18.657: Mathematics of Machine Learning

Lecturer: Philippe Rigollet — Lecture 21
Scribe: Ali Makhdoumi — Nov. 25, 2015

6. LINEAR BANDITS

Recall form last lectures that in prediction with expert advise, at each time t, the player plays $a_t \in \{e_1, \ldots, e_k\}$ and the adversary plays z_t such that $l(a_t, z_t) \leq 1$ for some loss function. One example of such loss function is linear function $l(a_t, z_t) = a_t^T z_t$ where $|z_t|_\infty \leq 1$. Linear bandits are a more general setting where the player selects an action $a_t \in \mathcal{A} \subset \mathbb{R}^k$, where $\mathcal{A}$ is a convex set and the adversary selects $z_t \in \mathcal{Z}$ such that $|z_t^T a_t| \leq 1$. Similar to the prediction with expert advise, the regret is defined as

$$R_n = \mathbb{E}\left[\sum_{t=1}^n A_t^T z_t\right] - \min_{a \in \mathcal{K}} \sum_{t=1}^n a^T z_t,$$

where A_t is a random variable in $\mathcal{A}$. Note that in the prediction with expert advise, the set $\mathcal{A}$ was essentially a polyhedron and we had $\min_{a\in\mathcal{K}} \sum_{t=1}^n a^T z_t = \min_{1 \leq j \leq k} e_j^T z_t$. However, in the linear bandit setting the minimizer of $a^T z_t$ can be any point of the set $\mathcal{A}$ and essentially the umber experts that the player tries to "compete" with are infinity. Similar, to the prediction with expert advise we have two settings:

1 **Full feedback:** after time t, the player observes z_t.

2 **Bandit feedback:** after time t, the player observes $A_t^T z_t$, where A_t is the action that player has chosen at time t.

We next, see if we can use the bounds we have developed in the prediction with expert advise in this setting. In particular, we have shown the following bounds for prediction with expert advise:

1 **Prediction with k expert advise, full feedback:** $R_n \leq \sqrt{2n \log k}$.

2 **Prediction with k expert advise, bandit feedback:** $R_n \leq \sqrt{2nk \log k}$.

The idea to deal with linear bandits is to discretize the set $\mathcal{A}$. Suppose that $\mathcal{A}$ is bounded (e.g., $\mathcal{A} \subset B_2$, where B_2 is the l_2 ball in $\mathbb{R}^k$). We can use a $\frac{1}{n}$-covering of $\mathcal{A}$ which we have shown to be of size (smaller than)$O(n^k)$. This means there exist $y_1, \ldots, y_{|\mathcal{N}|}$ such that for any $a \in \mathcal{A}$, there exist y_i such that $||y_i - a|| \leq \frac{1}{n}$. We now can bound the regret for general case, where the experts can be any point in $\mathcal{A}$, based on the regret on the discrete set, $\mathcal{N} = \{y_1, \ldots, y_{|\mathcal{N}|}\}$,as follows.

$$\begin{aligned} R_n &= \mathbb{E}\left[\sum_{t=1}^n A_t^T z_t\right] - \min_{a \in \mathcal{A}} \sum_{t=1}^n a^T z_t \\ &= \mathbb{E}\left[\sum_{t=1}^n A_t^T z_t\right] - \min_{a \in \mathcal{N}} \sum_{t=1}^n a^T z_t + o(1). \end{aligned}$$

Therefore, we restrict actions A_t to a combination of the actions that belong to $\{y_1, \ldots, y_{|\mathcal{N}|}\}$ (we can always do this), then using the bounds for the prediction with expert advise, we obtain the following bounds:

1 **Linear bandit, full feedback:** $R_n \leq \sqrt{2n\log(n^k)} = O(\sqrt{kn\log n})$, which in terms of dependency to n is of order $O(\sqrt{n})$ that is what we expect to have.

2 **Linear bandit, bandit feedback:** $R_n \leq \sqrt{2nn^k\log(n^k)} = \Omega(n)$, which is useless in terms of dependency of n as we expect to obtain $O(\sqrt{n})$ behavior.

