The End of Optimism?
An Asymptotic Analysis of Finite-Armed Linear Bandits

Tor Lattimore
Indiana University, Bloomington

Csaba Szepesvári
University of Alberta, Edmonton

Abstract

Stochastic linear bandits are a natural and simple generalisation of finite-armed bandits with numerous practical applications. Current approaches focus on generalising existing techniques for finite-armed bandits, notably the optimism principle and Thompson sampling. Prior analysis has mostly focussed on the worst-case setting. We analyse the asymptotic regret and show matching upper and lower bounds on what is achievable. Surprisingly, our results show that no algorithm based on optimism or Thompson sampling will ever achieve the optimal rate. In fact, they can be arbitrarily far from optimal, even in very simple cases. This is a disturbing result because these techniques are standard tools that are widely used for sequential optimisation, for example, generalised linear bandits and reinforcement learning.

1 INTRODUCTION

The linear bandit is a simple generalisation of the finite-armed bandit. Let $\mathcal{A} \subset \mathbb{R}^d$ be a finite set that spans $\mathbb{R}^d$ with $|\mathcal{A}| = k$ and $\|x\|_2 \leq 1$ for all $x \in \mathcal{A}$. A learner interacts with the bandit over n rounds. In each round t the learner chooses an action (arm) $A_t \in \mathcal{A}$ and observes a payoff $Y_t = \langle A_t, \theta \rangle + \eta_t$ where $\eta_t \sim \mathcal{N}(0, 1)$ is Gaussian noise and $\theta \in \mathbb{R}^d$ is an unknown parameter. The optimal action is $x^* = \arg\max_{x \in \mathcal{A}} \langle x, \theta \rangle$, which is not known since it depends on θ. The assumption that $\mathcal{A}$ spans $\mathbb{R}^d$ is non-restrictive, since if $\mathrm{span}(\mathcal{A})$ has rank $r < d$, then one can simply use a different basis for which all but r coordinates are always zero and then drop them from the analysis. The Gaussian assumption can be relaxed to 1-subgaussian (see Rivasplata, 2012) for our upper bound, but is needed for

the lower bound. Our performance measure is the expected pseudo-regret (from now on just the regret), which is

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

where π is the strategy that determines the actions (a mapping from observations to a distribution over the actions A_t) and the expectation is taken with respect to the actions of the strategy and the noise. There are a number of algorithms designed for minimising the regret in this setting, most of which use one of two algorithmic designs. The first is the principle of optimism in the face of uncertainty, which was originally applied to finite-armed bandits by Agrawal [1995], Katehakis and Robbins [1995], Auer et al. [2002] and many others, and more recently to linear bandits [Auer, 2002, Dani et al., 2008, Abbasi-Yadkori et al., 2011, 2012]. The second algorithm design is Thompson sampling, which is an old algorithm [Thompson, 1933] that has experienced a resurgence in popularity because of its impressive practical performance and theoretical guarantees for finite-armed bandits [Kaufmann et al., 2012, Korda et al., 2013]. Thompson sampling has also recently been applied to linear bandits with good empirical performance [Chapelle and Li, 2011] and near-minimax theoretical guarantees [Agrawal and Goyal, 2013].

While both approaches lead to practical algorithms (especially Thompson sampling), we will show they are flawed in that algorithms based on these ideas cannot be close to asymptotically optimal. Along the way we characterise the optimal achievable asymptotic regret and design an impractical strategy achieving it. This is an important message because optimism and Thompson sampling are widely used beyond the finite-armed case. Examples include generalised linear bandits [Filippi et al., 2010], spectral bandits [Valko et al., 2014], and even learning in Markov decision processes [Auer et al., 2010, Gopalan and Mannor, 2015].

The disadvantages of these approaches are obscured in the worst-case regime, where both are quite close to optimal. One might question whether or not the asymptotic analysis is relevant in practice. The gold standard would be instance-dependent finite-time guarantees like what is available for finite-armed bandits, but historically

the asymptotic analysis has served as a useful guide towards understanding the trade-offs in finite-time. Besides hiding the structure of specific problems, pushing for optimality in the worst-case regime can also lead to sub-optimal instance-dependent guarantees. For example, the MOSS algorithm for finite-armed bandits is minimax optimal, but far from finite-time optimal [Audibert and Bubeck, 2009]. For these reasons we believe that understanding the asymptotics of a problem is a useful first step towards optimal finite-time instance-dependent guarantees that are most desirable. Note that finite-time problem-dependent guarantees are known for linear bounds, but none are close to optimal [Rusmevichientong and Tsitsiklis, 2010, Abbasi-Yadkori et al., 2011].

It is worth mentioning that partial monitoring (a more complicated online learning setting) is a well known example of the failure of optimism [Bartók et al., 2014]. Although related, the partial monitoring framework is more general than the bandit setting because the learner may not observe the reward even for the action they take, which means that additional exploration is usually necessary in order to gain information. Basic results in partial monitoring are concerned with characterizing whether an instance is easier or harder than bandit instances. More recently, the question of asymptotic instance optimality was studied in finite stochastic partial monitoring [Komiyama et al., 2015], and the special setting of learning with side information [Wu et al., 2015]. While the algorithms derived in these works served as inspiration, the analysis and the algorithms do not generalise in a simple direct fashion to the linear setting, which requires a careful study of how information is transferred between actions in a linear setting. Optimism has also been shown to fail in hand-crafted bandit-like problems where information generalises between actions [Russo and Van Roy, 2014].

This last paper introduces another algorithmic approach called information directed sampling, which is a promising candidate to overcome the failures of optimism and Thompson sampling. So far, however, the theoretical analysis of this algorithm has been restricted to the Bayesian regret, with problem-dependent frequentist bounds remaining an interesting open problem. Yet another algorithmic approach is the explore-then-exploit idea, which involves committing to periods of exploration followed by exploitation (see, for example, the work by Rusmevichientong and Tsitsiklis [2010]). Our algorithm uses this idea, but differs from from previous attempts by using a carefully chosen data-dependent exploration strategies. A closely related setting is the best-arm identification problem, where the objective is not to minimise the regret, but rather to identify the optimal arm in as few rounds as possible. This problem was studied by Soare et al. [2014], who propose an algorithm that also refines its exploration distribution in a data-dependent way. Of course, the different objective leads to a different algorithm and analysis, but nevertheless provides inspiration. We note that our new concentration bounds may also be applied to that setting, where we believe they may provide the means for an optimal asymptotic analysis as seen in the finite-armed unstructured setting [Garivier and Kaufmann, 2016].

2 NOTATION

For positive semidefinite G (written as $G \succeq 0$) and vector x we write $\|x\|_G^2 = x^\top G x$. The Euclidean norm of a vector $x \in \mathbb{R}^d$ is $\|x\|$ and the spectral norm of a matrix A is $\|A\|$. The largest eigenvalue is $\lambda_{\max}(A)$ and the smallest is $\lambda_{\min}(A)$. The pseudo-inverse of a matrix A is denoted by $A^\dagger$. The mean of arm $x \in \mathcal{A}$ is $\mu_x = \langle x, \theta \rangle$ and the optimal mean is $\mu^* = \max_{x \in \mathcal{A}} \mu_x$. Let $x^* \in \mathcal{A}$ be any *optimal action* such that $\mu_{x^*} = \mu^*$. The sub-optimality gap of arm x is $\Delta_x = \mu^* - \mu_x$ and $\Delta_{\min} = \min\{\Delta_x : \Delta_x > 0, x \in \mathcal{A}\}$ and $\Delta_{\max} = \max\{\Delta_x : x \in \mathcal{A}\}$. The number of times arm x has been chosen after round t is denoted by $T_x(t) = \sum_{s=1}^t \mathbb{1}\{A_t = x\}$ and $T_*(t) = \sum_{s=1}^t \mathbb{1}\{\mu_{A_t} = \mu^*\}$. A policy π is *consistent* if for all θ and $p > 0$ it holds that $R_\theta^\pi(n) = o(n^p)$. Note that this is equivalent to $R_\theta^\pi(n) = O(n^p)$ and also to $\limsup_{n\to\infty} \log(R_\theta^\pi(n))/\log(n) \leq 0$. When more appropriate, we will use the more precise Landau notation $a_n \in O(b_n)$ (also with Ω, o and ω). Vectors in $\mathbb{R}^k$ will often be indexed by the action set, which we assume has an arbitrary fixed order. For example, we might write $\alpha \in \mathbb{R}^k$ and refer to $\alpha_x \in \mathbb{R}$ for some $x \in \mathcal{A}$. If $S \subseteq \mathbb{N}$ is infinite, then we write $\lim_{n \in S} f(n)$ for the limit of f taken over the elements of S in increasing order. The limit inferior and superior are used in the same way.

3 LOWER BOUND

We note first that the finite-armed UCB algorithm of Agrawal [1995], Katehakis and Robbins [1995] can be used on this problem by disregarding the structure on the arms to achieve an asymptotic regret of

$$\limsup_{n\to\infty} \frac{R_\theta^{\text{UCB}}(n)}{\log(n)} = \sum_{x \in \mathcal{A}:\Delta_x > 0} \frac{2}{\Delta_x}.$$

This quantity depends *linearly* on the number of suboptimal arms, which may be very large (much larger than the dimension) and is very undesirable. Nevertheless we immediately observe that the asymptotic regret should be logarithmic. The following theorem and its corollary characterises the optimal asymptotic regret.

Theorem 1. *Fix $\theta \in \mathbb{R}^d$ such that there is a unique optimal arm. Let π be a consistent policy and let*

$$\bar{G}_n = \mathbb{E}\left[\sum_{t=1}^n A_t A_t^\top\right].$$

Then $\liminf_{n\to\infty} \lambda_{\min}(\bar{G}_n)/\log(n) > 0$ *(and so $\bar{G}_n$ is non-singular for sufficiently large n). Furthermore, for all suboptimal $x \in \mathcal{A}$ it holds that*

$$\limsup_{n\to\infty} \log(n) \|x - x^*\|_{\bar{G}_n^{-1}}^2 \leq \frac{\Delta_x^2}{2}.$$

The astute reader may recognize $\|x - x^*\|_{\bar{G}_n^{-1}}$ as the leading factor in the width of the confidence interval for estimating the gap Δ_x using a linear least squares estimator. The result says that this width has to shrink at least logarithmically with a specific constant. Before the proof of Theorem 1 we present a trivial corollary and some consequences.

