25 Asymptotic Lower Bounds for Stochastic Linear Bandits

The lower bounds in the previous chapter were derived by analyzing the worst case for specific action sets and/or constraints on the unknown parameter. In this chapter we focus on the asymptotics of the problem and aim to understand the influence of the action set on the regret. We assume that $\mathcal{A} \subset \mathbb{R}^d$ is finite with $|\mathcal{A}| = K$ and that the reward is $X_t = \langle A_t, \theta \rangle + \eta_t$ where $\theta \in \mathbb{R}^d$ and η_t is a sequence of independent standard Gaussian noise. Of course the regret of a policy in this setting is

$$R_n(\mathcal{A}, \theta) = \mathbb{E}_\theta \left[\sum_{t=1}^n \Delta_{A_t} \right], \qquad \Delta_a = \max_{a' \in \mathcal{A}} \langle a' - a, \theta \rangle ,$$

where the dependence on the policy is omitted for readability and $\mathbb{E}_\theta[\cdot]$ is the expectation with respect to the measure on outcomes induced by the interaction of the policy and the linear bandit determined by θ. Like the asymptotic lower bounds in the classical finite-armed case (Chapter 16), the results of this chapter are proven only for consistent policies. Recall that a policy is consistent in some class of bandits $\mathcal{E}$ if the regret is subpolynomial for any bandit in that class. Here this means that

$$R_n(\mathcal{A}, \theta) = o(n^p) \qquad \text{for all } p > 0 \text{ and } \theta \in \mathbb{R}^d . \tag{25.1}$$

The main objective of the chapter is to prove the following theorem on the behaviour of any consistent policy and discuss the implications.

THEOREM 25.1 *Assume that $\mathcal{A} \subset \mathbb{R}^d$ is finite and spans $\mathbb{R}^d$ and suppose a policy is consistent (satisfies Eq. 25.1). Let $\theta \in \mathbb{R}^d$ be any parameter such that there is a unique optimal action and let $\bar{G}_n = \mathbb{E}_\theta \left[\sum_{t=1}^n A_t A_t^\top \right]$ be the expected Gram matrix . Then $\liminf_{n\to\infty} \lambda_{\min}(\bar{G}_n)/\log(n) > 0$. Furthermore, for any $a \in \mathcal{A}$ it holds that:*

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

The reader should recognize $\|a\|^2_{\bar{G}_n^{-1}}$ as the key term in the width of the confidence interval for the least squares estimator (Chapter 20). This is quite intuitive. The theorem is saying that any consistent algorithm must prove statistically that all suboptimal arms are indeed suboptimal by making the size of the confidence interval smaller than the suboptimality gap. Before the

proof of this result we give a corollary that characterizes the asymptotic regret that must be endured by any consistent policy.

COROLLARY 25.1 *Let $\mathcal{A} \subset \mathbb{R}^d$ be a finite set that spans $\mathbb{R}^d$ and $\theta \in \mathbb{R}^d$ be such that there is a unique optimal action. Then for any consistent policy*

$$\liminf_{n\to\infty} \frac{R_n(\mathcal{A},\theta)}{\log(n)} \geq c(\mathcal{A},\theta)\,,$$

where $c(\mathcal{A},\theta)$ is defined as

$$c(\mathcal{A},\theta) = \inf_{\alpha\in[0,\infty)^{\mathcal{A}}} \sum_{a\in\mathcal{A}} \alpha(a)\Delta_a$$
$$\text{subject to } \|a\|^2_{H_\alpha^{-1}} \leq \frac{\Delta_a^2}{2} \text{ for all } a \in \mathcal{A} \text{ with } \Delta_a > 0\,,$$

where $H = \sum_{a\in\mathcal{A}} \alpha(a) a a^\top$.

The lower bound is complemented by a matching upper bound that we will not prove.

THEOREM 25.2 *Let $\mathcal{A} \subset \mathbb{R}^d$ be a finite set that spans $\mathbb{R}^d$. Then there exists a policy such that*

$$\limsup_{n\to\infty} \frac{R_n(\mathcal{A},\theta)}{\log(n)} \leq c(\mathcal{A},\theta)\,,$$

where $\mathcal{A}$ is defined as in Corollary 25.1.