The topic of this lecture is to provide bounds for the linear bandit in the bandit feedback.
Problem Setup: Let us recap the problem formulation:

- at time t, player chooses action $a_t \in \mathcal{A} \subset [-1,1]^k$.
- at time t, adversary chooses $z_t \in \mathcal{Z} \subset \mathbb{R}^k$, where $a_t^T z_t = \langle a_t, z_t\rangle \in [0,1]$.
- Bandit feedback: player observes $\langle a_t, z_t\rangle$(rather than z_t in the full feedback setup).

Literature: $O(n^{3/4})$ regret bound has been shown in [BB04]. Later on this bound has been improved to $O(n^{2/3})$ in [BK04] and [VH06] with "Geometric Hedge algorithm", which we will describe and analyze below. We need the following assumption to show the results:
Assumption: There exist δ such that $\delta e_1, \ldots, \delta e_k \in \mathcal{A}$. This assumption guarantees that $\mathcal{A}$ has full-dimension around zero.
We also discretize $\mathcal{A}$ with a $\frac{1}{n}$-net of size Cn^k and only consider the resulting discrete set and denote it by $\mathcal{A}$, where $|\mathcal{A}| \leq (3n)^k$. All we need to do is to bound

$$R_n = \mathbb{E}\left[\sum_{t=1}^n A_t^T z_t\right] - \min_{a\in\mathcal{A}} \sum_{t=1}^n a^T z_t.$$

For any t and a, we define

$$p_t(a) = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \hat{z}_s^T a\right)}{\sum_{a\in\mathcal{A}} \exp\left(-\eta \sum_{s=1}^{t-1} \hat{z}_s^T a\right)},$$

where η is a parameter (that we will choose later) and $\hat{z}_t$ is defined to incorporate the idea of exploration versus exploitation. The algorithm which is termed *Geometric Hedge Algorithm* is as follows:
At time t we have

- **Exploitation:** with probability $1-\gamma$ draw a_t according to p_t and let $\hat{z}_t = 0$.
- **Exploration:** with probability $\frac{\gamma}{k}$ let $a_t = \delta e_j$ for some $1 \leq j \leq k$ and $\hat{z}_t = \frac{k}{\delta^2\gamma}\langle a_t, z_t\rangle a_t = \frac{k}{\gamma} z_t^{(j)} e_j$.

Note that δ is the the parameter that we have by assumption on the set $\mathcal{A}$, and η and γ are the parameters of the algorithm that we shall choose later.

Theorem: Using Geometric Hedge algorithm for linear bandit with bandit feedback, with $\gamma = \frac{1}{n^{1/3}}$ and $\eta = \sqrt{\frac{\log n}{kn^{4/3}}}$, we have

$$\mathbb{E}[R_n] \leq Cn^{2/3}\sqrt{\log n}\; k^{3/2}.$$

Proof. Let the overall distribution of a_t be q_t defined as $q_t = (1-\gamma)p_t + \gamma U$, where U is a uniform distribution over the set $\{\delta e_1, \ldots, \delta e_k\}$. Under this distribution, $\hat{z}_t$ is an unbiased estimator of z_t, i.e.,

$$\mathbb{E}_{a_t \sim q_t}[\hat{z}_t] = 0(1-\gamma) + \sum_{j=1}^{k} \frac{\gamma}{k}\frac{k}{\gamma} z_t^{(j)} e_j = z_t.$$

following the same lines of the proof that we had for analyzing exponential weight algorithm, we will define

$$w_t = \sum_{a\in\mathcal{A}} \exp\left(-\eta \sum_{s=1}^{t-1} a^T \hat{z}_s\right).$$

