}$$

where $\mathcal{N}(\mu, \sigma^2)$ is a Gaussian probability distribution with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 \geq 0$. The proof requires to upper bound of the above divergence but a simple computation shows that the divergence D_{KL} for the reward distributions cannot be upper bounded since the variances can go arbitrarily close to 0 and we only get the trivial bound $D_{KL} \leq \infty$.