Corollary 2. *Let π be a consistent policy, $\theta \in \mathbb{R}^d$ such that there is a unique optimal arm in $\mathcal{A}$. Then*

$$\limsup_{n\to\infty} \log(n) \|x\|_{\bar{G}_n^{-1}}^2 \leq \frac{\Delta_x^2}{2} \qquad (1)$$

and also $\liminf_{n\to\infty} \dfrac{R_\theta^\pi(n)}{\log(n)} \geq c(\mathcal{A}, \theta),$

where $c(\mathcal{A}, \theta)$ is the solution to the optimisation problem:

$$\inf_{\alpha \in [0,\infty)^{\mathcal{A}}} \sum_{x \in \mathcal{A}^-} \alpha(x) \Delta_x \text{ subject to}$$
$$\|x\|_{H^{-1}(\alpha)}^2 \leq \frac{\Delta_x^2}{2}, \quad \forall x \in \mathcal{A}^-, \qquad (2)$$

where $H(\alpha) = \sum_{x \in \mathcal{A}} \alpha(x) x x^\top$ and $\mathcal{A}^- = \mathcal{A} - \{x^*\}$.

As with the previous result, in (1) the reader may recognize the leading term of the confidence width for estimating the mean reward of x. Unsurprisingly, the width of this confidence interval has to shrink at least as fast as the width of the confidence interval for estimating the gap Δ_x. The intuition underlying the optimisation problem (2) is that no consistent strategy can escape allocating samples so that the gaps of all suboptimal actions are identified with high confidence, while a good strategy will also minimise the regret subject to the identifiability condition. The proof of Corollary 2 is in Appendix B. These results are related to previous analysis in the best-arm identification version of the problem, where the goal is not to minimise regret, but rather to find an optimal arm with as few plays as possible [Soare et al., 2014]. In that paper, however, the optimisation problem has a different form because the objective has changed.

Example 3 (Finite armed bandits)**.** Suppose $k = d$ and $\mathcal{A} = \{e_1, \ldots, e_k\}$ be the standard basis vectors. Then

$$c(\mathcal{A}, \theta) = \sum_{x \in \mathcal{A}: \Delta_x > 0} \frac{2}{\Delta_x},$$

which recovers the lower bound by Lai and Robbins [1985].

Example 4. Let $\alpha > 1$ and $d = 2$ and $\mathcal{A} = \{x_1, x_2, x_3\}$ with $x_1 = (1, 0)$ and $x_2 = (0, 1)$ and $x_3 = (1 - \varepsilon, \alpha\varepsilon)$ and $\theta = (1, 0)$ (see Figure 1). Then $c(\mathcal{A}, \theta) = 2\alpha^2$ for all sufficiently small ε (see Appendix F for the analysis showing this). The example serves to illustrate the interesting fact that $c(\mathcal{A} - \{x_2\}, \theta) = 2\varepsilon^{-1} \gg c(\mathcal{A}, \theta)$, which means that the problem becomes significantly harder if x_2 is removed from the action-set. The reason is that x_1 and x_3 are pointing in nearly the same direction, so learning the difference is very challenging. But determining which of x_1 and x_3 is optimal is easy by playing x_2. So we see that in linear bandits there is a complicated trade-off between information and regret that makes the structure of the optimal strategy more interesting than in the unstructured setting.

The closest prior work to our lower bound is by Komiyama et al. [2015] and Agrawal et al. [1989]. The latter consider stochastic partial monitoring when the reward is part of the observation. In this setting in each round, the learner selects one of finitely many actions and receives an observation from a distribution that depends on the action chosen and an unknown parameter, but is otherwise known. While this model could cover our setting, the results in the paper are developed only for the case when the unknown parameter belongs to a finite set, an assumption that all the results of the paper heavily depend on. Komiyama et al. [2015] on the other hand restricts partial monitoring to the case when the observations belong to a finite set, while the parameter belongs to the unit simplex. While this problem also has a linear structure, their results do not generalize beyond the discrete observation setting.

4 PROOF OF THEOREM 1

We make use of two standard results from information theory. The first is a high probability version of Pinsker's inequality.

Lemma 5. *Let $\mathbb{P}$ and $\mathbb{P}'$ be measures on the same measurable space $(\Omega, \mathcal{F})$. Then for any event $A \in \mathcal{F}$,*

$$\mathbb{P}(A) + \mathbb{P}'(A^c) \geq \frac{1}{2} \exp(-\operatorname{KL}(\mathbb{P}, \mathbb{P}')), \qquad (3)$$

where A^c is the complementary event of A ($A^c = \Omega \setminus A$) and $\operatorname{KL}(\mathbb{P}, \mathbb{P}')$ is the relative entropy between $\mathbb{P}$ and $\mathbb{P}'$, which is defined as $+\infty$, if $\mathbb{P}$ is not absolutely continuous with respect to $\mathbb{P}'$, and is $\int_\Omega d\mathbb{P}(\omega) \log \frac{d\mathbb{P}}{d\mathbb{P}'}(\omega)$ otherwise.

This result follows easily from Lemma 2.6 of Tsybakov [2008]. The second lemma is sometimes called the information processing lemma and shows that the relative entropy between measures on sequences of outcomes for the same algorithm interacting with different bandits can be decomposed in terms of the expected number of times each arm is chosen and the relative entropies of the distributions of the arms. There are many versions of this

result (e.g., Auer et al. [1995] and Gerchinovitz and Lattimore [2016]). To state the result, assume without the loss of generality that the measure space underlying the action-reward sequence $(A_1, Y_1, \ldots, A_n, Y_n)$ is $\Omega_n \doteq (\mathcal{A} \times \mathbb{R})^n$ and A_t and Y_t are the respective coordinate projections: $A_t(a_1, y_1, \ldots, a_n, y_n) = a_t$ and $Y_t(a_1, y_1, \ldots, a_n, y_n) = y_t, 1 \le t \le n$.

Lemma 6. *Let $\mathbb{P}$ and $\mathbb{P}'$ be the probability measures on the sequence $(A_1, Y_1, \ldots, A_n, Y_n) \in \Omega_n$ for a fixed bandit policy π interacting with a linear bandit with standard Gaussian noise and parameters θ and θ' respectively. Under these conditions the KL divergence of $\mathbb{P}$ and $\mathbb{P}'$ can be computed exactly and is given by*

$$\mathrm{KL}(\mathbb{P}, \mathbb{P}') = \frac{1}{2} \sum_{x \in \mathcal{A}} \mathbb{E}[T_x(n)] \langle x, \theta - \theta' \rangle^2, \quad (4)$$

where $\mathbb{E}$ is the expectation operator induced by $\mathbb{P}$.

Proof of Theorem 1. The proof of the first part is deferred to Appendix C. Recall that x^* is the optimal arm, which we assumed to be unique. Let $x \in \mathcal{A}$ be a suboptimal arm (so $\Delta_x > 0$) and $A \subset \Omega_n$ be an event to be chosen later. Rearranging (3) gives $\mathrm{KL}(\mathbb{P}, \mathbb{P}') \ge \log(\frac{1}{2\mathbb{P}(A) + 2\mathbb{P}'(A^c)})$ and recalling that $\bar{G}_n = \mathbb{E}\left[\sum_{t=1}^n A_t A_t^\top\right]$, together with Lemma 6 we get that

$$\frac{1}{2} \|\theta - \theta'\|_{\bar{G}_n}^2 = \mathrm{KL}(\mathbb{P}, \mathbb{P}') \ge \log\left(\frac{1}{2\mathbb{P}(A) + 2\mathbb{P}'(A^c)}\right). \quad (5)$$

Now we choose θ' close to θ, but in a such a way that $\langle x - x^*, \theta' \rangle > 0$, meaning in the bandit determined by θ' the optimal action is not x^*. Selecting $A = \{T_{x^*}(n) \le n/2\}$ ensures that $\mathbb{P}(A) + \mathbb{P}'(A^c)$ is small, because π is consistent. Intuitively, this holds because if $\mathbb{P}(A)$ is large then x^* is not used much in θ, hence $R_n \doteq R_\theta^\pi(n)$ must be large. If $\mathbb{P}'(A^c)$ is large, then x^* is used often in θ', hence $R_n' \doteq R_{\theta'}^\pi(n)$ must be large. But from the consistency of π we know that both R_n and R_n' are sub-polynomial. Let $\varepsilon \in (0, \Delta_{\min})$ and $H \succeq 0$ satisfy $\|x - x^*\|_H > 0$ and define θ' by

$$\theta' = \theta + \frac{H(x - x^*)}{\|x - x^*\|_H^2} (\Delta_x + \varepsilon), \quad (6)$$

Then the sub-optimality gap for x^* in bandit θ' can be bounded by,

$$\langle x - x^*, \theta' \rangle = \langle x - x^*, \theta \rangle + \Delta_x + \varepsilon = \varepsilon > 0. \quad (7)$$

Now we control the regret in terms of the number of times the optimal action is not chosen.

$$R_n = \sum_{x \in \mathcal{A}} \Delta_x \mathbb{E}[T_x(n)] \ge \Delta_{\min} \mathbb{E}[(n - T_*(n))]$$

$$\ge \Delta_{\min} \mathbb{E}\left[\mathbb{1}\{T_*(n) \le n/2\} \frac{n}{2}\right]$$

$$\ge \frac{\varepsilon n}{2} \mathbb{P}(T_*(n) \le n/2).$$

On the other hand, introducing $\Delta_y' = \max_z \langle z - y, \theta' \rangle$ and $\mathbb{E}'$ to denote the expectation operator induced by $\mathbb{P}'$ and using that by (7), x^* is suboptimal in θ', we also have

$$R_n' = \sum_x \Delta_x' \mathbb{E}'[T_x(n)] \ge \Delta_{x^*}' \mathbb{E}'[T_*(n)]$$

$$\ge \varepsilon \mathbb{E}'[\mathbb{1}\{T_*(n) > n/2\} T_*(n)] \ge \frac{\varepsilon n}{2} \mathbb{P}'(T_*(n) > n/2).$$

Adding up the two inequalities we get

$$\frac{2R_n + 2R_n'}{\varepsilon n} \ge \mathbb{P}\left(T_*(n) \le \frac{n}{2}\right) + \mathbb{P}'\left(T_*(n) > \frac{n}{2}\right), \quad (8)$$

which completes the proof that $\mathbb{P}(T_*(n) \le n/2) + \mathbb{P}'(T_*(n) > n/2)$ is indeed small. Now we calculate the term on the left-hand side of (5). Using the definition of θ', we get

$$\frac{1}{2} \|\theta - \theta'\|_{\bar{G}_n}^2 = \frac{(\Delta_x + \varepsilon)^2}{2} \frac{\|x - x^*\|_{H\bar{G}_n H}^2}{\|x - x^*\|_H^4}$$

$$= \frac{(\Delta_x + \varepsilon)^2}{2 \|x - x^*\|_{\bar{G}_n^{-1}}^2} \rho_n(H)$$

where in the last line we introduced

$$\rho_n(H) \doteq \frac{\|x - x^*\|_{\bar{G}_n^{-1}}^2 \|x - x^*\|_{H\bar{G}_n H}^2}{\|x - x^*\|_H^4}.$$

