Lecture 16

Instructor: Haipeng Luo

1 Adversarial Linear Bandit and Exp2

Previously we have discussed bandit problems under some stochastic assumptions. In this lecture we come back to adversarial setting and discuss the adversarial linear bandit problem, a natural generalization of adversarial multi-armed bandit. Specifically, the setting is as follows: at each time $t = 1, \ldots, T$,

1. learner picks action $a_t \in A \subset \mathbb{R}^d$ while simultaneously environment picks $\ell_t \in \mathbb{R}^d$;
2. learner suffers and observes $a_t^\top \ell_t$.

We assume that $\max_{a \in A} \|a\|_2 \leq B$ for some constant $B > 0$, $|a_t^\top \ell_t| \leq 1$, and A is a finite set with cardinality $|A| = K$. For simplicity we also assume that the environment is oblivious. The expected regret of the learner is

$$\mathbb{E}[\mathcal{R}_T] = \mathbb{E}\left[\sum_{t=1}^T a_t^\top \ell_t\right] - \min_{a \in A} \sum_{t=1}^T a^\top \ell_t.$$

Compared to the stochastic linear bandit, the main difference is that the action set A for the learner is now fixed while the loss vector ℓ_t is changing over time. Adversarial multi-armed bandit a clearly a special case of this model with A being the standard basis of $\mathbb{R}^d$. However, on the other hand one can also see linear bandit as a special case of multi-armed bandit with K actions, completely ignoring the underlying linear structure of the losses. This gives a trivial solution with regret of order $\mathcal{O}(\sqrt{TK})$, independent of d. However, as we will see soon, by exploiting the linear structure, one can in fact achieve regret $\mathcal{O}(\sqrt{dT \ln K})$, much better than the trivial solution as long as $d \ll K$.

Nevertheless, we can still borrow the idea of Exp3, the classic solution for adversarial multi-armed bandit. Recall that Exp3 is simply feeding the Hedge algorithm with unbiased loss estimators. Due to the similarity, it is natural to try the same idea here:

1. play $a_t \sim p_t \in \Delta(K)$ and observe $a_t^\top \ell_t$;
2. construct unbiased loss estimator $\widehat{\ell}_t$ based on $a_t^\top \ell_t$;
3. update $p_{t+1}(a) \propto \exp(\eta \sum_{\tau=1}^t a^\top \widehat{\ell}_t)$.

Note that instead of estimating the loss of each action, here we estimate the underlying loss vector ℓ_t directly. The key is therefore to come up with this estimator. Recall that in the last lecture we also constructed some estimator for the true parameter $\theta^\star$ for stochastic linear bandit, and it was simply based on least square regression using observed data. The difficulty here is that we have only one single observation about ℓ_t. However, because of the randomization in picking a_t, it turns out that one can construct the following estimator using a very similar formula as for stochastic linear bandit

$$\widehat{\ell}_t = M_t^{-1} a_t a_t^\top \ell_t \quad \text{where} \quad M_t = \sum_{a \in A} p_t(a) a a^\top = \mathbb{E}_{a \sim p_t}\left[a a^\top\right].$$

Note that although ℓ_t appears in this formula, the dependence is only through $a_t^\top \ell_t$, a quantity that we indeed observe. M_t is assumed to be invertible here, which is equivalent to assuming A is full

rank. Unlike the case for stochastic linear bandit, this is in fact without loss of generality. Indeed, if A is not full rank, then before the game starts one can do a preprocessing step to project the actions into a subspace with lower dimension. (Alternatively, one can also simply replace "inverse" by "pseudo-inverse" and verify that it does not change the final results.)

With $\mathbb{E}_t$ being the conditional expectation with respect to the random draw of a_t, direct calculations show that this estimator is indeed unbiased:

$$\mathbb{E}_t[\widehat{\ell}_t] = M_t^{-1}\mathbb{E}\left[a_t a_t^\top\right]\ell_t = M_t^{-1}M_t\ell_t = \ell_t.$$

Also, when A is the standard basis of $\mathbb{R}^d$, M_t is a diagonal matrix with $M_t(a,a) = p_t(a)$ and thus this recovers the importance weighted estimator used in Exp3.

