22 Stochastic Linear Bandits with Finitely Many Arms

The optimal design problem has immediate applications to stochastic linear bandits. In Chapter 19 we developed a linear version of the upper confidence bound algorithm that achieves a regret of $R_n = \tilde{O}(d\sqrt{n})$. The only required assumptions were that the sequence of available action-sets were bounded. In this short chapter we consider a more restricted setting where:

1 *Fixed finite action set:* The set of actions available in round t is $\mathcal{A} \subset \mathbb{R}^d$ and $|\mathcal{A}| = K$ for some natural number K.
2 *Subgaussian rewards:* The reward is $X_t = \langle \theta_*, A_t \rangle + \eta_t$ where η_t is conditionally 1-subgaussian:

$$\mathbb{E}[\exp(\lambda \eta_t) | A_1, \eta_1, \ldots, A_{t-1}] \leq \exp(\lambda^2/2) \qquad \text{almost surely for all } \lambda \in \mathbb{R}\,.$$

3 *Bounded mean rewards:* $\Delta_a = \max_{b \in \mathcal{A}} \langle \theta_*, b - a \rangle \leq 1$ for all $a \in \mathcal{A}$.

The key difference is that now the set of actions is finite and does not change with time. Under these conditions it becomes possible to design a policy such that

$$R_n = O\left(\sqrt{dn \log(nK)}\right)\,.$$

When K is small this bound improves the regret by a factor of $d^{1/2}$, which in some regimes is large enough to be worth the effort. The core idea is to introduce phases of determinisim into the algorithm so that within each phase the actions are chosen independently from the rewards. This decoupling allows us to make use of the tighter confidence bounds available in the fixed design setting as discussed in the previous chapter. The choice of policy within each phase uses the solution to an optimal design problem to minimize the number of required samples to eliminate arms that are far from optimal.

THEOREM 22.1 *With probability at least $1 - \delta$ the regret of Algorithm 11 is at most:*

$$R_n \leq C \sqrt{nd \log\left(\frac{K \log(n)}{\delta}\right)}\,,$$

where $C > 0$ is a universal constant. If $\delta = O(1/n)$, then $\mathbb{E}[R_n] \leq C\sqrt{nd \log(Kn)}$ for appropriately chosen universal constant $C > 0$.

Input $\mathcal{A} \subset \mathbb{R}^d$ and δ

Step 0 Set $\ell = 1$ and let $\mathcal{A}_1 = \mathcal{A}$

Step 1 Let $t_\ell = t$ be the current timestep and find G-optimal design $\pi_\ell : \mathcal{A}_\ell \to [0, 1]$ that maximizes

$$\log\det V(\pi_\ell) \text{ subject to } \sum_{a \in \mathcal{A}_\ell} \pi_\ell(a) = 1$$

Step 2 Let $\varepsilon_\ell = 2^{-\ell}$ and

$$T_\ell(a) = \left\lceil \frac{2\pi(a)}{\varepsilon_\ell^2} \log\left(\frac{K\ell(\ell+1)}{\delta}\right) \right\rceil \text{ and } T_\ell = \sum_{a \in \mathcal{A}_\ell} T_\ell(a)$$

Step 3 Choose each action $a \in \mathcal{A}_\ell$ exactly $T_a(\ell)$ times

Step 4 Calculate empirical estimate:

$$\hat{\theta} = V_\ell^{-1} \sum_{t=t_\ell}^{t_\ell + T_\ell} A_t X_t$$

Step 5 Eliminate low rewarding arms:

$$\mathcal{A}_{\ell+1} = \left\{ a \in \mathcal{A}_\ell : \max_{b \in \mathcal{A}_\ell} \langle \hat{\theta}_\ell, b - a \rangle \geq 2\varepsilon_\ell \right\} .$$

Algorithm 11: Phased elimination with G-optimal exploration

The proof of this theorem follows relatively directly from the high-probability correctness of the confidence intervals used to eliminate low-rewarding arms. We leave the details to the reader in Exercise 22.1.

22.1 Bibliographic remarks

Algorithm 11 is a combination of several existing ideas. The use of phases to decouple the dependence of the design and the outcomes is originally due to Auer [2002], where a more complicated version of the presented problem is solved in which the action set is permitted to change with time. The complexity of the analysis unfortunately prohibited us from presenting these ideas here. Phased approaches have since appeared in many places, but the most similar is the work on spectral bandits by Valko et al. [2014]. Neither of these works used the Kiefer–Wolfowitz theorem. This idea is taken from the literature on adversarial linear bandits where John's ellipsoid has been used to define exploration policies [Bubeck et al., 2012]. For more details on adversarial linear bandits read on to Part VI.

SupLinRel, LinRel, Chu et al. [2011].

22.2 Exercises

22.1 In this exercise you will prove Theorem 22.1.

(a) The first step is to use Theorem 21.1 (and the preceding comments) to show that the length of the ℓth phase is bounded by

$$T_\ell \le \frac{2d}{\varepsilon_\ell^2} \log\left(\frac{K\ell(\ell+1)}{\delta}\right) + \frac{d(d+3)}{2}$$

(b) Let $a^* \in \operatorname{argmax}_{a\in\mathcal{A}} \langle a, \theta_* \rangle$ be the optimal arm and use Theorem 21.1 to show that

$$\mathbb{P}\left(\text{exists phase } \ell \text{ such that } a^* \notin A_\ell\right) \le \frac{\delta}{K}\,.$$

(c) For action a define $\ell_a = \min\{\ell : \Delta_a < 2\varepsilon_\ell\}$ to be the first phase where the suboptimality gap of arm a is smaller than $2\varepsilon_\ell$. Show that

$$\mathbb{P}\left(a \in \mathcal{A}_{\ell_a}\right) \le \frac{\delta}{K}$$

(d) Show that with probability at least $1-\delta$ the regret is bounded by

$$R_n \le C\sqrt{dn \log\left(\frac{K}{\delta}\right)}\,,$$

where $C > 0$ is a universal constant.

(e) Show that this implies Theorem 22.1 for the given choice of δ.