Combining this with (8), (5) and some algebra gives

$$\frac{(\Delta_x + \varepsilon)^2 \rho_n(H)}{2 \log(n) \|x - x^*\|_{\bar{G}_n^{-1}}^2} \ge 1 - \frac{\log(\frac{\varepsilon}{4}) + \log(R_n + R_n')}{\log(n)}. \quad (9)$$

Since π is consistent, $\limsup_{n \to \infty} \frac{\log(R_n + R_n')}{\log(n)} \le 0$. Hence, by making ε arbitrarily small we have

$$\liminf_{n \to \infty} \frac{\rho_n(H)}{\log(n) \|x - x^*\|_{\bar{G}_n^{-1}}^2} \ge \frac{2}{\Delta_x^2}. \quad (10)$$

The proof is finished by way of contradiction. Suppose

$$\limsup_{n \to \infty} \log(n) \|x - x^*\|_{\bar{G}_n^{-1}}^2 > \frac{\Delta_x^2}{2}.$$

Then there exists an $\varepsilon > 0$ and infinite subset $S \subseteq \mathbb{N}$ such that

$$\log(n) \|x - x^*\|_{\bar{G}_n^{-1}}^2 \ge \frac{(\Delta_x + \varepsilon)^2}{2} \text{ for all } n \in S. \quad (11)$$

Thus, by (10), for any $H \succeq 0$ such that $\|x - x^*\|_H > 0$,

$$\liminf_{n \in S} \rho_n(H) > 1. \quad (12)$$

Now choose H as a cluster point of the sequence $\{\bar{G}_n^{-1}/\|\bar{G}_n^{-1}\|\}_{n \in S}$, which exists by the compactness of

matrices with bounded spectral norm. We let $S' \subseteq S$ be a subset of S on which $\bar{G}_n^{-1} / \|\bar{G}_n^{-1}\|$ converges to H. We have to check that $\|x - x^*\|_H > 0$, which follows since

$$
\begin{aligned}
\|x - x^*\|_H^2 &= \lim_{n \in S'} \frac{\|x - x^*\|_{\bar{G}_n^{-1}}^2}{\|\bar{G}_n^{-1}\|} \\
&\geq \lim_{n \in S'} \frac{(\Delta_x + \varepsilon)^2 \lambda_{\min}(\bar{G}_n)}{2 \log(n)} > 0 \,, \quad (13)
\end{aligned}
$$

where in the second last inequality we used the fact that $\|\bar{G}_n^{-1}\| = 1/\lambda_{\min}(\bar{G}_n)$, (11). The last inequality follows from the first part of the theorem. Now we derive a contradiction from (12) and the definition of ρ_n and H.

$$
\begin{aligned}
1 &< \liminf_{n \in S} \rho_n(H) \\
&\leq \liminf_{n \in S'} \frac{\|x - x^*\|_{\bar{G}_n^{-1}}^2 \|x - x^*\|_{H\bar{G}_n H}^2}{\|x - x^*\|_H^4} = 1 \,.
\end{aligned}
$$

Therefore we have a contradiction and so

$$
\limsup_{n \to \infty} \log(n) \|x - x^*\|_{\bar{G}_n^{-1}}^2 \leq \frac{\Delta_x^2}{2} \,. \qquad \square
$$

Remark 7. The uniqueness assumption of the theorem can be lifted at the price of more work and by slightly changing the theorem statement. In particular, the theorem statement must be restricted to those suboptimal actions $x \in \mathcal{A}^-$ that can be made optimal by changing θ to θ', while none of the optimal actions $\mathcal{A}^*(\theta) = \{x \in \mathcal{A} : \langle x, \theta \rangle = \max_{y \in \mathcal{A}} \langle y, \theta \rangle\}$ are optimal. That is, the statement only concerns $x \in \mathcal{A}$ such that $x \notin \mathcal{A}^*(\theta)$ but there exists $\theta' \in \mathbb{R}^d$ such that $\mathcal{A}^*(\theta') \cap \mathcal{A}^*(\theta) = \emptyset$ and $x \in \mathcal{A}^*(\theta')$. The choice of θ' would still be as before, except that x^* is selected as the optimal action under θ that maximizes $c(H, \theta) = \inf_{x' \in \mathcal{A}^*(\theta)} \langle x - x', x - x^* \rangle_H$. Then, in the proof, $T_*(n)$ has to be redefined to be $\sum_{x \in \mathcal{A}^*(\theta)} T_x(n)$ (the total number of times an optimal action is chosen), and at the end one also needs to show that the chosen H satisfies $c(H, \theta) > 0$.

5 CONCENTRATION

Before introducing the new algorithm we analyse the concentration properties of the least squares estimator. Our results refine the existing guarantees by Abbasi-Yadkori et al. [2011], and are necessary in order to obtain asymptotic optimality. Let G_t be the Gram matrix after round t defined by $G_t = \sum_{s \leq t} A_s A_s^\top$ and $\hat{\theta}(t) = G_t^{-1} \sum_{s=1}^t A_s Y_s$ be the empirical (least squares) estimate, where A_s is selected based on $A_1, Y_1, \ldots, A_{s-1}, Y_{s-1}$ and $Y_s = \langle A_s, \theta \rangle + \eta_s$, $\eta_s \sim N(0, 1)$. We will only use $\hat{\theta}(t)$ for rounds t when G_t is invertible. The empirical estimate of the sub-optimality gap for arm x is $\hat{\Delta}_x(t) = \max_{y \in \mathcal{A}} \hat{\mu}_y(t) - \hat{\mu}_x(t)$, where $\hat{\mu}_x(t) = \langle x, \hat{\theta}(t) \rangle$. We will also use the notation $\hat{\mu}(t)$ and $\hat{\Delta}(t) \in \mathbb{R}^k$ for vectors of empirical means and sub-optimality gaps (indexed by the arms).

Theorem 8. *For any $\delta \in [1/n, 1)$, n sufficiently large and $t_0 \in \mathbb{N}$ such that G_{t_0} is almost surely non-singular,*

$$
\mathbb{P}\left(\exists t \geq t_0, x : |\hat{\mu}_x(t) - \mu_x| \geq \sqrt{\|x\|_{G_t^{-1}}^2 f_{n,\delta}}\right) \leq \delta \,,
$$

where for some universal constant $c > 0$,

$$
f_{n,\delta} = 2\left(1 + \frac{1}{\log(n)}\right) \log(1/\delta) + cd \log(d \log(n)) \,.
$$

The result improves on the elegant concentration guarantee of Abbasi-Yadkori et al. [2011] because asymptotically we have $f_{n,1/n} \sim 2d \log(n)$, while there it was $2d \log(n)$. Note that the restriction on δ may be relaxed with a small additional argument. The proof of Theorem 8 relies on a peeling argument and is given in Appendix A. For the remainder we abbreviate $f_n = f_{n,1/n}$ and $g_n = f_{n,1/\log(n)}$, which are chosen so that

$$
\mathbb{P}\left(\exists t \geq t_0, x : |\hat{\mu}_x(t) - \mu_x| \geq \sqrt{\|x\|_{G_t^{-1}}^2 f_n}\right) \leq \frac{1}{n} \,,
$$

$$
\mathbb{P}\left(\exists t \geq t_0, x : |\hat{\mu}_x(t) - \mu_x| \geq \sqrt{\|x\|_{G_t^{-1}}^2 g_n}\right) \leq \frac{1}{\log(n)} \,.
$$
(14)

6 OPTIMAL STRATEGY

A barycentric spanner of the action space is a set $B = \{x_1, \ldots, x_d\} \subseteq \mathcal{A}$ such that for any $x \in \mathcal{A}$ there exists an $\alpha \in [-1, 1]^d$ with $x = \sum_{i=1}^d \alpha_i x_i$. The existence of a barycentric spanner is guaranteed because $\mathcal{A}$ is finite and spans $\mathbb{R}^d$ [Awerbuch and Kleinberg, 2004]. We propose a simple strategy that operates in three phases called the *warm-up* phase, the *success* phase and the *recovery* phase. In the warm-up the algorithm deterministically chooses its actions from a barycentric spanner to obtain a rough estimate of the sub-optimality gaps. The algorithm then uses the estimated gaps as a substitute for the true gaps to determine the optimal pull counts for each action, and starts implementing this strategy. Finally, if an anomaly is detected that indicates the inaccuracy of the estimated gaps then the algorithm switches to the recovery phase where it simply plays UCB.

Definition 9. Assume for this definition that $0 \times \infty = 0$. For any $\Delta \in [0, \infty)^k$ define $T_n(\Delta) \in [0, \infty]^k$ to be a solution to the optimisation problem

$$
\min_{T \in [0,\infty]^k} \sum_{x \in \mathcal{A}} T_x \Delta_x \quad \text{subject to}
$$

$$
\|x\|_{H_T^{-1}}^2 \leq \frac{\Delta_x^2}{f_n} \quad \text{for all } x \in \mathcal{A}, \quad \text{where } H_T = \sum_{x \in \mathcal{A}} T_x x x^\top
$$

and $H_T^{-1} \doteq \lim_{u \to \infty} H_{T \wedge u}^{-1}$ with $(T \wedge u)_x = \min\{T_x, u\}$.

The optimisation problem is convex (proof in Appendix G), so in principle there is hope for an efficient solution. The problem is challenging when k is large, but one might hope that most constraints are easily satisfied. We leave the question of how to solve this optimisation problem in practice for another day.

Algorithm 1 Optimal Algorithm

1: **Input:** $\mathcal{A}$ and n
2: // Warmup phase
3: Find a barycentric spanner: $B = \{x_1, \dots, x_d\}$
4: Choose each arm in B exactly $\lceil \log^{1/2}(n) \rceil$ times
5: // Success phase
6: $\varepsilon_n \leftarrow \max_{x \in \mathcal{A}} \|x\|_{G_{t-1}^{-1}} \, g_n^{1/2}$
7: $\hat{\Delta} \leftarrow \hat{\Delta}(t-1)$ and $\hat{T} \leftarrow T_n(\hat{\Delta})$ and $\hat{\mu} \leftarrow \hat{\mu}(t-1)$
8: **while** $t \le n$ and $\|\hat{\mu} - \hat{\mu}(t-1)\|_\infty \le 2\varepsilon_n$ **do**
9:     Play actions x in a round-robin fashion with $T_x(t) \le \hat{T}_x$, $t \leftarrow t+1$
10: **end while**
11: // Recovery phase
12: Discard all data and play UCB until $t = n$.