There is one more detail we need to take care of before applying the result of Hedge. Recall that in the analysis of Hedge, we use the inequality $e^{-x} \le 1 - x + x^2$ for $x \ge 0$ where x corresponds to $\eta a^\top \widehat{\ell}_t$ here. While for multi-armed bandit this is indeed non-negative (or at least can be made to be nonnegative by shifting the losses), this is not true anymore for the general linear case, even if all $a \in A$ and ℓ_t have nonnegative coordinates.

Fortunately, the inequality in fact holds whenever $x \ge -1$. So at least negativity is not necessarily an issue. We do, however, still need to control the magnitude of $\eta a^\top \widehat{\ell}_t$. We will ensure this by enforcing an explicit exploration. Specifically, let $q \in \Delta(K)$ be a fixed exploration distribution over the actions in A and γ be some exploration parameter. We modify the algorithm as (first two steps remain the same):

1. play $a_t \sim p_t \in \Delta(K)$ and observe $a_t^\top \ell_t$;
2. construct unbiased loss estimator $\widehat{\ell}_t = M_t^{-1} a_t a_t^\top \ell_t$;
3. update $p'_{t+1}(a) \propto \exp(\eta \sum_{\tau=1}^t a^\top \widehat{\ell}_t)$;
4. compute $p_{t+1} = (1-\gamma)p'_{t+1} + \gamma q$.

With the explicit exploration, we can show that the magnitude of $|a^\top \widehat{\ell}_t|$ is controlled by the minimum eigenvalue of $\mathbb{E}_{a\sim q}[aa^\top]$ as shown by the following lemma.

Lemma 1. *If $\eta \le \frac{\gamma \lambda_{min}}{B^2}$ where λ_{min} is the minimum eigenvalue of $\mathbb{E}_{a\sim q}[aa^\top]$, then $\eta|a^\top \widehat{\ell}_t| \le 1$.*

Proof. Let $M_t = \sum_{i=1}^d \lambda_i v_i v_i^\top$ be the eigendecomposition of M_t so that $\lambda_1 \le \ldots, \lambda_d$. Note that because $M_t = (1-\gamma)\mathbb{E}_{a\sim p'_t}[aa^\top] + \gamma \mathbb{E}_{a\sim q}[aa^\top]$, its smallest eigenvalue λ_1 is lower bounded by $\gamma\lambda_{\min}$. Therefore, we have

$$|a^\top \widehat{\ell}_t| = |a^\top M_t^{-1} a_t||a_t^\top \ell_t| \le |a^\top M_t^{-1} a_t| = \left|\sum_{i=1}^d \frac{1}{\lambda_i}(a^\top v_i)(a_t^\top v_i)\right| \le \frac{B^2}{\lambda_1} \le \frac{B^2}{\gamma\lambda_{\min}},$$

where the first inequality is by the assumption $|a_t^\top \ell_t| \le 1$ and the second inequality is by Cauchy-Schwarz inequality

$$\left|\sum_{i=1}^d \frac{1}{\lambda_i}(a^\top v_i)(a_t^\top v_i)\right| \le \frac{1}{\lambda_1}\sum_{i=1}^d \left|(a^\top v_i)(a_t^\top v_i)\right| \le \frac{1}{\lambda_1}\sqrt{\left(\sum_{i=1}^d (a^\top v_i)^2\right)\left(\sum_{i=1}^d (a_t^\top v_i)^2\right)}$$

$$= \frac{1}{\lambda_1}\sqrt{\left(a^\top\left(\sum_{i=1}^d v_i v_i^\top\right)a\right)\left(a_t^\top\left(\sum_{i=1}^d v_i v_i^\top\right)a_t\right)} = \frac{1}{\lambda_1}\|a\|_2\|a_t\|_2 \le \frac{B^2}{\lambda_1}.$$

This finishes the proof. □

Roughly speaking, the lemma above implies that it is desirable to pick q such that it explores every direction with reasonable probability. The resulting algorithm is called by many names in different works, such as Exp2 (Expanded Exp) or GeometricHedge [Dani et al., 2008, Cesa-Bianchi and Lugosi, 2012, Bubeck et al., 2012]. We are now ready to prove the following regret bound.