Proof of Theorem 25.1 The proof of the first part is simply omitted (see the reference below for details). It follows along similar lines to what follows, essentially that if G_n is not sufficiently large in every direction, then some alternative parameter is not sufficiently identifiable. Let $a^* = \operatorname{argmax}_{a\in\mathcal{A}} \langle a, \theta\rangle$ be the optimal action, which we assumed to be unique. Let $\theta' \in \mathbb{R}^d$ be an alternative parameter to be chosen subsequently and let $\mathbb{P}$ and $\mathbb{P}'$ be the measures on the sequence of outcomes $A_1, Y_1, \ldots, A_n, Y_n$ induced by the interaction between the policy and the bandit determined by θ and θ' respectively. Let $\mathbb{E}[\cdot]$ and $\mathbb{E}'[\cdot]$ be the expectation operators of $\mathbb{P}$ and $\mathbb{P}'$ respectively. By Theorem 14.2 and Lemma 15.1 for any event E we have

$$\begin{aligned}\mathbb{P}(E) + \mathbb{P}'(E^c) &\geq \frac{1}{2}\exp\left(-\mathrm{D}(\mathbb{P},\mathbb{P}')\right) \\ &= \frac{1}{2}\exp\left(-\frac{1}{2}\mathbb{E}\left[\sum_{t=1}^n \langle A_t, \theta - \theta'\rangle^2\right]\right) = \frac{1}{2}\exp\left(-\frac{1}{2}\|\theta-\theta'\|^2_{\bar{G}_n}\right)\,.\end{aligned} \tag{25.2}$$

A simple re-arrangement shows that

$$\frac{1}{2}\|\theta-\theta'\|^2_{\bar{G}_n} \geq \log\left(\frac{1}{2\mathbb{P}(E) + 2\mathbb{P}'(E^c)}\right)\,.$$

Now we follow the usual plan of choosing θ' to be close to θ, but so that the

optimal action in the bandit determined by θ' is not a^*. Let $\Delta_{\min} = \min\{\Delta_a : a \in \mathcal{A}, \Delta_a > 0\}$ and $\varepsilon \in (0, \Delta_{\min})$ and H be a positive definite matrix to be chosen later such that $\|a - a^*\|_H^2 > 0$. Then define

$$\theta' = \theta + \frac{\Delta_a + \varepsilon}{\|a - a^*\|_H^2} H(a - a^*) \,,$$

which is chosen so that

$$\langle a - a^*, \theta' \rangle = \langle a - a^*, \theta \rangle + \Delta_a + \varepsilon = \varepsilon \,.$$

This means that a^* is ε-suboptimal action for bandit θ'. We abbreviate $R_n = R_n(\mathcal{A}, \theta)$ and $R'_n = R_n(\mathcal{A}, \theta')$. Then

$$R_n = \mathbb{E}\left[\sum_{a \in \mathcal{A}} T_a(n) \Delta_a\right] \geq \frac{n \Delta_{\min}}{2} \mathbb{P}\left(T_{a^*}(n) < n/2\right) \geq \frac{n\varepsilon}{2} \mathbb{P}\left(T_{a^*}(n) < n/2\right) \,,$$

where $T_a(n) = \sum_{t=1}^n \mathbb{I}\{A_t = a\}$. Similarly, a^* is at least ε-suboptimal in bandit θ' so that

$$R'_n \geq \frac{n\varepsilon}{2} \mathbb{P}'\left(T_{a^*}(n) \geq n/2\right) \,.$$

Therefore

$$\mathbb{P}\left(T_{a^*}(n) < n/2\right) + \mathbb{P}'\left(T_{a^*}(n) \geq n/2\right) \leq \frac{2}{n\varepsilon}\left(R_n + R'_n\right) \,. \tag{25.3}$$