We then have

$$\begin{aligned}
\log\left(\frac{w_{t+1}}{w_t}\right) &= \log\left(\sum_{a\in\mathcal{A}} p_t(a)\exp\left(-\eta a^T \hat{z}_t\right)\right) \\
&\overset{e^{-x}\le 1-x+\frac{x^2}{2}}{\le} \log\left(\sum_{a\in\mathcal{A}} p_t(a)\left(1-\eta a^T\hat{z}_t + \frac{1}{2}\eta^2 (a^T\hat{z}_s)^2\right)\right) \\
&= \log\left(1+\sum_{a\in\mathcal{A}} p_t(a)\left(-\eta a^T\hat{z}_t + \frac{1}{2}\eta^2 (a^T\hat{z}_t)^2\right)\right) \\
&\overset{\log(1+x)\le x}{\le} \sum_{a\in\mathcal{A}} p_t(a)\left(-\eta a^T\hat{z}_t + \frac{1}{2}\eta^2 (a^T\hat{z}_t)^2\right).
\end{aligned}$$

Taking expectation from both sides leads to

$$\begin{aligned}
&\mathbb{E}_{a_t\sim q_t}\left[\log\left(\frac{w_{t+1}}{w_t}\right)\right] \le -\eta\mathbb{E}_{a_t\sim q_t}\left[\sum_{a\in\mathcal{A}} p_t(a)a^T\hat{z}_t\right] + \frac{\eta^2}{2}\mathbb{E}_{a_t\sim q_t}\left[\sum_{a\in\mathcal{A}} p_t(a)(a^T\hat{z}_t)^2\right] \\
&= -\eta\mathbb{E}_{a_t\sim p_t}\left[a_t^T\hat{z}_t\right] + \frac{\eta^2}{2}\mathbb{E}_{a_t\sim q_t}\left[\sum_{a\in\mathcal{A}} p_t(a)(a^T\hat{z}_t)^2\right] \\
&\overset{q_t=(1-\gamma)p_t+\gamma U}{=} \frac{-\eta}{1-\gamma}\mathbb{E}_{a_t\sim q_t}\left[a_t^T\hat{z}_t\right] + \eta\frac{\gamma}{1-\gamma}\mathbb{E}_{a_t\sim U}\left[a_t^T\hat{z}_t\right] + \frac{\eta^2}{2}\mathbb{E}_{a_t\sim q_t}\left[\sum_{a\in\mathcal{A}} p_t(a)(a^T\hat{z}_t)^2\right] \\
&\overset{a_t^T z_t\le 1}{\le} \frac{-\eta}{1-\gamma}\mathbb{E}_{a_t\sim q_t}\left[a_t^T\hat{z}_t\right] + \frac{\eta\gamma}{1-\gamma} + \frac{\eta^2}{2}\mathbb{E}_{a_t\sim q_t}\left[\sum_{a\in\mathcal{A}} p_t(a)(a^T\hat{z}_t)^2\right].
\end{aligned}$$

We next, take summation of the previous relation for $t = 1$ up to n and use a telescopic cancellation to obtain

$$\begin{aligned}
\mathbb{E}\left[\log w_{n+1}\right] &\le \mathbb{E}\left[\log w_1\right] - \frac{\eta}{1-\gamma}\mathbb{E}\left[\sum_{t=1}^{n} a_t^T\hat{z}_t\right] + \frac{\eta\gamma}{1-\gamma}n + \frac{\eta^2}{2}\mathbb{E}\left[\sum_{t=1}^{n}\sum_{a\in\mathcal{A}} p_t(a)(a^T\hat{z}_t)^2\right] \\
&\le \mathbb{E}\left[\log w_1\right] - \eta\mathbb{E}\left[\sum_{t=1}^{n} a_t^T\hat{z}_t\right] + \frac{\eta\gamma}{1-\gamma}n + \frac{\eta^2}{2}\mathbb{E}\left[\sum_{t=1}^{n}\sum_{a\in\mathcal{A}} p_t(a)(a^T\hat{z}_t)^2\right]. \qquad (6.1)
\end{aligned}$$

Note that for all $a^* \in \mathcal{A}$ we have

$$\log(w_{n+1}) = \log\left(\sum_{a\in\mathcal{A}} \exp\left(-\eta\sum_{s=1}^{n} a^T\hat{z}_s\right)\right) \geq -\eta\sum_{s=1}^{n}\langle a^*, \hat{z}_s\rangle.$$