Theorem 10. *Assuming that x^* is unique, the strategy given in Algorithm 1 satisfies*

$$\limsup_{n \to \infty} \frac{R_\theta^\pi(n)}{\log(n)} \le c(\mathcal{A}, \theta) \text{ for all } \theta \in \mathbb{R}^d \,.$$

7 PROOF OF THEOREM 10

We analyse the regret in each of the three phases. The warm-up phase has length $d\lceil \log^{1/2}(n) \rceil$, so its contribution to the asymptotic regret is negligible. There are two challenges. The first is to show that the recovery phase happens with probability at most $1/\log(n)$. Then, since the regret in the recovery phase is logarithmic by known results for UCB, this ensures that the expected regret incurred in the recovery phase is also negligible. The second challenge is to show that the expected regret incurred during the success phase is asymptotically matching the lower bound in Theorem 1. The set of rounds when the algorithm is in the warm-up/success/recovery phases are denoted by $T_{\text{warm.}}$, $T_{\text{succ.}}$ and $T_{\text{rec.}}$ respectively. We introduce two failure events that occur when the errors in the empirical estimates of the arms are excessively large. Let F_n be the event that there exists an arm x and round $t \ge d$ such that

$$|\hat{\mu}_x(t) - \mu_x| \ge \sqrt{\|x\|_{G_t^{-1}}^2 \, g_n} \,.$$

Similarly, let F_n' be the event that there exists an arm x and round $t \ge d$ such that

$$|\hat{\mu}_x(t) - \mu_x| \ge \sqrt{\|x\|_{G_t^{-1}}^2 \, f_n} \,.$$

Theorem 8 with $t_0 = d$ and (14) imply that $\mathbb{P}(F_n) \le 1/\log(n)$ and $\mathbb{P}(F_n') \le 1/n$. The failure events determine the quality of the estimates throughout time. The following two lemmas show that if F_n does not occur then the regret is asymptotically optimal, while if F_n' occurs then the regret is logarithmic with some constant factor that depends only on the problem (determined by the action set $\mathcal{A}$ and the parameter θ). Since F_n' occurs with probability at most $1/n$, the contribution of the latter component is negligible asymptotically.

Lemma 11. *If F_n does not occur then Algorithm 1 never enters the recovery phase. Furthermore,*

$$\limsup_{n \to \infty} \mathbb{E}\left[\frac{\mathbb{1}\{\text{not } F_n\} \sum_{t \in T_{\text{succ.}}} \Delta_{A_t}}{\log(n)} \right] \le c(\mathcal{A}, \theta) \,.$$

Before proving Lemma 11 we need a naive bound on the solution to the optimisation problem, the proof of which is given in Appendix D.

Lemma 12. *Let $T = T_n(\Delta)$ for any n. Then*

$$\sum_{x : \Delta_x > 0} T_x \le 2d^3 f_n \Delta_{\max} / \Delta_{\min}^3 \,.$$

Proof of Lemma 11. First, if $t = d\lceil \log^{1/2}(n) \rceil$ is the round at the end of the warm-up period then by the definition of the algorithm there is a barycentric spanner $B = \{x_1, \dots, x_d\}$ and $T_{x_i}(t) = \lceil \log^{1/2}(n) \rceil$ for $1 \le i \le d$. Let $x \in \mathcal{A}$ be arbitrary. Then, by the definition of the barycentric spanner, we can write $x = \sum_{i=1}^d \alpha_i x_i$ where $\alpha_i \in [-1, 1]$ for all i. Therefore,

$$\|x\|_{G_t^{-1}} \le \sum_{i=1}^d \|x_i\|_{G_t^{-1}} \le \frac{d}{\log^{1/4}(n)} \,.$$

Recalling the definition of ε_n in the algorithm we have

$$\varepsilon_n = \max_{x \in \mathcal{A}} \|x\|_{G_t^{-1}} \sqrt{g_n} = O\left(\frac{d \log^{1/2}(\log(n))}{\log^{1/4}(n)} \right) \,.$$

Consider the case when F_n does not hold. Then, for all arms x and rounds t after the warm-up period we have $|\hat{\mu}_x(t) - \mu_x| \le \|x\|_{G_t^{-1}} \sqrt{g_n} \le \varepsilon_n$. Therefore for all s, t after the warm-up period we have $|\hat{\mu}_x(t) - \hat{\mu}_x(s)| \le 2\varepsilon_n$, which means the success phase never ends and so the first part of the lemma is proven. It remains to bound the regret. Since we are only concerned with the asymptotics we may take n to be large enough so that $2\varepsilon_n \le \Delta_{\min}/2$, which implies that $\hat{\Delta}_{x^*} = 0$. For $T_n(\Delta)$, the solution to the optimisation problem in Def. 9 with the true gaps, we have

$$\limsup_{n \to \infty} \frac{\sum_{x \ne x^*} T_{n,x}(\Delta)\Delta_x}{\log(n)} = c(\mathcal{A}, \theta) \,. \tag{15}$$

Letting $T^* = T_n(\Delta)$ and $1 + \delta_n = \max_{x:\hat{\Delta}_x>0} \Delta_x^2/\hat{\Delta}_x^2$,

$$\|x\|^2_{H^{-1}_{(1+\delta_n)T^*}} = \frac{\|x\|^2_{H^{-1}_{T^*}}}{1+\delta_n} \leq \frac{\Delta_x^2}{(1+\delta_n)f_n} \leq \frac{\hat{\Delta}_x^2}{f_n} \,.$$

Therefore, $\sum_{x\neq x^*} T_x\hat{\Delta}_x \leq (1+\delta_n)\sum_{x\neq x^*} T_x^*\Delta_x$, where $T \doteq (T_x)_x \doteq T_x(n)$. Also,

$$\begin{aligned}
1+\delta_n &= \max_{x:\hat{\Delta}_x>0} \frac{\Delta_x^2}{\hat{\Delta}_x^2} \leq \max_{x:\hat{\Delta}_x>0} \frac{\Delta_x^2}{(\Delta_x-2\varepsilon_n)^2} \\
&= \max_{x:\hat{\Delta}_x>0}\left(1+\frac{4(\Delta_x-\varepsilon_n)\varepsilon_n}{(\Delta_x-2\varepsilon_n)^2}\right) \leq 1+\frac{16\varepsilon_n}{\Delta_{\min}}\,, \quad (16)
\end{aligned}$$

where in the last inequality we used the fact that $0 \leq 2\varepsilon_n \leq \Delta_{\min}/2$. Then the regret in the success phase is

$$\begin{aligned}
\sum_{t\in T_{\text{succ.}}} \Delta_{A_t} &\leq \sum_{x\neq x^*} T_x\Delta_x \\
&= \sum_{x\neq x^*} T_x\hat{\Delta}_x + \sum_{x\neq x^*} T_x(\Delta_x-\hat{\Delta}_x) \\
&\leq (1+\delta_n)\sum_{x\neq x^*} T_x^*\hat{\Delta}_x + 2\varepsilon_n\sum_{x\neq x^*} T_x \\
&\leq (1+\delta_n)\sum_{x\neq x^*} T_x^*\Delta_x + 2\varepsilon_n\sum_{x\neq x^*}((1+\delta_n)T_x^*+T_x)\,.
\end{aligned}$$

The result follows by taking the limit as n tends to infinity and from Lemma 12 and (15) and (16), together with the reverse Fatou lemma. $\qquad\square$

Our second lemma shows that provided F_n' fails, the regret in the success phase is at most logarithmic:

Lemma 13. *It holds that:*

$$\limsup_{n\to\infty} \frac{\mathbb{E}\left[\mathbb{1}\{F_n \text{ and not } F_n'\}\sum_{t\in T_{succ.}}\Delta_{A_t}\right]}{\log(n)} = 0\,.$$

The proof follows by showing the existence of a constant m that depends on $\mathcal{A}$ and θ, but not n such that the regret suffered in the success phase whenever F_n' does not hold is almost surely at most $m\log(n)$. The result follows from this because $\mathbb{P}(F_n) \leq 1/\log(n)$. Details in Appendix E.

Proof of Theorem 10. We decompose the regret into the regret suffered in each of the phases:

$$R_\theta^\pi(n) = \mathbb{E}\left[\sum_{t\in T_{\text{warm.}}}\Delta_{A_t} + \sum_{t\in T_{\text{succ.}}}\Delta_{A_t} + \sum_{t\in T_{\text{rec.}}}\Delta_{A_t}\right]. \quad (17)$$

The warm-up phase has length $d\lceil\log^{1/2}(n)\rceil$, which contributes asymptotically negligibly to the regret:

$$\limsup_{n\to\infty} \frac{\mathbb{E}\left[\sum_{t\in T_{\text{warm.}}}\Delta_{A_t}\right]}{\log(n)} = 0\,. \quad (18)$$

By Lemma 11, the recovery phase only occurs if F_n occurs and $\mathbb{P}(F_n) \leq 1/\log(n)$. Therefore by well-known guarantees for UCB [Bubeck and Cesa-Bianchi, 2012] there exists a universal constant $c > 0$ such that

$$\begin{aligned}
\mathbb{E}\left[\sum_{t\in T_{\text{rec.}}}\Delta_{A_t}\right] &= \mathbb{E}\left[\sum_{t\in T_{\text{rec.}}}\Delta_{A_t}\middle| T_{\text{rec.}} \neq \emptyset\right]\mathbb{P}(T_{\text{rec.}} \neq \emptyset) \\
&\leq \frac{ck\log(n)}{\Delta_{\min}}\mathbb{P}(T_{\text{rec.}} \neq \emptyset) \leq \frac{ck}{\Delta_{\min}}\,.
\end{aligned}$$

Therefore $\displaystyle\limsup_{n\to\infty} \frac{\mathbb{E}\left[\sum_{t\in T_{\text{rec.}}}\Delta_t\right]}{\log(n)} = 0\,.$ $\quad (19)$

Finally we use the previous lemmas to analyse the regret in the success phase:

$$\begin{aligned}
\mathbb{E}\left[\sum_{t\in T_{\text{succ.}}}\Delta_{A_t}\right] &= \mathbb{E}\left[\mathbb{1}\{\text{not } F_n\}\sum_{t\in T_{\text{succ.}}}\Delta_{A_t}\right] \\
&+ \mathbb{E}\left[\mathbb{1}\{F_n \text{ and not } F_n'\}\sum_{t\in T_{\text{succ.}}}\Delta_{A_t}\right] \\
&+ \mathbb{E}\left[\mathbb{1}\{F_n'\}\sum_{t\in T_{\text{succ.}}}\Delta_{A_t}\right]. \quad (20)
\end{aligned}$$

