29 The Relation Between Adversarial and Stochastic Linear Bandits

As we have seen in the preceding chapters, adversarial and stochastic linear bandits do share certain similarities. For example, the least squares estimator plays a fundamental role in both, as does the machinery from experimental design for optimizing the exploration distribution. There are also surprising differences, however. Theorem 24.2 shows that the regret for stochastic linear bandits on a ball is lower bounded by $\Omega(d\sqrt{n})$, while for the adversarial bandits the upper bound is $O(\sqrt{dn})$ as shown in Theorem 28.4. As we will keep referring to these results, we added Table 29.1 to summarize the situation. We hope the reader is at least mildly surprised that the regret upper for the adversarial environment is of lower order than the regret lower bound for the stochastic environment. After all, was not the purpose of working with adversarial environments to enlarge the scope of algorithms beyond stochastic environments? The purpose of this chapter is to explain why our intuition fails us in this case.

To make the notation consistent we present the stochastic and adversarial linear bandit frameworks again, but this time using losses for both. Let $\mathcal{A} \subset \mathbb{R}^d$ be the action set. In each round the learner should choose $A_t \in \mathcal{A}$ and receives the loss Y_t, where

$$Y_t = \langle A_t, \theta \rangle + \eta_t \,, \qquad \text{(Stochastic setting)} \tag{29.1}$$

$$Y_t = \langle A_t, \theta_t \rangle \,, \qquad \text{(Adversarial setting)} \tag{29.2}$$

where $(\eta_t)_t$ is a sequence of independent and identically distributed 1-subgaussian random variables and $(\theta_t)_t$ is a sequence of loss vectors chosen by the adversary. As noted earlier, the assumptions on the noise can be relaxed significantly. For example, if $\mathcal{F}_t = \sigma(A_1, Y_1, \ldots, A_t, Y_t, A_{t+1})$, then the results of the previous chapters hold as soon as $\eta_t | \mathcal{F}_{t-1} \sim \text{subG}(1)$. The expected regret for the two

	Stochastic environment	Adversarial environment
Regret	$\Omega(d\sqrt{n})$	$O(\sqrt{dn})$

Table 29.1 The behavior of regret as a function of dimension d and number of rounds n for linear bandits when the action set $\mathcal{A} = B_2^d$ is the d-dimension Euclidean ball.

cases are defined as follows:

$$R_n = \sum_{t=1}^{n} \mathbb{E}\left[\langle A_t, \theta\rangle\right] - n \inf_{a\in\mathcal{A}} \langle a, \theta\rangle\,, \qquad \text{(Stochastic setting)}$$

$$R_n = \sum_{t=1}^{n} \mathbb{E}\left[\langle A_t, \theta_t\rangle\right] - n \inf_{a\in\mathcal{A}} \langle a, \bar{\theta}_n\rangle\,. \qquad \text{(Adversarial setting)}$$

In the last display, $\bar{\theta}_n = \frac{1}{n}\sum_{t=1}^{n} \theta_t$ is the average of the loss vectors chosen by the adversary. We chose to write the adversarial form with the help of the average loss vector to emphasize the similarity between the two settings.

29.1 Reducing stochastic linear bandits to adversarial linear bandits

To formalize the intuition that adversarial environments are harder than stochastic environments one may try to find a **reduction** where learning in the stochastic environments is reduced to learning in the adversarial algorithms. Here, reducing problem E ('**e**asy') to problem H ('**h**ard') just means that we can use algorithms designed for problem H to solve instances of problem E. In order to do this we need to transform instances of problem E into instances of problem H and translate back the actions of algorithms to actions for problem E. To get a regret bound for problem E from regret bound for problem H, one needs to ensure that the losses translate properly between the problem classes.

Of course, based on our previous discussion we know that if there is a reduction from stochastic linear bandits to adversarial linear bandits then somehow the adversarial problem must change so that no contradiction is created in the curious case of the unit ball. To be able to use an adversarial algorithm in the stochastic environment, we need to specify a sequence $(\theta_t)_t$ so that the adversarial feedback matches the stochastic one. Comparing Eq. (29.1) and Eq. (29.2), we can see that the crux of the problem is incorporating the noise η_t into θ_t while satisfying the other requirements. One simple way of doing this is by introducing an extra dimension for the adversarial problem.