Note that this holds for practically any choice of H as long as $\|a - a^*\|_H > 0$. The logical next step is to select H (which determines θ') to make (25.2) as large as possible. The main difficulty is that this depends on n, so instead we aim to choose an H so the quantity is large enough infinitely often. We starting by just re-arranging things:

$$\frac{1}{2}\|\theta - \theta'\|_{\bar{G}_n}^2 = \frac{(\Delta_a + \varepsilon)^2}{2} \cdot \frac{\|a - a^*\|_{H\bar{G}_n H}^2}{\|a - a^*\|_H^4} = \frac{(\Delta_a + \varepsilon)^2}{2\|a - a^*\|_{\bar{G}_n^{-1}}^2} \rho_n(H) \,,$$

where we introduced

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

Therefore by choosing E to be the event that $T_{a^*}(n) < n/2$ and using (25.3) and (25.2) we have

$$\frac{(\Delta_a + \varepsilon)^2}{2\|a - a^*\|_{\bar{G}_n^{-1}}^2} \rho_n(H) \geq \log\left(\frac{n\varepsilon}{4R_n + 4R'_n}\right) \,,$$

which after re-arrangement leads to

$$\frac{(\Delta_a + \varepsilon)^2}{2\log(n)\|a - a^*\|_{\bar{G}_n^{-1}}^2} \rho_n(H) \geq 1 - \frac{\log((4R_n + 4R'_n)/\varepsilon)}{\log(n)} \,.$$

The definition of consistency means that R_n and R'_n are both sub-polynomial,

which implies that the second term in the previous expression tends to zero for large n and so by sending ε to zero we see that

$$\liminf_{n\to\infty} \frac{\rho_n(H)}{\log(n)\|a-a^*\|^2_{\bar{G}_n^{-1}}} \geq \frac{2}{\Delta_a^2}\,. \tag{25.4}$$

We complete the result using proof by contradiction. Suppose that

$$\limsup_{n\to\infty} \log(n)\|a-a^*\|^2_{\bar{G}_n^{-1}} > \frac{\Delta_a^2}{2}\,. \tag{25.5}$$

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

$$\log(n)\|a-a^*\|^2_{\bar{G}_n^{-1}} \geq \frac{(\Delta_a+\varepsilon)^2}{2} \quad \text{for all } n \in S\,.$$

Therefore by (25.4), $\liminf_{n\in S} \rho_n(H) > 1$. We now choose H to be a cluster point of the sequence $(\bar{G}_n^{-1}/\|\bar{G}_n^{-1}\|)_{n\in S}$ where $\|\bar{G}_n^{-1}\|$ is the spectral norm of the matrix $\bar{G}_n^{-1}$. Such a point must exist, since matrices in this sequence have unit spectral norm by definition, and the set of matrices with bounded spectral norm is compact. We let $S' \subseteq S$ be a subset so that $\bar{G}_n^{-1}/\|\bar{G}_n^{-1}\|$ converges to H on $n \in S'$. We now check that $\|a-a^*\|_H > 0$.

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

where the last inequality follows from the assumption in (25.5) and the first part of the theorem. Therefore

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

which is a contradiction, and so we conclude that (25.5) does not hold and so

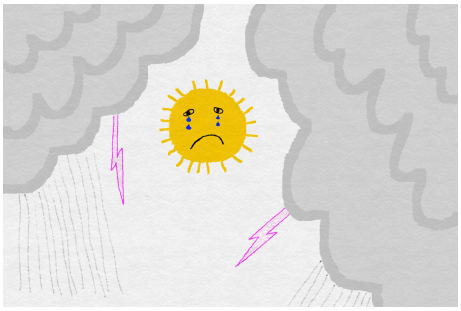
$$\limsup_{n\to\infty} \log(n)\|a-a^*\|^2_{\bar{G}_n^{-1}} \leq \frac{\Delta_a^2}{2}\,. \qquad \square$$

We leave the proof of the corollary as an exercise for the reader. Essentially though, any consistent algorithm must choose its actions so that in expectation

$$\|a-a^*\|^2_{\bar{G}_n^{-1}} \leq (1+o(1))\frac{\Delta_a^2}{2\log(n)}\,.$$

Now since a^* will be chosen linearly often it is easily shown for suboptimal a that $\lim_{n\to\infty} \|a-a^*\|_{\bar{G}_n^{-1}}/\|a\|_{\bar{G}_n^{-1}} \to 1$. This leads to the required constraint on the actions of the algorithm, and the optimization problem in the corollary is derived by minimizing the regret subject to this constraint.