By (14), the last term satisfies

$$\begin{aligned}
&\limsup_{n\to\infty} \frac{\mathbb{E}\left[\mathbb{1}\{F_n'\}\sum_{t\in T_{\text{succ.}}}\Delta_{A_t}\right]}{\log(n)} \\
&\leq \limsup_{n\to\infty} \frac{n\Delta_{\max}\mathbb{P}(F_n')}{\log(n)} = 0\,.
\end{aligned}$$

The first two terms in (20) are bounded using Lemmas 11 and 13, leading to

$$\limsup_{n\to\infty} \frac{\mathbb{E}\left[\sum_{t\in T_{\text{succ.}}}\Delta_{A_t}\right]}{\log(n)} \leq c(\mathcal{A},\theta)\,.$$

Substituting the above display together with (18) and (19) into (17) completes the result. $\qquad\square$

8 SUB-OPTIMALITY OF OPTIMISM AND THOMPSON SAMPLING

We now argue that algorithms based on optimism or Thompson sampling cannot be close to asymptotically optimal. In each round t an optimistic algorithm constructs a confidence set $\mathcal{C}_t \subseteq \mathbb{R}^d$ and chooses A_t according to $A_t = \arg\max_{x\in\mathcal{A}}\max_{\tilde{\theta}\in\mathcal{C}_{t-1}}\langle x,\tilde{\theta}\rangle$. In order to proceed we need to make some assumptions on $\mathcal{C}_t$, otherwise one can define a "confidence set" to ensure any behaviour at all. First of all, we will assume that $\mathbb{P}(\exists t \leq n : \theta \notin \mathcal{C}_t) = O(1/n)$. That is, the probability that the true parameter is ever outside the confidence set is not too large. Second, we

assume that $\mathcal{C}_t \subseteq \mathcal{E}_t$ where $\mathcal{E}_t$ is the ellipsoid around the least squares estimator given by

$$\mathcal{E}_t = \left\{ \tilde{\theta} : \|\hat{\theta}(t) - \tilde{\theta}\|_{G_t}^2 \leq \alpha \log(n) \right\},$$

where α is some constant and $\hat{\theta}(t)$ is the empirical estimate of θ based on the observations so far. Existing algorithms based on confidence all use such confidence sets. Standard wisdom when designing optimistic algorithms is to use the smallest confidence set possible, so an alternative algorithm that used a different form of confidence set would normally be advised to use the intersection $\mathcal{C}_t \cap \mathcal{E}_t$, which remains valid with high probability by a union bound. If the optimistic algorithm is not consistent, then its regret is not logarithmic on some problem and so diverges relative to the optimal strategy. Suppose now that the algorithm is consistent. Then we design a bandit on which its asymptotic regret is worse than optimal by an arbitrarily large constant factor. The following counter-example is the same as used by Soare et al. [2014] for the best arm identification version of the problem. Let $d = 2$ and $e_1 = (1, 0)$ and $e_2 = (0, 1)$ be the standard basis vectors. The counter-example (illustrated in Figure 1) is very simple with $\mathcal{A} = \{e_1, e_2, x\}$ where $x = (1 - \varepsilon, 8\alpha\varepsilon)$. The true parameter is given by $\theta = e_1$, which means that $x^* = e_1$ and $\Delta_{e_2} = 1$ and $\Delta_x = \varepsilon$. Suppose a consistent optimistic algorithm has chosen $T_{e_2}(t - 1) \geq 4\alpha \log(n)$ and that $\theta \in \mathcal{C}_t$. Using the definition of the confidence interval and the fact that $\langle e_2, \theta \rangle = 0$ we have

$$\max_{\tilde{\theta} \in C_{t-1}} \langle e_2, \tilde{\theta} \rangle \leq \langle e_2, \hat{\theta}(t-1) \rangle + \sqrt{\|e_2\|_{G_{t-1}^{-1}}^2 \alpha \log(n)}$$

$$< 2\sqrt{\|e_2\|_{G_{t-1}^{-1}}^2 \alpha \log(n)} \leq 1.$$

But because $\theta \in \mathcal{C}_t$, the optimistic value of the optimal action is at least $\langle e_1, \theta \rangle = 1$, which means that $A_t \neq e_2$. We conclude that if $\theta \in \mathcal{C}_t$ for all rounds, then the optimistic algorithm satisfies $T_{e_2}(t - 1) \leq 1 + 4\alpha \log(n)$. By the assumption that $\theta \in \mathcal{C}_t$ with probability at least $1 - 1/n$ we bound $\mathbb{E}[T_{e_2}(n)] \leq 2 + 4\alpha \log(n)$. By consistency of the optimistic algorithm and our lower bound (Theorem 1) we have

$$\limsup_{n \to \infty} \log(n) \|x - e_1\|_{\bar{G}_n^{-1}}^2 \leq \frac{\varepsilon^2}{2}.$$

Therefore by choosing ε sufficiently small we conclude that $\limsup_{n \to \infty} \mathbb{E}[T_x(n)]/\log(n) = \Omega(1/\varepsilon^2)$ and so the asymptotic regret of the optimistic algorithm is at least

$$\limsup_{n \to \infty} \frac{R_\theta^{\text{OPTIMISTIC}}(n)}{\log(n)} = \Omega\left(\frac{1}{\varepsilon}\right).$$

However, for small ε the optimal regret for this problem is $c(\mathcal{A}, \theta) \log(n) = 128\alpha^2 \log(n)$ (as follows from the calculation in Appendix F) and so by choosing $\varepsilon \ll \alpha$ we

can see that the optimistic approach is sub-optimal by an arbitrarily large constant factor. The intuition is that the optimistic algorithms very quickly learn that e_2 is a sub-optimal arm and stop playing it. But as it turns out, the information gained by choosing e_2 is sufficiently valuable that an optimal algorithm should use it for exploration.

Thompson sampling has also been proposed for the linear bandit problem [Agrawal and Goyal, 2013]. The standard approach uses a nearly flat Gaussian prior (and so posterior), which means that essentially the algorithm operates by sampling θ_t from $\mathcal{N}(\hat{\mu}(t), \alpha G_t^{-1})$
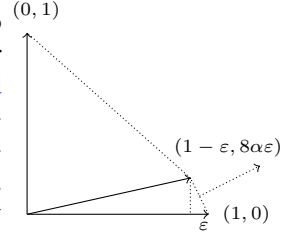

Figure 1: Counter-example

and choosing the arm $A_t = \arg\max_{x \in \mathcal{A}} \langle x, \theta_t \rangle$. Why does this approach fail? Very briefly, by the assumption of consistency we expect that the optimal arm will be played all but logarithmically often, which means that the posterior will concentrate quickly about the value of the optimal action so that $\langle x^*, \theta_t \rangle \approx \mu^*$. Then using the same counter-example as for the optimistic algorithm we see that the likelihood that $\langle e_2 - e_1, \theta_t \rangle \geq 0$ is vanishingly small once $T_{e_2}(t - 1) = \Omega(\alpha \log(n))$ and so Thompson sampling will also fail to sample action e_2 sufficiently often.

9 SUMMARY

We characterised the optimal asymptotic regret for linear bandits with Gaussian noise and finitely many actions in the sense of Lai and Robbins [1985]. The results highlight a surprising fact that reasonable algorithms based on optimism can be arbitrarily worse than optimal. While this behaviour has been observed before (notably, in partial monitoring or hand-crafted counter-examples), our results are the first to illustrate this issue in a popular setting only barely more complicated than finite-armed bandits. Besides this we improve the self-normalised concentration guarantees by Abbasi-Yadkori et al. [2011] by a factor of d asymptotically. As usual, we open more questions than we answer. While the proposed strategy is asymptotically optimal, it is also extraordinarily naive and the analysis is far from showing finite-time optimality. For this reason we think the most pressing task is to develop efficient and practical algorithms that exploit the structure of the problem in a way that Thompson sampling and optimism do not. There are two natural research directions towards this goal. The first is to push the optimisation approach used here and also by Wu et al. [2015], but applied more "smoothly" without discarding data or long phases. The second is to generalise information-theoretic ideas used (for instance) by Russo and Van Roy [2014] or Reddy et al. [2016].

Acknowledgment

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2312–2320, 2011.

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Online-to-confidence-set conversions and application to sparse stochastic bandits. In *AISTATS*, pages 1–9, 2012.

Rajeev Agrawal. Sample mean based index policies with $O(\log n)$ regret for the multi-armed bandit problem. *Advances in Applied Probability*, pages 1054–1078, 1995.

Rajeev Agrawal, Demosthenis Teneketzis, and Venkatachalam Anantharam. Asymptotically efficient adaptive allocation schemes for controlled i.i.d. processes: Finite parameter space. *IEEE Transaction on Automatic Control*, 34:258–267, 1989.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *ICML*, pages 127–135, 2013.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *Proceedings of Conference on Learning Theory (COLT)*, pages 217–226, 2009.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

Peter Auer, Nicoló Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of the 36th Annual Symposium on Foundations of Computer Science*, pages 322–331, 1995.

Peter Auer, Nicoló Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 99:1563–1600, August 2010. ISSN 1532-4435.

Baruch Awerbuch and Robert D Kleinberg. Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches. In *Proceedings of the 36th Annual ACM Symposium on the Theory of Computing*, pages 45–53, 2004.

Gábor Bartók, Dean P Foster, Dávid Pál, Alexander Rakhlin, and Csaba Szepesvári. Partial monitoring-classification, regret bounds, and algorithms. *Mathematics of Operations Research*, 39(4):967–997, 2014.

Sébastien Bubeck and Nicolò Cesa-Bianchi. *Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems*. Foundations and Trends in Machine Learning. Now Publishers Incorporated, 2012. ISBN 9781601986269.

Olivier Chapelle and Lihong Li. An empirical evaluation of Thompson sampling. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2249–2257, 2011.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. In *Proceedings of Conference on Learning Theory (COLT)*, pages 355–366, 2008.

Sarah Filippi, Olivier Cappé, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In *NIPS*, pages 586–594, December 2010.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory (COLT)*, 2016.

Sébastien Gerchinovitz and Tor Lattimore. Refined lower bounds for adversarial bandits. *arXiv preprint arXiv:1605.07416*, 2016.

Aditya Gopalan and Shie Mannor. Thompson sampling for learning parameterized Markov decision processes. In *Proceedings of the 28th Conference on Learning Theory (COLT)*, pages 861–898, 2015.