In particular, suppose that the stochastic problem is d-dimensional so that $\mathcal{A} \subset \mathbb{R}^d$. For the sake of simplicity, assume furthermore that the noise and parameter vector satisfy $|\langle \mathcal{A}, \theta\rangle + \eta_t| \le 1$ almost surely and that $a_* = \operatorname{argmin}_{a\in\mathcal{A}} \langle a, \theta\rangle$ exists. Then define $\mathcal{A}_{\text{aug}} = \{(a, 1) : a \in \mathcal{A}\} \subset \mathbb{R}^{d+1}$ and let the adversary choose $\theta_t = (\theta, \eta_t) \in \mathbb{R}^{d+1}$. The reduction is now straightforward:

1 Initialize adversarial bandit policy with action set $\mathcal{A}_{\text{aug}}$.

2 Collect action $A'_t = (A_t, 1)$ from the policy.

3 Play A_t and observe loss Y_t.

4 Feed Y_t to the adversarial bandit policy and repeat from step 2.

Suppose the adversarial policy guarantees a bound B_n on the expected regret:

$$R'_n = \mathbb{E}\left[\sum_{t=1}^{n}\langle A'_t, \theta_t\rangle - \inf_{a'\in\mathcal{A}_{\rm aug}}\sum_{t=1}^{n}\langle a', \theta_t\rangle\right] \le B_n\,.$$

Let $a'_* = (a_*, 1)$. Note that for any $a' = (a, 1) \in \mathcal{A}_{\rm aug}$, $\langle A_t, \theta\rangle - \langle a, \theta\rangle = \langle A'_t, \theta_t\rangle - \langle a', \theta_t\rangle$ and thus adversarial regret, and eventually B_n, will upper bound the stochastic regret:

$$\mathbb{E}\left[\sum_{t=1}^{n}\langle A_t, \theta\rangle] - n\langle a_*, \theta\rangle\right] = \mathbb{E}\left[\sum_{t=1}^{n}\langle A'_t, \theta_t\rangle - n\langle a'_*, \bar\theta_n\rangle\right] \le R'_n \le B_n\,.$$

Therefore the expected regret in the stochastic bandit is also at most B_n. We have to emphasize that this reduction changes the geometry of the decision sets for both the learner and the adversary. For example, if $\mathcal{A} = B_2^d$ is the unit ball, then neither $\mathcal{A}_{\rm aug}$ nor its polar $\mathcal{A}^\circ_{\rm aug}$ are unit balls. It does not seem like this should make much difference, but at least in the case of the ball, from our $\Omega(d\sqrt{n})$ lower bound on the regret for the stochastic case, we see that the changed geometry must make the adversary more powerful. This reinforces the importance of the geometry of the action set, which we have already seen in the previous chapter.

While the reduction shows one way to use adversarial algorithms in stochastic environments, the story seems to be unfinished. When facing a linear bandit problem with some action set $\mathcal{A}$, the user is forced to make a choice of whether to believe the environment to be stochastic. Strangely enough, if the environment is believed to be stochastic, the recommendation seems to be to run one's favorite adversarial linear bandit algorithm on the *augmented* action set. What if the environment may or may not be stochastic? One can still try to run the adversarial linear bandit algorithm with no changes. At present we cannot guarantee that this lead to a small regret. In fact, the regret may get larger than what it needs to be. For example, if the mirror descent algorithm of the last chapter is run on a stochastic environment with the learning rate of the algorithm set as recommended in Theorem 28.4, the regret upper bound increases to $\tilde O(d^2\sqrt{n})$. We believe that this increase of the regret is real. By tuning the learning rate, the regret can be brought back to $\tilde O(d\sqrt{n})$.

We see a case here when the cost of using an algorithm prepared to deal with a larger class of environments pays a nontrivial cost for its increased robustness. At least as far as minimax regret is concerned, this was not the case for finite-armed bandits.

The real reason for all these discrepancies is that that the adversarial linear bandit model is better viewed as relaxation of another class of stochastic linear bandits, which we discuss in the next section.

29.2 Stochastic linear bandits with parameter noise

Another way to relate the adversarial and stochastic linear bandit frameworks is to start from the adversarial model and add stochasticity by assuming that θ_t is chosen from some fixed distribution $\nu \in \mathbb{R}^d$. We call the resulting stochastic linear bandit model the **stochastic linear bandit with parameter noise**. This new problem can be trivially reduced to adversarial bandits (assuming the support of ν is bounded). In particular, there is no need to change the action sets. We note in passing that constructing a stochastic environment like this is often the way lower bounds are constructed for adversarial models.