Using $\mathbb{E}[\hat{z}_s] = z_s$, leads to

$$\mathbb{E}\left[\log(w_{n+1})\right] \geq -\eta\sum_{s=1}^{n}\langle a^*, z_s\rangle. \tag{6.2}$$

We also have that

$$\log(w_1) = \log|\mathcal{A}| \leq 2k\log n. \tag{6.3}$$

Plugging (6.2) and (6.3) into (6.1), leads to

$$\mathbb{E}[R_n] \leq \frac{\gamma}{1-\gamma}n + \frac{\eta}{2}\mathbb{E}\left[\sum_{t=1}^{n}\sum_{a\in\mathcal{A}} p_t(a)(a^T\hat{z}_t)^2\right] + \frac{2k\log n}{\eta}. \tag{6.4}$$

It remains to control the quadratic term $\mathbb{E}\left[\sum_{t=1}^{n}\sum_{a\in\mathcal{A}} p_t(a)(a^T\hat{z}_t)^2\right]$. We use the fact that $|z_t^{(j)}|, |a_t^{(j)}| \leq 1$ to obtain

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{n}\sum_{a\in\mathcal{A}} p_t(a)(a^T\hat{z}_t)^2\right] &= \sum_{t=1}^{n}\sum_{a\in\mathcal{A}} p_t(a)\mathbb{E}_{q_t}[(a^T\hat{z}_t)^2] \\
&= \sum_{t=1}^{n}\sum_{a\in\mathcal{A}} p_t(a)\left((1-\gamma)0 + \sum_{j=1}^{k}\frac{\gamma}{k}\left(\frac{k}{\gamma}\right)^2[a^j z_t^{(j)}]^2\right) \\
&\overset{|a^j z_t^{(j)}|\leq 1}{\leq} \sum_{t=1}^{n}\sum_{a\in\mathcal{A}} p_t(a)\left(\frac{k^2}{\gamma}\right) = n\frac{k^2}{\gamma}.
\end{aligned}$$

Plugging this bound into (6.4), we have

$$\mathbb{E}[R_n] \leq \gamma n + \frac{\eta}{2}n\frac{k^2}{\gamma} + \frac{2k\log n}{\eta}.$$

Letting $\gamma = \frac{1}{n^{1/3}}$ and $\eta = \sqrt{\frac{\log n}{kn^{4/3}}}$ leads to

$$\mathbb{E}[R_n] \leq Ck^{3/2}n^{2/3}\sqrt{\log n}.$$

□

Literature: The bound we just proved has been improved in [VKH07] where they show $O(d^{3/2}\sqrt{n\log n})$ bound with a better exploration in the algorithm. The exploration that we used in the algorithm was coordinate-wise. The key is that we have a linear problem and we can use better tools from linear regression such as least square estimation. However, we will describe a slightly different approach in which we never explore and the exploration is completely done with the exponential weighting. This approach also gives a better performance in terms of the dependency on k. In particular, we obtain the bound $O(d\sqrt{n\log n})$ which coincides with the bound recently shown in [BCK 12] using a John's ellipsoid.

Theorem: Let $C_t = \mathbb{E}_{a_t \sim q_t}[a_t a_t^T]$, $\hat{z}_t = (a_t^T z_t) C_t^{-1} a_t$, and $\gamma = 0$ (so that $p_t = q_t$). Using Geometric Hedge algorithm with $\eta = 2\sqrt{\frac{\log n}{n}}$ for linear bandit with bandit feedback leads to

$$\mathbb{E}[R_n] \leq CK\sqrt{n \log n}.$$

Proof. We follow the same lines of the proof as the previous theorem to obtain (6.4). Note that the only fact that we used in order to obtain (6.4) is unbiasedness, i.e., $\mathbb{E}[\hat{z}_t] = z_t$, which holds here as well since

$$\mathbb{E}[\hat{z}_t] = \mathbb{E}[C_t^{-1} a_t a_t^T z_t] = C_t^{-1} \mathbb{E}[a_t a_t^T] z_t = z_t.$$