Michael N Katehakis and Herbert Robbins. Sequential choice from several populations. *Proceedings of the National Academy of Sciences of the United States of America*, 92(19):8584, 1995.

Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In Nader H. Bshouty, Gilles Stoltz, Nicolas Vayatis, and Thomas Zeugmann, editors, *Algorithmic Learning Theory*, pages 199–213, 2012.

Junpei Komiyama, Junya Honda, and Hiroshi Nakagawa. Regret lower bound and optimal algorithm in finite stochastic partial monitoring. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1792–1800, 2015.

Nathaniel Korda, Emilie Kaufmann, and Rémi Munos. Thompson sampling for 1-dimensional exponential family bandits. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1448–1456, 2013.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

Gautam Reddy, Antonio Celani, and Massimo Vergassola. Infomax strategies for an optimal balance between exploration and exploitation. *Journal of Statistical Physics*, 163(6):1454–1476, 2016.

Omar Rivasplata. Subgaussian random variables: An expository note. Arxiv preprint, 2012.

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

Dan Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. In *Advances in Neural Information Processing Systems*, pages 1583–1591, 2014.

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

William Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

Alexandre B Tsybakov. *Introduction to nonparametric estimation.* Springer Science & Business Media, 2008.

Michal Valko, Rémi Munos, Branislav Kveton, and Tomas Kocak. Spectral bandits for smooth graph functions. In *ICML*, pages 46–54, 2014.

Yifan Wu, András György, and Csaba Szepesvári. Online Learning with Gaussian Payoffs and Side Observations. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1360–1368, 2015.

A PROOF OF THEOREM 8

Recall that A_t is the action chosen in round t and that $\eta_t = Y_t - \langle A_t, \theta \rangle$ is the noise term, which we assumed to be a standard Gaussian. Let $S_t = \sum_{s=1}^{t} A_s \eta_s$. By assumption, $\|A_t\| \leq 1$ for all $t \geq 1$.

Lemma 14. *Let* $n \in \mathbb{N}$ *and* $\varepsilon > 0$ *and* $\sigma^2 > 0$. *Let* $X_1, X_2, \ldots, X_n$ *be a sequence of Gaussian random variables adapted to filtration* $\mathcal{F}_1, \mathcal{F}_2, \ldots$ *such that* $\mathbb{E}[X_t | \mathcal{F}_{t-1}] = 0$. *Define* $\sigma_t^2 = \mathrm{Var}[X_t | \mathcal{F}_{t-1}]$ *and assume that* $\sigma_t^2 \leq \sigma^2$ *almost surely. Then*

$$\mathbb{P}\left(\exists t \leq n : \sum_{s=1}^{t} X_s \geq \sqrt{2\gamma_n V_t \log\left(\frac{N}{\delta}\right)}\right) \leq \delta,$$

where $V_t = \max\left\{\varepsilon, \sum_{s=1}^{t} \sigma_t^2\right\}$ *and*

$$\gamma_n = 1 + \frac{1}{\log(n)} \quad and \quad N = 1 + \left\lceil \frac{\log(n\sigma^2/\varepsilon)}{\log(\gamma_n)} \right\rceil.$$

Proof. For $\psi \in \mathbb{R}$ define

$$M_{t,\psi} = \exp\left(\sum_{s=1}^{t} \psi X_t - \frac{\psi^2 \sigma_t^2}{2}\right).$$

If $\tau \leq n$ is a stopping time with respect to $\mathcal{F}$, then as in the proof [Abbasi-Yadkori et al., 2011, Lemma 8] we have $\mathbb{E}[M_{\tau,\psi}] \leq 1$. Therefore, by Markov's inequality we have

$$\mathbb{P}\left(M_{\tau,\psi} \geq 1/\delta\right) \leq \delta. \tag{21}$$

For $k \in \{1, 2, \ldots, N\}$ define

$$\psi_k = \sqrt{\frac{2}{\varepsilon \gamma_n^{k-1}} \log\left(\frac{N}{\delta}\right)}$$

Then rearranging (21) leads to

$$\mathbb{P}\left(\exists k \in [N] : \sum_{t=1}^{\tau} X_t \geq \frac{1}{\psi_k} \log\left(\frac{N}{\delta}\right) + \frac{\psi_k V_\tau}{2}\right) \leq \delta.$$

Therefore letting

$$k^* = \min\left\{k \in [N] : \psi_k \geq \sqrt{2\log(N/\delta)/V_\tau}\right\}$$

leads to

$$\delta \geq \mathbb{P}\left(\sum_{t=1}^{\tau} X_t \geq \frac{1}{\psi_{k^*}} \log\left(\frac{N}{\delta}\right) + \frac{\psi_{k^*} V_\tau}{2}\right)$$

$$\geq \mathbb{P}\left(\sum_{t=1}^{\tau} X_t \geq \sqrt{2\gamma_n V_\tau \log\left(\frac{N}{\delta}\right)}\right).$$

The result is completed by choosing stopping time τ by $\tau = \min(n, \tau_n)$, where

$$\tau_n = \min\left\{t \leq n : \sum_{s=1}^{t} X_s \geq \sqrt{2\gamma_n V_t \log\left(\frac{N}{\delta}\right)}\right\}.$$
□

Lemma 15. *Let* $\delta \in [1/n, 1)$ *and* $\lambda \in \mathbb{R}^d$ *with* $\|\lambda\| \leq 1$. *Then*

$$\mathbb{P}\left(\exists t \leq n : \langle \lambda, S_t \rangle \geq \sqrt{\frac{1}{n^2} \vee \|\lambda\|_{G_t}^2 \, h_{n,\delta}}\right) \leq \delta,$$

where

$$h_{n,\delta} = 2\left(1 + \frac{1}{\log(n)}\right) \log\left(\frac{c\log(n)}{\delta}\right)$$

with some universal constant $c \geq 1$.

Proof. We prepare to use the previous lemma. First note that

$$\langle \lambda, S_t \rangle = \sum_{s=1}^{t} \eta_s \langle \lambda, A_t \rangle.$$

Since η_s is a standard Gaussian, the predictable variance of the term inside the sum is $\sigma_t^2 = \langle \lambda, A_t \rangle^2 \leq \|\lambda\|^2 \|A_t\|^2 \leq 1$. Therefore

$$\sum_{s=1}^{t} \sigma_s^2 = \lambda^\top \sum_{s=1}^{t} A_s A_s^\top \lambda = \|\lambda\|_{G_t}^2.$$

Therefore the result follows by the previous lemma with $X_t = \eta_t \langle \lambda, A_t \rangle$ and $\varepsilon = 1/(n^2 \log(n)^3)$ and $\sigma^2 = 1$. □

The following lemma can be extracted from the proof of Theorem 1 in Abbasi-Yadkori et al. [2011].

Lemma 16. *Assume that $\{A_s\}$ is such that for some $t_0 > 0$, G_{t_0} is non-singular almost surely. Then, for some $c > 0$ universal constant,*

$$\mathbb{P}\left(\exists t \geq t_0 : \|S_t\|^2_{G_t^{-1}} \geq cd\log(n/\delta)\right) \leq \delta.$$

Proof of Theorem 8. Let $\varepsilon > 0$ be some small real number to be tuned subsequently and choose $\mathcal{C} \subset \mathbb{R}^d$ to be a finite non-random covering set such that for all $x \in \mathcal{A}$ and t with G_t non-singular there exists a $\lambda \in \mathcal{C}$ such that $\lambda = (I + \mathcal{E})G_t^{-1}x$, where $\mathcal{E}$ is some diagonal matrix (possibly depending on x and G_t^{-1}) with entries bound in $[0, \varepsilon]$. Of course G_t is a random variable, so we insist the existence of λ is almost sure (that is, no matter how the actions are taken). We defer calculating the necessary size $N = |\mathcal{C}|$ until later. Let $\delta_1 = \delta/(N + 1)$ and F_λ be the event that

$$F_\lambda = \left\{ \exists t : \langle \lambda, S_t \rangle \geq \sqrt{\frac{1}{n^2} \vee \|\lambda\|^2_{G_t} \, h_{n,\delta_1}} \right\}.$$

Then a union bound and Lemma 15 leads to

$$\mathbb{P}\left(\cup_{\lambda \in \mathcal{C}} F_\lambda\right) \leq N\delta_1. \tag{22}$$

By Lemma 16, for $\mathcal{G} = \{\exists t \geq t_0 : \|S_t\|^2_{G_t^{-1}} \geq cd\log(n/\delta_1)\}$, we have

$$\mathbb{P}(\mathcal{G}) \leq \delta_1. \tag{23}$$

Another union bound shows that the $\mathbb{P}(\cup_{\lambda \in \mathcal{C}} F_\lambda \cup \mathcal{G}) \leq (N + 1)\delta_1 = \delta$. From now on we assume that neither $\mathcal{F} \doteq \cup_{\lambda \in \mathcal{C}} F_\lambda$, nor $\mathcal{G}$ occurs and let $x \in \mathcal{A}$ be arbitrary and for $t \geq t_0$ let $\lambda \in \mathcal{C}$ be such that $\lambda = (I + \mathcal{E})G_t^{-1}x$ where $\mathcal{E}$ is diagonal with entries in $[0, \varepsilon]$. Then

$$\hat{\mu}_x(t) - \mu_x = \langle G_t^{-1}x, S_t \rangle \tag{24}$$
$$= \langle G_t^{-1}x - \lambda, S_t \rangle + \langle \lambda, S_t \rangle$$
$$\leq \|G_t^{-1}x - \lambda\|_{G_t} \|S_t\|_{G_t^{-1}} + \sqrt{\frac{1}{n^2} \vee \|\lambda\|^2_{G_t} \, h_{n,\delta_1}}. \tag{25}$$

We bound each term separately using matrix algebra and the assumption that the failure events $\mathcal{F}$ and $\mathcal{G}$ do not occur:

$$\|G_t^{-1}x - \lambda\|_{G_t} = \|\mathcal{E}G_t^{-1}x\|_{G_t}$$
$$= \|G_t^{1/2}\mathcal{E}G_t^{-1/2}G_t^{-1/2}x\|$$
$$\leq \|G_t^{1/2}\mathcal{E}G_t^{-1/2}\|_F \|x\|_{G_t^{-1}},$$

where $\|\cdot\|_F$ is the Frobenius norm. Then

$$\|G_t^{1/2}\mathcal{E}G_t^{-1/2}\|_F = \sqrt{\mathrm{tr}(G_t\mathcal{E}G_t^{-1}\mathcal{E})}$$
$$\leq \sqrt{d}\|\mathcal{E}\|_\infty \leq \varepsilon\sqrt{d}.$$