25.1 Clouds looming for optimism



The theorem and its corollary have disturbing implications for policies based on the principle of optimism in the face of uncertainty, which is that they can never be asymptotically optimal. The reason is that these policies do not choose actions for which they have collected enough statistics to prove they are suboptimal, but in the linear setting it can still be worthwhile playing these actions in case they are very informative about other actions for which the statistics are not yet so clear. As we shall see, a problematic example appears in the simplest case where there is information sharing between the arms. Namely, when the dimension is $d = 2$ and there are $K = 3$ arms.

Let $\mathcal{A} = \{a_1, a_2, a_3\}$ where $a_1 = e_1$ and $a_2 = e_2$ and $a_3 = (1 - \varepsilon, \gamma\varepsilon)$ where $\gamma \geq 1$ and $\varepsilon > 0$ is small. Let $\theta = (1, 0)$ so that the optimal action is $a^* = a_1$ and $\Delta_{a_2} = 1$ and $\Delta_{a_3} = \varepsilon$. Clearly if ε is very small, then a_1 and a_3 point in nearly the same direction and so choosing only these arms does not provide sufficient information to quickly learn which of a_1 or a_3 is optimal. On the other hand, a_2 and $a_1 - a_3$ point in very different directions and so choosing a_2 allows a learning agent to quickly identify that a_1 is in fact optimal. We now show how the theorem and corollary demonstrate this. First we calculate what is the optimal solution to the optimization problem in Corollary 25.1. Recall we are trying to minimize

$$\sum_{a \in \mathcal{A}} \alpha(a)\Delta_a \qquad \text{subject to } \|a\|^2_{H(\alpha)^{-1}} \leq \frac{\Delta_a^2}{2} \text{for all } a \in \mathcal{A}\,,$$

where $H = \sum_{a \in \mathcal{A}} \alpha(a) a a^\top$. Clearly we should choose $\alpha(a_1)$ arbitrarily large, then a computation shows that

$$\lim_{\alpha(a_1) \to \infty} H(\alpha)^{-1} = \begin{bmatrix} 0 & 0 \\ 0 & \frac{1}{\alpha(a_3)\varepsilon^2\gamma^2 + \alpha(a_2)} \end{bmatrix}.$$

The constraints mean that

$$\frac{1}{\alpha(a_3)\varepsilon^2\gamma^2 + \alpha(a_2)} = \lim_{\alpha(a_1) \to \infty} \|a_2\|^2_{H(\alpha)^{-1}} \leq \frac{1}{2}$$
$$\frac{\gamma^2\varepsilon^2}{\alpha(a_3)\varepsilon^2\gamma^2 + \alpha(a_2)} = \lim_{\alpha(a_1) \to \infty} \|a_3\|^2_{H(\alpha)^{-1}} \leq \frac{\varepsilon^2}{2}\,.$$

Provided that $\gamma \geq 1$ this reduces simply to the constraint that

$$\alpha(a_3)\varepsilon^2 + \alpha(a_2) \geq 2\gamma^2\,.$$

Since we are minimizing $\alpha(a_2) + \varepsilon\alpha(a_3)$ we can easily see that $\alpha(a_2) = 2\gamma^2$ and $\alpha(a_3) = 0$ provided that $2\gamma^2 \leq 2/\varepsilon$. Therefore if ε is chosen sufficiently small relative to γ, then the optimal rate of the regret is $c(\mathcal{A}, \theta) = 2\gamma^2$ and so by