Note that we can use pseudo-inverse instead of inverse so that invertibility is not an issue. Therefore, rewriting (6.4) with $\gamma = 0$, we obtain

$$\mathbb{E}[R_n] \leq \frac{\eta}{2} \mathbb{E}_{a_t \sim p_t} \left[\sum_{t=1}^n \sum_{a \in \mathcal{A}} p_t(a) (a^T \hat{z}_t)^2 \right] + \frac{2k \log n}{\eta}.$$

We now bound the quadratic term as follows

$$\begin{aligned}
&\mathbb{E}_{a_t \sim p_t} \left[\sum_{t=1}^n \sum_{a \in \mathcal{A}} p_t(a) (a^T \hat{z}_t)^2 \right] = \sum_{t=1}^n \sum_{a \in \mathcal{A}} p_t(a) \mathbb{E}_{a_t \sim p_t} \left[(a^T \hat{z}_t)^2 \right] \\
&\overset{C_t^T = C_t,\ \hat{z}_t = (a_t^T z_t) C_t^{-1} a_t}{=} \sum_{t=1}^n \sum_{a \in \mathcal{A}} p_t(a) a^T \mathbb{E}\left[\hat{z}_t \hat{z}_t^T \right] a = \sum_{t=1}^n \sum_{a \in \mathcal{A}} p_t(a) a^T \mathbb{E}\left[(a_t^T z_t)^2 C_t^{-1} a_t a_t^T C_t^{-1} \right] a \\
&\overset{|a_t^T z_t| \leq 1}{\leq} \sum_{t=1}^n \sum_{a \in \mathcal{A}} p_t(a) a^T C_t^{-1} \mathbb{E}\left[a_t a_t^T \right] C_t^{-1} a \overset{\mathbb{E}[a_t a_t^T] = C_t}{=} \sum_{t=1}^n \sum_{a \in \mathcal{A}} p_t(a) a^T C_t^{-1} a \\
&= \sum_{t=1}^n \sum_{a \in \mathcal{A}} p_t(a) \mathrm{tr}(a^T C_t^{-1} a) \overset{\mathrm{tr}(AB) = \mathrm{tr}(BA)}{=} \sum_{t=1}^n \sum_{a \in \mathcal{A}} p_t(a) \mathrm{tr}(C_t^{-1} a a^T) \\
&= \sum_{t=1}^n \mathrm{tr}(C_t^{-1} \mathbb{E}_{a \sim p_t}[a a^T]) = \sum_{t=1}^n \mathrm{tr}(C_t^{-1} C_t) = \sum_{t=1}^n \mathrm{tr}(I_k) = kn.
\end{aligned}$$

Plugging this bound into previous bound yields

$$\mathbb{E}[R_n] \leq \frac{\eta}{2} nk + \frac{2k \log n}{\eta}.$$

Letting $\eta = 2\sqrt{\frac{\log n}{n}}$, leads to $\mathbb{E}[R_n] \leq Ck\sqrt{n \log n}$. ☐

References

[BCK 12] Bubeck, Sbastien, Nicolo Cesa-Bianchi, and Sham M. Kakade. *Towards minimax policies for online linear optimization with bandit feedback.* arXiv preprint arXiv:1202.3079 (2012). APA

[BB04] McMahan, H. Brendan, and Avrim Blum. *Online geometric optimization in the bandit setting against an adaptive adversary.* Conference on Learning theory (COLT) 2004.

[VH06] Dani, Varsha, and Thomas P. Hayes. *Robbing the bandit: Less regret in online geometric optimization against an adaptive adversary.* Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm. Society for Industrial and Applied Mathematics, 2006.

[BK04] Awerbuch, Baruch, and Robert D. Kleinberg. *Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches.*Proceedings of the thirty-sixth annual ACM symposium on Theory of computing. ACM, 2004.

[VKH07] Dani, Varsha, Sham M. Kakade, and Thomas P. Hayes, *The price of bandit information for online optimization*, Advances in Neural Information Processing Systems. 2007.