Theorem 1. *If $\eta \leq \frac{\gamma\lambda_{min}}{B^2}$, then Exp2 ensures*

$$\mathbb{E}\left[\sum_{t=1}^{T} a_t^\top \ell_t\right] - \min_{a\in A}\sum_{t=1}^{T} a^\top \ell_t \leq \frac{\ln K}{\eta} + 2\gamma T + \eta T d.$$

Setting $\eta = \sqrt{\frac{\ln K}{\left(\frac{2B^2}{\lambda_{min}}+d\right)T}}$ and $\gamma = \frac{B^2\eta}{\lambda_{min}}$ leads to a regret of order $\mathcal{O}\left(\sqrt{\left(\frac{2B^2}{\lambda_{min}}+d\right)T\ln K}\right)$.

Proof. By Lemma 1 and the analysis of Hedge, we have for any $a_\star \in A$,

$$\sum_{t=1}^{T}\sum_{a\in A} p'_t(a)(a^\top \widehat{\ell}_t) - \sum_{t=1}^{T} a_\star^\top \widehat{\ell}_t \leq \frac{\ln K}{\eta} + \eta\sum_{t=1}^{T}\sum_{a\in A} p'_t(a)(a^\top \widehat{\ell}_t)^2.$$

Plugging $p'_t(a) = \frac{p_t(a)-\gamma q(a)}{1-\gamma}$, multiplying both sides by $1-\gamma$, and rearranging give

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{a\in A} p_t(a)(a^\top \widehat{\ell}_t) - \sum_{t=1}^{T} a_\star^\top \widehat{\ell}_t \\
&\leq \frac{(1-\gamma)\ln K}{\eta} + \gamma\sum_{t=1}^{T}\sum_{a\in A} q(a)(a^\top \widehat{\ell}_t) - \gamma\sum_{t=1}^{T} a_\star^\top \widehat{\ell}_t + \eta\sum_{t=1}^{T}\sum_{a\in A}(p_t(a)-\gamma q(a))(a^\top \widehat{\ell}_t)^2 \\
&\leq \frac{\ln K}{\eta} + \gamma\sum_{t=1}^{T}\sum_{a\in A} q(a)(a^\top \widehat{\ell}_t) - \gamma\sum_{t=1}^{T} a_\star^\top \widehat{\ell}_t + \eta\sum_{t=1}^{T}\sum_{a\in A} p_t(a)(a^\top \widehat{\ell}_t)^2.
\end{aligned}$$

By the unbiasedness of $\widehat{\ell}_t$, taking expectation on both sides and using $|a^\top \ell_t| \leq 1$ lead to

$$\mathbb{E}\left[\sum_{t=1}^{T} a_t^\top \ell_t\right] - \min_{a\in A}\sum_{t=1}^{T} a^\top \ell_t \leq \frac{\ln K}{\eta} + 2\gamma T + \eta\sum_{t=1}^{T}\sum_{a\in A}\mathbb{E}\left[p_t(a)(a^\top \widehat{\ell}_t)^2\right].$$

To bound the last term, note that

$$\begin{aligned}
&\mathbb{E}_t\left[p_t(a)(a^\top \widehat{\ell}_t)^2\right] = p_t(a)\mathbb{E}_t\left[(a_t^\top \ell_t)^2 a^\top M_t^{-1} a_t a_t^\top M_t^{-1} a\right] \\
&\leq p_t(a) a^\top M_t^{-1}\mathbb{E}_t\left[a_t a_t^\top\right] M_t^{-1} a = p_t(a) a^\top M_t^{-1} a = \mathrm{TR}(M_t^{-1}(p_t(a) a a^\top))
\end{aligned}$$

and thus

$$\sum_{a\in A}\mathbb{E}_t\left[p_t(a)(a^\top \widehat{\ell}_t)^2\right] \leq \mathrm{TR}(M_t^{-1} M_t) = d,$$

which completes the proof. □

Note that the sum of the eigenvalues of $\mathbb{E}_{a\sim q}[aa^\top]$ is bounded by B^2 (since $\mathrm{TR}(\mathbb{E}_{a\sim q}[aa^\top]) = \mathbb{E}_{a\sim q}[\mathrm{TR}(aa^\top)]$). Therefore, if the eigenvalues of $\mathbb{E}_{a\sim q}[aa^\top]$ are all close to each other, we have $\lambda_{\min} = \Omega(B^2/d)$, which then leads to a regret bound of $\mathcal{O}(\sqrt{dT\ln K})$, proven to be optimal in [Dani et al., 2008] (note that the parameter B in fact does not play a role in the optimal regret). This again suggests that q should uniformly explore different directions in $\mathbb{R}^d$.