Therefore if $\varepsilon = 1/(d^{3/2}\log(n))$, then the first term in (25) is bounded by

$$\|G_t^{-1}x - \lambda\|_{G_t} \|S_t\|_{G_t^{-1}} = O(1) \cdot \|x\|_{G_t^{-1}}. \tag{26}$$

For the second term we proceed similarly:

$$\|\lambda\|^2_{G_t} = \|G_t^{-1}x + \mathcal{E}G_t^{-1}x\|^2_{G_t}$$
$$\leq \|x\|^2_{G_t^{-1}}\left(1 + \varepsilon\sqrt{d}\right)^2$$
$$= (1 + o(1))\|x\|^2_{G_t^{-1}}.$$

Therefore, assuming n is large enough so that $1/n^2 \leq \|x\|/n \leq \|x\|^2_{G_t^{-1}}$ (in the unique case that $\|x\| = 0$ we simply note that the following equality holds trivially), we have

$$\sqrt{\frac{1}{n^2} \vee \|\lambda\|^2_{G_t} \, h_{n,\delta_1}} = (1 + o(1))\sqrt{\|x\|^2_{G_t^{-1}} \, h_{n,\delta_1}}.$$

Substituting the above expression along with (26) into (25) leads to

$$\hat{\mu}_x(t) - \mu_x = (1 + o(1))\sqrt{\|x\|^2_{G_t^{-1}} \, h_{n,\delta_1}}.$$

Finally we note that $\mathcal{C}$ can be chosen in such a way that for suitably large universal constant $c > 0$ its size is $\log N = O(d\log d\log(n))$. This follows by treating each arm $x \in \mathcal{A}$ separately and noting that $\|x\|/n \leq \|G_t^{-1}x\| \leq \|x\|$ almost surely. Then letting $J = \lceil\log(n)/\log(1 + \varepsilon)\rceil = O(d^{3/2}\log^2(n))$, the covering set is given by $\mathcal{C} = \bigcup_{x \in \mathcal{A}} \mathcal{C}_x$ where $\mathcal{C}_x$ is a product covering space with a geometrical grid.

$$\mathcal{C}_x = \bigtimes_{i=1}^d \left\{ \frac{\|x\|(1 + \varepsilon)^j}{n} : 0 \leq j \leq J \right\}.$$

The theorem is completed by using the definition of h_{n,δ_1} in Lemma 15. $\qquad\square$

B PROOF OF COROLLARY 2

Fix $x \in \mathcal{A}$ a suboptimal action. It suffices to prove that $\limsup_{n\to\infty} \log^{1/2}(n)\|x\|_{\bar{G}_n^{-1}} \leq \Delta_x/\sqrt{2}$. By the triangle inequality we have

$$\|x\|_{\bar{G}_n^{-1}} \leq \|x - x^*\|_{\bar{G}_n^{-1}} + \|x^*\|_{\bar{G}_n^{-1}}$$
$$\leq \|x - x^*\|_{\bar{G}_n^{-1}} + \frac{\|x^*\|}{\sqrt{\mathbb{E}[T_{x^*}(n)]}}, \tag{27}$$

where the last inequality follows since the expected Gram matrix satisfies $\bar{G}_n \succeq \mathbb{E}[T_{x^*}(n)]x^*(x^*)^\top$, which implies that

$$\|x^*\|^2_{\bar{G}_n^{-1}} \leq (x^*)^\top \left(\mathbb{E}[T_{x^*}(n)]x^*(x^*)^\top\right)^\dagger x^*$$
$$= \frac{\|x^*\|^2}{\mathbb{E}[T_{x^*}(n)]}.$$

Now multiply both sides of (27) by $\log^{1/2}(n)$ and take the limes superior as n tends to infinity. By assumption, $\limsup_{n\to\infty} \log(n) \|x - x^*\|_{\bar{G}_n^{-1}} \leq \Delta_x/\sqrt{2}$, hence the result is proved if $\limsup_{n\to\infty} \sqrt{\frac{\log(n)}{\mathbb{E}[T_{x^*}(n)]}} = 0$. This follows because π is consistent and the optimal action is assumed to be unique, hence it holds that $\lim_{n\to\infty} \mathbb{E}[T_{x^*}(n)]/n = 1$.

For the second part we start with

$$\frac{R_\theta^\pi(n)}{\log(n)} = \sum_{x \in \mathcal{A}^-} \frac{\mathbb{E}[T_x(n)]}{\log(n)} \Delta_x \,.$$

Then $\alpha_n(x) = \mathbb{E}[T_x(n)] / \log(n)$ is asymptotically feasible for n large. Indeed, $\bar{G}_n = \log(n) H(\alpha_n)$, hence $\bar{G}_n^{-1} = H^{-1}(\alpha_n)/\log(n)$ and so

$$\frac{\Delta_x^2}{2} \geq \limsup_{n\to\infty} \log(n) \|x\|_{\bar{G}_n^{-1}}^2 = \limsup_{n\to\infty} \|x\|_{H^{-1}(\alpha_n)}^2 \,.$$

Thus for any $\varepsilon > 0$ and n large enough, $\|x\|_{H^{-1}(\alpha_n)}^2 \leq \Delta_x^2/2 + \varepsilon$ and also

$$\frac{R_\theta^\pi(n)}{\log(n)} = \sum_{x \in \mathcal{A}^-} \frac{\mathbb{E}[T_x(n)]}{\log(n)} \Delta_x \geq c_\varepsilon(\mathcal{A}, \theta) \,,$$

where $c_\varepsilon(\mathcal{A}, \theta)$ is the solution to the optimisation problem (2) where $\Delta_x^2/2$ is replaced by $\Delta_x^2/2 + \varepsilon$. Hence, $\liminf_{n\to\infty} \frac{R_\theta^\pi(n)}{\log(n)} \geq c_\varepsilon(\mathcal{A}, \theta)$. Since $\varepsilon > 0$ was arbitrary and $\inf_{\varepsilon>0} c_\varepsilon(\mathcal{A}, \theta) = c(\mathcal{A}, \theta)$, we get the desired result. $\square$

C PROOF OF THE FIRST PART OF THEOREM 1

We start by proving that the expected Gram matrix is eventually non-singular. Let π be a consistent strategy and $\mathcal{A}$ and θ be the action-set and parameter for a linear bandit. Define $\mathcal{A}' = \{x : \mathbb{E}[\sum_{t=1}^n \mathbb{1}\{A_t = x\}] > 0\}$ to be the set of arms that are played at least once with non-zero probability. We proceed by contradiction. Suppose that $\bar{G}_n$ is singular for all n. Then there exists an $x \in \mathcal{A}$ such that $x \notin \text{span}\,\mathcal{A}'$. Decompose $x = y + z$ where $y \in \text{span}\,\mathcal{A}'$ and $z \in (\text{span}\,\mathcal{A}')^\perp$ is non-zero and in the orthogonal complement of the subspace spanned by $\mathcal{A}'$. Therefore $\langle x', z \rangle = 0$ for all $x' \in \mathcal{A}'$. Define an alternative bandit with the same action-set and parameter $\theta' = \theta + 2\Delta_{\max} \frac{z}{\|z\|^2}$. Then for any $x' \in \mathcal{A}'$, $\langle x, \theta' \rangle = \langle x, \theta \rangle + 2\Delta_{\max} \geq \langle x', \theta \rangle - \Delta_x + 2\Delta_{\max} \geq \langle x', \theta' \rangle + \Delta_{\max}$, hence in the bandit determined by θ' the suboptimality gaps of the actions in $\mathcal{A}'$ is at least $\Delta_{\max}$. Let $\mathbb{P}'$ be the bandit distribution over $\Omega_n = (\mathcal{A} \times \mathcal{R})^n$ defined by θ' and let $\mathbb{E}'$ be the corresponding expectation operator. Since the distribution of the noise of rewards is independent of the action chosen and $\langle x', \theta - \theta' \rangle = 0$ for all $x' \in \mathcal{A}'$, $\mathbb{P} = \mathbb{P}'$ (the

bandits determined by θ and θ' appear identical to the algorithm). In particular, $\mathbb{E}'[\sum_{t=1}^n \mathbb{1}\{A_t \notin \mathcal{A}'\}] = 0$, and yet by construction we have

$$R_{\theta'}^\pi(n) \geq \Delta_{\max} \mathbb{E}'\left[\sum_{t=1}^n \mathbb{1}\{A_t \in \mathcal{A}'\}\right] = n\Delta_{\max} \,.$$

Therefore the regret is linear for θ', which implies that π is not consistent. Therefore for sufficiently large n, $\bar{G}_n$ must be non-singular.

Now that we are allowed to invert $\bar{G}_n$ for sufficiently large n we show the second part that $\liminf_{n\to\infty} \lambda_{\min}(G_n)/\log(n) > 0$. Suppose on the contrary, that

$$\liminf_{n\to\infty} \frac{1}{\|\bar{G}_n^{-1}\| \log(n)} = \liminf_{n\to\infty} \frac{\lambda_{\min}(\bar{G}_n)}{\log(n)} = 0 \,.$$

Then because $\mathcal{A}$ spans $\mathbb{R}^d$ there exists an $x \in \mathcal{A}$ and an infinite set $S \subseteq \mathbb{N}$ such that

$$\lim_{n \in S} \log(n) \|x - x^*\|_{\bar{G}_n^{-1}}^2 = \infty$$

$$\lim_{n \in S} \frac{\|x - x^*\|_{\bar{G}_n^{-1}}^2}{\|\bar{G}_n^{-1}\|} > 0 \,.$$

Now one simply repeats the analysis used in the proof of the second part up to (13), but now the condition that $\|x - x^*\|_H > 0$ is satisfied by the second line above and a contradiction is reached.

D PROOF OF LEMMA 12

Let $B \subseteq \mathcal{A}$ be a barycentric spanner and let $S \in [0, \infty]^k$ be an alternative to T given by

$$S_x = \begin{cases} \infty, & \text{if } x = x^*; \\ \frac{2d^2 f_n}{\Delta_{\min}^2}, & \text{if } x \in B; \\ 0, & \text{otherwise} \,. \end{cases}$$

Then $\|x^*\|_{H_s^\dagger} = 0$ and for $x^* \neq y \in \mathcal{A}$ we have

$$\|y\|_{H_{S\dagger}}^2 \leq \left(\sum_{x \in B} \|x\|_{H_{S\dagger}}\right)^2$$

$$\leq \left(\frac{\Delta_{\min}}{\sqrt{2f_n}}\right)^2 \leq \frac{\Delta_y^2}{2f_n} \,.$$

Therefore

$$\sum_{x:\Delta_x>0} T_x \leq \frac{1}{\Delta_{\min}} \sum_{x:\Delta_x>0} T_x \Delta_x$$

$$\leq \sum_{x:\Delta_x>0} S_x \Delta_x \leq \frac{2d^3 \Delta_{\max} f_n}{\Delta_{\min}^3} \,. \qquad \square$$

E   PROOF OF LEMMA 13

The proof of Lemma 13 requires one more technical result.