Parameter noise environments form a subset of all possible stochastic environments. To see this, let $\theta = \int x\nu(dx)$ be the mean parameter vector under ν. Then, the loss (or reward) in round t is

$$\langle A_t, \theta_t \rangle = \langle A_t, \theta \rangle + \langle A_t, \theta_t - \theta \rangle \,.$$

Let $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$. By our assumption that ν has mean θ the second term vanishes in expectation, $\mathbb{E}_t[\langle A_t, \theta_t - \theta \rangle] = 0$. This implies that we can make a connection to the 'vanilla' stochastic setting by letting $\tilde{\eta}_t = \langle A_t, \theta_t - \theta \rangle$. Now consider the conditional variance of $\tilde{\eta}_t$:

$$\mathbb{V}_t[\tilde{\eta}_t] = \mathbb{E}_t[\langle A_t, \theta_t - \theta \rangle^2] = A_t^\top \mathbb{E}_t[(\theta_t - \theta)(\theta_t - \theta)^\top] A_t = A_t^\top \Sigma A_t \,, \tag{29.3}$$

where Σ is the covariance matrix of multivariate distribution ν. Eq. (29.3) implies that the variance of the noise $\tilde{\eta}_t$ now depends on the choice of action and in particular the noise variance scales with the length of A_t. This can make parameter noise problems easier. For example, if ν is a Gaussian with identity covariance, then $\mathbb{V}_t[\tilde{\eta}_t] = \|A_t\|_2^2$ so that long actions have more noise than short actions. By contrast, in the usual stochastic linear bandit, the variance of the noise is unrelated to the length of the action. In particular, even the noise accompanying short actions can be large. This makes quite a bit of difference in cases when the action set has both short and long actions. In the standard stochastic model, shorter actions have the disadvantage of having a worse signal-to-noise ratio, which an adversary can exploit.

This calculation also provides the reason for the different guarantees for the unit ball. For stochastic linear bandits with 1-subgaussian noise the regret is $\tilde{O}(d\sqrt{n})$ while in the last chapter we showed that for adversarial linear bandits the regret is $\tilde{O}(\sqrt{dn})$. This discrepancy is explained by the variance of the noise. Suppose that ν is supported on the unit sphere, then the eigenvalues of its covariance matrix sum to 1 and if the learner chooses A_t from the uniform probability measure μ on the sphere, then

$$\mathbb{E}[\mathbb{V}_t[\tilde{\eta}_t]] = \int a^\top \Sigma a d\mu(a) = 1/d \,,$$

By contrast, in the standard stochastic model with 1-subgaussian noise the predictable variation of the noise is just 1. If the adversary were allowed to choose

its loss vectors from the sphere of radius $\sqrt{d}$, then the expected predictable variation would be 1, matching the standard stochastic case, and the regret would scale linearly in d, which also matches the vanilla stochastic case. This example further emphasizes the importance of the assumptions that restrict the choices of the adversary.

The main takeaway of this chapter that the best way to think about the standard adversarial linear model is that it generalizes the stochastic linear bandit model under parameter noise, which is a special case stochastic linear bandits, which oftentimes is easier than the full stochastic linear bandit problem because parameter noise limits the adversary's control of the signal-to-noise ratio experienced by the learner.

29.3 Notes

1 One obvious issue with the stochastic linear bandit model is that the feedbacks may not follow it! It is tempting to try and use adversarial bandits to resolve the resulting unrealizable linear bandit problem where

$$X_t = \langle A_t, \theta \rangle + \eta_t + \varepsilon(A_t)\,,$$

with $\varepsilon : \mathcal{A} \to \mathbb{R}$ some function with small supremum norm. Because $\varepsilon(A_t)$ depends on the chosen action it is not possible to write $X_t = \langle A_t, \theta_t \rangle$ for some θ_t that does not depend on A_t. However, at least in some cases, the idea of using an adversarial linear bandit can be shown to work (cf. Exercise 29.4).

2 For the reduction in Section 29.1 we assumed that $|\langle A_t, \theta \rangle + \eta_t| \leq 1$ almost surely. This is not true for many classical noise models like the Gaussian. One way to overcome this annoyance is to apply the adversarial analysis on the event that $|\langle A_t, \theta \rangle + \eta_t| \leq C$ for some constant $C > 0$ that is sufficiently large that the probability that this event occurs is high. For example, if η_t is a standard Gaussian and $\sup_{a \in \mathcal{A}} |\langle a, \theta \rangle| \leq 1$, then C may be chosen to be $1 + \sqrt{4 \log(n)}$ and the failure event that there exists a t such that $|\langle A_t, \theta \rangle + \eta_t| \geq C$ has probability at most $1/n$ by Theorem 5.1 and a union bound.