Theorem 25.2 there exists a policy such that

$$\limsup_{n\to\infty} \frac{R_n(\mathcal{A},\theta)}{\log(n)} = 2\gamma^2 \,.$$

Now we argue that for γ sufficiently large and ε arbitrarily small that the regret for any consistent optimistic algorithm is at least

$$\limsup_{n\to\infty} \frac{R_n(\mathcal{A},\theta)}{\log(n)} = \Omega(1/\varepsilon) \,,$$

which can be arbitrarily worse than the optimal rate! So why is this so? Recall that optimistic algorithms choose

$$A_t = \operatorname{argmax}_{a\in\mathcal{A}} \max_{\tilde\theta\in\mathcal{C}_t} \left\langle a, \tilde\theta \right\rangle \,,$$

where $\mathcal{C}_t \subset \mathbb{R}^d$ is a confidence set that we assume contains the true θ with high probability. So far this does not greatly restrict the class of algorithms that we might call optimistic. We now assume that there exists a constant $c > 0$ such that

$$\mathcal{C}_t \subseteq \left\{ \tilde\theta : \|\hat\theta_t - \tilde\theta\|_{G_t} \le c\sqrt{\log(n)} \right\} \,.$$

So now we ask how often can we expect the optimistic algorithm to choose action $a_2 = e_2$ in the example described above? Since we have assumed $\theta \in \mathcal{C}_t$ with high probability we have that

$$\max_{\tilde\theta\in\mathcal{C}_t} \langle a_1, \tilde\theta\rangle \ge 1 \,.$$

On the other hand, if $T_{a_2}(t-1) > 4c^2\log(n)$, then

$$\max_{\tilde\theta\in\mathcal{C}_t} \langle a_2, \tilde\theta\rangle = \max_{\tilde\theta\in\mathcal{C}_t} \langle a_2, \tilde\theta - \theta\rangle \le 2c\sqrt{\|a_2\|_{G_t^{-1}} \log(n)} \le 2c\sqrt{\frac{\log(n)}{T_{a_2}(t-1)}} < 1 \,,$$

which means that a_2 will not be chosen more than $1 + 4c^2\log(n)$ times. So if $\gamma = \Omega(c^2)$, then the optimistic algorithm will not choose a_2 sufficiently often and a simple computation shows it must choose a_3 at least $\Omega(\log(n)/\varepsilon^2)$ times and suffers regret of $\Omega(\log(n)/\varepsilon)$. The key take-away from this is that optimistic algorithms do not choose actions that are statistically suboptimal, but for linear bandits it can be optimal to choose these actions more often to gain information about *other actions*.

Needless to say this conclusion should generalize to various other structured bandits problems. In other words, when choosing an action makes you learn about other actions, optimism, while often providing some basic guarantees, may be missing out in better exploiting the structure of the problem.

25.2 Notes

1 The algorithm that realizes Theorem 25.2 is a complicated three-phase affair that we cannot recommend in practice. A practical asymptotically optimal algorithm for linear bandits is a fascinating open problem.
2 In Chapter 35 we will introduce the randomized Bayesian algorithm called Thompson sampling algorithm for finite-armed and linear bandits. While Thompson sampling comes with several benefits over UCB, it does not overcome the issues described here.
3 The main difficulty in designing asymptotically optimal algorithms is how to balance the tradeoff between information and regret. One algorithm that tries to this in an explicit way is "Information-Directed Sampling" by Russo and Van Roy [2014a], which we also discuss in Chapter 35. It is not known if the algorithm proposed there is optimal when adapted to linear bandits.

25.3 Bibliographic remarks

The theorems of this chapter are by the authors: Lattimore and Szepesvári [2017]. The example in Section 25.1 first appeared in a paper by Soare et al. [2014], which deals with the problem of best arm identification for linear bandits (for an introduction to best arm identification see Chapter 33).

25.4 Exercises

25.1 Prove Corollary 25.1.

25.2 Prove the first part of Theorem 25.1.

25.3 Give an example of an action set $\mathcal{A} \subset \mathbb{R}^d$ and $\theta \in \mathbb{R}^d$ and vector $a \in \mathbb{R}^d$ where the asymptotic regret for the same θ and action-set $\mathcal{A} \cup \{a\}$ is:

(a) Makes the asymptotic regret larger.
(b) Makes the asymptotic regret smaller.