Lemma 17. *Let $\varepsilon > 0$ and recall the definition of $T_n(\hat{\Delta})$ given in Definition 9. For $m \in \mathbb{N}$ define*

$$S_{n,m}(\hat{\Delta}) = \min\left\{ m f_n, T_n(\hat{\Delta}) \right\} .$$

Then there exists an m such that for all $n \in \mathbb{N}$ and $\hat{\Delta} \in [0, \infty)^k$ and $x \in \mathcal{A}$

$$\|x\|^2_{H^{-1}_{S_{n,m}(\hat{\Delta})}} \le \max\left\{ \frac{\varepsilon^2}{f_n}, \frac{\hat{\Delta}_x^2}{f_n} \right\} .$$

Proof of Lemma 13. Assume that F'_n does not hold. We consider three cases.

Case 1. $\hat{\Delta}_{x^*} > 0$.

Case 2. $\hat{\Delta}_{x^*} = 0$ and $\hat{\Delta}_{\min} > \Delta_{\min}/4$.

Case 3. $\hat{\Delta}_{x^*} = 0$ and $\hat{\Delta}_{\min} \le \Delta_{\min}/4$.

The idea is to show that in each case the regret is at most logarithmic, with a leading constant that depends on θ and $\mathcal{A}$, but not on the observed samples. Treating each case separately.

Case 1   Recall that $\hat{\Delta} \in \mathbb{R}^k$ (indexed by the actions) is the empirical estimate of the sub-optimality gaps after the warm-up phase. Let x be the sub-optimal arm for which $\hat{\Delta}_x = 0$. By the definition of the optimisation problem this arm will be played in every while loop. Let t be the first round when for all x it holds that

$$\|x\|^2_{G_t^{-1}} \le \max\left\{ \frac{\hat{\Delta}_x^2}{f_n}, \frac{\Delta_{\min}^2}{16 f_n} \right\} .$$

By Lemma 17 there exists a constant m_1 depending only on $\mathcal{A}$ and θ such that

$$t \le m_1 f_n .$$

By the assumption that F'_n does not hold (and its definition) we have

$$
\begin{aligned}
\hat{\mu}_{x^*}(t) &\ge \mu_{x^*} - \max\left\{ \hat{\Delta}_{x^*}, \Delta_{\min}/4 \right\} \\
&\ge \mu_x + \Delta_x - \hat{\Delta}_{x^*} - \frac{\Delta_{\min}}{4} \\
&\ge \hat{\mu}_x(t) + \Delta_x - \frac{\Delta_{\min}}{2} - \hat{\Delta}_{x^*} \\
&\ge \hat{\mu}_x(t) + \frac{\Delta_{\min}}{2} + \hat{\mu}_{x^*}(t_0) - \hat{\mu}_x(t_0) ,
\end{aligned}
$$

where $t_0 = d\lceil \log^{1/2}(n) \rceil$ is the round at the end of the warm-up phase. Therefore if n is sufficiently large that $\Delta_{\min}/2 \ge 4\varepsilon_n$, then

$$\hat{\mu}_{x^*}(t) - \hat{\mu}_{x^*}(t_0) + \hat{\mu}_x(t_0) - \hat{\mu}_x(t) \ge \frac{\Delta_{\min}}{2} \ge 4\varepsilon_n ,$$

which by the fact that $\max\{a, b\} \ge (a + b)/2$ for all $a, b \in \mathbb{R}$ implies that the success phase of the algorithm ends. Therefore if n is sufficiently large, then in case 1 the regret in the success phase is at most

$$\sum_{t \in T_{\text{succ.}}} \Delta_{A_t} \le \Delta_{\max} m_1 f_n . \tag{28}$$

Case 2   Recall that $\hat{T}$ is the strategy used in the success phase based on samples collected in the warm-up phase. Since $\hat{\Delta}_{x^*} = 0$ and $\hat{\Delta}_{\min} \ge \Delta_{\min}/4$, by Lemma 12 it holds that

$$\sum_{x \ne x^*} \hat{T}_x \le \frac{2 \cdot 4^3 d^3 f_n \Delta_{\max}}{\Delta_{\min}^3} .$$

And again we have that for sufficiently large n that the regret in the success phase is at most

$$\sum_{t \in T_{\text{succ.}}} \Delta_{A_t} \le \frac{2 \cdot 4^3 d^3 f_n \Delta_{\max}^2}{\Delta_{\min}^3} . \tag{29}$$

Case 3   For the final case we assume that $\hat{\Delta}_{x^*} = 0$ and there exists an x for which $\hat{\Delta}_x \le \Delta_{\min}/4$. Let t be the first time-step when for all $x \in \mathcal{A}$ it holds that

$$\|x\|^2_{G_t^{-1}} \le \max\left\{ \frac{\Delta_{\min}^2}{64 f_n}, \frac{\hat{\Delta}_x^2}{f_n} \right\} .$$

Then by Lemma 17 there exists a constant m_2 that is independent of $\hat{\Delta}$ and n such that $t \le m_2 f_n$. Then since F'_n does not hold we have

$$
\begin{aligned}
\hat{\mu}_{x^*}(t) &- \hat{\mu}_{x^*}(t_0) + \hat{\mu}_x(t_0) - \hat{\mu}_x(t) \\
&\ge \hat{\mu}_{x^*}(t) - \hat{\mu}_x(t) - \hat{\Delta}_x \ge \frac{\Delta_{\min}}{4} - \hat{\Delta}_x \\
&\ge \frac{\Delta_{\min}}{2} \ge 2\varepsilon_n .
\end{aligned}
$$

Therefore provided that n is sufficiently large, the success phase ends and by the same reasoning as in Case 1 the regret in the success phase is bounded by

$$\sum_{t \in T_{\text{succ.}}} \Delta_{A_t} \le \Delta_{\max} m_2 f_n . \tag{30}$$

The proof of the lemma is completed by combining (28), (29) and (30), which imply the existence of a constant m_3 that is independent of n and $\hat{\Delta}$ such that

$$\mathbb{1}\{\text{not } F'_n\} \sum_{t \in T_{\text{succ.}}} \Delta_{A_t} \le m_3 f_n .$$

Therefore by (14) and the definition of $f_n \sim 2\log(n)$ we have

$$\limsup_{n\to\infty} \frac{\mathbb{E}\left[\mathbb{1}\{F_n \text{ and not } F'_n\} \sum_{t\in T_{\text{succ.}}} \Delta_{A_t}\right]}{\log(n)}$$
$$\leq \limsup_{n\to\infty} \frac{\mathbb{E}[\mathbb{1}\{F_n\} m_3 f_n]}{\log(n)}$$
$$= \limsup_{n\to\infty} \frac{\mathbb{P}(F_n) m_3 f_n}{\log(n)}$$
$$\leq \limsup_{n\to\infty} \frac{m_3 f_n}{\log^2(n)}$$
$$= 0.$$ $\qquad\square$

F   SOLVING THE OPTIMISATION PROBLEM IN EXAMPLE 4

First note $\Delta_{x_1} = 0$ so that $\alpha(x_1)$ should be arbitrarily large in (2). Then

$$H(\alpha) = \begin{bmatrix} \alpha(x_1) + \alpha(x_3)(1-\varepsilon)^2 & \alpha(x_3)\varepsilon(1-\varepsilon)\alpha \\ \alpha(x_3)\varepsilon(1-\varepsilon)\alpha & \alpha(x_3)\varepsilon^2\alpha^2 + \alpha(x_2) \end{bmatrix}$$

Then

$$\lim_{\alpha(x_1)\to\infty} H(\alpha)^{-1} = \begin{bmatrix} 0 & 0 \\ 0 & \frac{1}{\alpha(x_3)\varepsilon^2\alpha^2+\alpha(x_2)} \end{bmatrix}$$

We need to choose $\alpha(x_2) \geq 0$ and $\alpha(x_3) \geq 0$ so that

$$\frac{1}{\alpha(x_3)\varepsilon^2\alpha^2 + \alpha(x_2)} = \|x_2\|^2_{H(\alpha)^{-1}} \leq \frac{1}{2}$$
$$\frac{\alpha^2\varepsilon^2}{\alpha(x_3)\varepsilon^2\alpha^2 + \alpha(x_2)} = \|x_3\|^2_{H(\alpha)^{-1}} \leq \frac{\varepsilon^2}{2},$$

which for $\alpha \geq 1$ is equivalent to

$$\alpha(x_3)\varepsilon^2\alpha^2 + \alpha(x_2) \geq 2\alpha^2.$$

The objective we should minimise becomes

$$\alpha(x_2) + \varepsilon\alpha(x_3).$$

So we see that $\alpha(x_2) = 2\alpha^2$ and $\alpha(x_3) = 0$ is optimal as long as $2/\varepsilon > 2\alpha^2$. Hence, $c(\mathcal{A}, \theta) = 2\alpha^2$.

G   CONVEXITY OF OPTIMISATION PROBLEM IN DEFINITION 9

The objective is linear, which means we only have to check the feasible set is convex. Let $G, H \in \mathbb{R}^{d\times d}$ be invertible and positive definite. Therefore via a simultaneous diagonalisation there exists a $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_d)$ and invertible symmetric $Q \in \mathbb{R}^{d\times d}$ such that

$$Q^\top Q = G \quad \text{and} \quad H = Q^\top \Lambda Q.$$

Then for any $p \in [0, 1]$ and $y = Q^{-1}x$ we have

$$\|x\|^2_{(pG+(1-p)H)^{-1}} = x^\top (pG + (1-p)H)^{-1} x$$
$$= y^\top (pI + (1-p)\Lambda)^{-1} y = \sum_{i=1}^d \frac{y_i^2}{p + (1-p)\lambda_i}$$
$$\leq \sum_{i=1}^d \left(py_i^2 + \frac{(1-p)y_i^2}{\lambda_i}\right) = p\|y\|^2 + (1-p)\|y\|^2_{\Lambda^{-1}}$$
$$= p\|x\|^2_{G^{-1}} + (1-p)\|x\|^2_{H^{-1}}.$$

The result follows easily from the definition of the constraints, though care is needed to deal with the limit definition of H_T^{-1} for T with $\|T\|_\infty = \infty$.