3 The mirror descent analysis of adversarial linear bandits also works for stochastic bandits. Recall that mirror descent samples A_t from a distribution with a conditional mean of $\bar{A}_t$ and suppose that $\hat{\theta}_t$ is a conditionally unbiased estimator of θ. Then the regret for a stochastic linear bandit with optimal action a^* can be rewritten as

$$R_n = \mathbb{E}\left[\sum_{t=1}^{n} \langle a^* - A_t, \theta \rangle\right] = \mathbb{E}\left[\sum_{t=1}^{n} \langle a^* - \bar{A}_t, \theta \rangle\right] = \mathbb{E}\left[\sum_{t=1}^{n} \langle a^* - \bar{A}_t, \hat{\theta}_t \rangle\right].$$

Except that we have switched from losses to gains, this is now in the standard format necessary for the analysis of mirror descent. In general for the stochastic

setting the covariance of the least squares estimator $\hat{\theta}_t$ will not be the same as in the adversarial setting, however, which leads to different results. When $\hat{\theta}_t$ is biased, the bias term can be incorporated into the above formula and then bounded separately.

4 Consider a stochastic bandit with $\mathcal{A}$ the unit ball and $X_t = \langle A_t, \theta \rangle + \eta_t$ where $\eta_t \in [-1, 1]$ almost surely and θ is also in the unit ball. Adapting the analysis of the algorithm in Section 28.4 leads to a bound of $O(d\sqrt{n \log(n)})$. Essentially the only change is the variance calculation in Eq. (28.12), which increases by roughly a factor of d. The details of this calculation are left to you in Exercise 29.5.

29.4 Bibliographic remarks

Linear bandits on the sphere with parameter noise have been studied by Carpentier and Munos [2012]. However they consider the case where the action-set is the sphere and the components of the noise are independent so that $X_t = \langle A_t, \theta + \eta_t \rangle$ where the coordinates of $\eta_t \in \mathbb{R}^d$ are independent with unit variance. In this case the predictable variation is $\mathbb{V}[X_t \mid A_t] = \sum_{i=1}^d A_{ti}^2 = 1$ for all actions A_t and the parameter noise is equivalent to the standard model. We are not aware of any systematic studies of parameter noise in the stochastic setting. With only a few exceptions, the impact on the regret of the action-set and adversaries choices is not well understood beyond the case where $\mathcal{A}$ is an ℓ_p-ball and the adversary chooses losses from the polar of $\mathcal{A}$. A variety of lower bounds illustrating the complications are given by Shamir [2015]. Perhaps the most informative is the observation that obtaining $O(\sqrt{dn})$ regret is not possible when $\mathcal{A} = \{a + x : \|x\|_2 \leq 1\}$ is a shifted unit ball with $a = (2, 0, \ldots, 0)$, which also follows from our reduction in Section 29.1.

29.5 Exercises

29.1 Complete the claims made at the end of Section 29.1. In particular, show that the bandit algorithm of Theorem 28.4 achieves an $O(d^2\sqrt{n})$ expected regret when applied to a stochastic bandit problem where the noise $(\eta_t)_t$ sequence is bounded by a constant and when used with the learning rate as described in that theorem. Further, show that by an appropriate adjustment of the learning rate, the regret can be improved to $O(d\sqrt{n})$.

29.2 Let $\mathcal{A} \subset \mathcal{R}^d$ be an action set. Take an adversarial linear bandit algorithm that enjoys a worst-case guarantee B_n on its n-round expected regret R_n when the adversary is restricted to playing $(\theta_t)_t$ in the polar $\mathcal{A}^\circ$ of $\mathcal{A}$. Show that if this algorithm is used in a stochastic linear bandit problem with parameter noise (that is, $\theta_t \sim \nu$) and the support of ν is a subset of the polar of $\mathcal{A}^\circ$ then the expected

regret R'_n is still bounded by B_n. Derive a bound on the expected regret R'_n in the stochastic problem when the restriction on ν is replaced by an assumption that $\sup_{a \in \mathcal{A}} \langle a, \theta_t - \theta \rangle$ has a bounded magnitude.

29.3 Modify LinUCB to make it (potentially) better for stochastic bandits under parameter noise. Show that the regret improves.

29.4 Complete the details to prove the claims made in Note 1. In particular, prove that for $\mathcal{A} = B_2^d$ there exists a universal constant $C > 0$ such that the expected regret R_n of an appropriately tuned version of the bandit algorithm of Theorem 28.4 satisfies $R_n \leq C(d\sqrt{n} + \varepsilon n)$, where $\varepsilon = \sup_{a \in \mathcal{A}} \varepsilon(a)$.

29.5 Complete the details to prove the claims made in Note 4.

You will need to repeat the analysis in Eq. (28.12), update the learning rate and check the bounds on the norm of the estimators.