In general, it turns out that one can always find a q over a subset of A with special geometric properties such that $\lambda_{\min} = \Omega(B^2/d)$ [Bubeck et al., 2012]. However, for many examples (such as those we discuss in the next section), simply setting q to be a uniform distribution over A is enough.

2 Examples

The most important example of linear bandit is the combinatorial bandit problem, where $A \subset \{0,1\}^d$ represents a set of combinatorial concepts chosen from d basic elements. Examples include spanning trees, paths, cuts, Hamiltonian cycles, permutations, and many more. Below we will apply the general results to two of these problems by computing $\lambda_{\min}$ in each case. In both examples

q is uniform over A and we will drop the subscript $a \sim q$ in the expectation for conciseness. We will use the fact $\lambda_{\min} = \min_{\|v\|_2=1} v^\top \mathbb{E}\left[aa^\top\right] v = \min_{\|v\|_2=1} \mathbb{E}\left[(a^\top v)^2\right]$ and

$$\mathbb{E}\left[(a^\top v)^2\right] = \mathbb{E}\left[\left(\sum_{i=1}^d a(i)v(i)\right)^2\right] = \mathbb{E}\left[\sum_{i=1}^d a(i)^2 v(i)^2 + \sum_{i\neq j} a(i)a(j)v(i)v(j)\right]$$
$$= \sum_{i=1}^d \Pr(a(i)=1)v(i)^2 + \sum_{i\neq j} \Pr(a(i)=a(j)=1)v(i)v(j).$$

Note that in general the time complexity of the algorithm is $\mathcal{O}(K)$ per round, which is often prohibitively large. However, in some cases one can actually implement the algorithm much more efficiently using techniques such as dynamic programming.

Hypercube. The first example is when A is the entire hypercube $\{0,1\}^d$. This corresponds to a setting where there are d items and each time we can pick any subset of them and observe the sum of the losses of the selected items. Now note that $\Pr(a(i)=1) = 1/2$ and $\Pr(a(i)=a(j)=1) = 1/4$, we have for any v with $\|v\|_2 = 1$,

$$\mathbb{E}\left[(a^\top v)^2\right] = \frac{1}{2}\|v\|_2^2 + \frac{1}{4}\sum_{i\neq j} v(i)v(j) = \frac{1}{4}\|v\|_2^2 + \frac{1}{4}\left(\sum_{i=1}^d v(i)\right)^2 \geq \frac{1}{4}.$$

The minimum is achievable as long as $\sum_{i=1}^d v(i) = 0$, which means $\lambda_{\min} = 1/4$. Together with $B = \sqrt{d}$ and $K = 2^d$, this implies a regret of order $\mathcal{O}(d\sqrt{T})$ for Exp2.

m-sets. The next example is when $A = \{a \in \{0,1\}^d : \|a\|_1 = m\}$, that is, each time we can only pick exactly m items. Similarly, noting that $\Pr(a(i)=1) = \binom{d-1}{m-1}/\binom{d}{m}$ and $\Pr(a(i)=a(j)=1) = \binom{d-2}{m-2}/\binom{d}{m}$, we have for any v with $\|v\|_2 = 1$

$$\mathbb{E}\left[(a^\top v)^2\right] = \frac{\binom{d-1}{m-1}}{\binom{d}{m}}\|v\|_2^2 + \frac{\binom{d-2}{m-2}}{\binom{d}{m}}\sum_{i\neq j} v(i)v(j)$$
$$= \left(\frac{m}{d} - \frac{m(m-1)}{d(d-1)}\right)\|v\|_2^2 + \frac{m(m-1)}{d(d-1)}\left(\sum_{i=1}^d v(i)\right)^2 \geq \frac{m(d-m)}{d(d-1)}.$$

Together with $K = \binom{d}{m}$ and $B = \sqrt{m}$, as long as $m = o(d)$ the regret of Exp2 is $\mathcal{O}(\sqrt{dmT\ln\frac{d}{m}})$.

References

Sébastien Bubeck, Nicolò Cesa-Bianchi, and Sham Kakade. Towards minimax policies for online linear optimization with bandit feedback. In *25th Annual Conference on Learning Theory*, 2012.

Nicolò Cesa-Bianchi and Gábor Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422, 2012.

Varsha Dani, Sham M Kakade, and Thomas P Hayes. The price of bandit information for online optimization. In *Advances in Neural Information Processing Systems 21*, 2008.