IE 498: Online Learning and Decision Making, Fall 2019 10/8/19

Lecture 13: Linear Contextual Bandits

Lecturer: Yuan Zhou *Scribe: Ebrahim Arian, Manuel Torres*

1 Recap

We first recall the setup for linear contextual bandits. There is an unknown vector $\vec{\theta} \in \mathbb{R}^d$, normalized such that $\left\|\vec{\theta}\right\|_2 \leq 1$. There are n actions, or arms, and we assume the time horizon T is known. At time t,

1. For all arms $i \in [n]$, player observes context $\vec{x}_{t,i} \in \mathbb{R}^d$
2. The player decides on arm i_t
3. The player receives reward $r_t = \vec{x}_{t,i_t}^\top \vec{\theta} + \epsilon_t$ where $\epsilon_t \sim \mathcal{N}(0,1)$

Note that we normalize the contexts such that $\|\vec{x}_{t,i}\|_2 \leq 1$ for all $t \in [T]$ and $i \in [n]$. We can now define the regret $\mathcal{R}(T)$ in this setting. We have

$$\mathcal{R}(T) := \mathbb{E}\left[\sum_{t=1}^{T} \max_{j\in[n]} \vec{x}_{t,j}^\top \vec{\theta} - r_t\right].$$

As a matter of notational convenience, we let $\vec{y}_t = \vec{x}_{t,i_t}$.

In the last lecture, we discussed two key lemmas regarding linear contextual bandits. We define $v_t = \sum_{z=1}^{t} \vec{y}_z \vec{y}_z^\top$, and $\hat{\vec{\theta}} = (I + v_t)^{-1} \sum_{z=1}^{t} r_z \vec{y}_z$. Based on these definitions, we can state the two key lemmas.

Lemma 1. *There exists $c > 0$, such that for all $\delta \in (0, \frac{1}{2})$, and $\{\epsilon_t\}$ is independent from $\{\vec{y}_t\}$, then for all $x \in \mathbb{R}^d$ we have*

$$\Pr\left[\left|(\hat{\vec{\theta}} - \vec{\theta})^\top \vec{x}\right| \leq c \left\|(I + v_t)^{-\frac{1}{2}} \vec{x}\right\|_2 \sqrt{\ln(1/\delta)}\right] \geq 1 - \delta.$$

Lemma 2 (Elliptical potential lemma)**.** *Let u_t be $I + v_t$. Then*

$$\sum_{t=1}^{T} \left\|u_{t-1}^{-\frac{1}{2}} \vec{y}_t\right\|_2 \leq \sqrt{2Td \ln\left(\frac{T}{d} + 1\right)}.$$

Lemma 1 shows that the probability the error will be bounded by the given confidence interval is more than $1 - \delta$. In this lemma, it is crucial that the noise $\{\epsilon_t\}$ and action $\{y_t\}$ are independent.

2 SupLinUCB Algorithm

For each time t, we construct a partition of size $\log T$ of all the past times $\{1, 2, .., t-1\}$ and denote this partition as $\{\Psi_t^s\}_{s=1}^{\log T}$ where each Ψ_t^s is a subset of $[t-1]$.[1] Now, let $\hat{\vec{\theta}}_t^s$ be the estimator for a set of a

[1] We assume $\log T$ is an integer, but one can easily verify that all results still hold when you consider $\lceil \log T \rceil$. We also assume the base of the log is 2.

Algorithm 1 SupLinUCB

1. $s \leftarrow 1$, $A_1 \leftarrow [n]$.
2. Repeat
   IF $2^{-s} \leq \frac{1}{\sqrt{T}}$ THEN $i_t = \text{argmax}_{i \in A_s}\{\hat{r}^{s-1}_{t,i}\}$ (case 1)
   ELSIF $\exists i \in A_s$ where $\omega^s_{t,i} > 2^{-s}$ THEN choose any $i_t \in A_s$ s.t. $\omega^s_{t,i_t} > 2^{-s}$ (case 2)
   ELSE set $A_{s+1} \leftarrow \{i \in A_s : \hat{r}^s_{t,i} + \omega^s_{t,i} \geq \max_{j \in A_s}\{\hat{r}^s_{t,j} + \omega^s_{t,j} - 2^{1-s}\}$ and $s \leftarrow s+1$
   UNTIL i_t found
3. (update partition) $\Psi^{s'}_{t+1} = \begin{cases} \Psi^s_t \cup \{t\}, & \text{if } s = s' \\ \Psi^s_t & \text{otherwise.} \end{cases}$

Figure 1: Pseudocode for the SupLinUCB algorithm initially proposed by Chu et al. [CLRS11]. Note that we only include the pseudocode for time step t.

particular partition where

$$\hat{\vec{\theta}}^s_t = (u^s_t)^{-1} \sum_{z \in \Psi^s_t} r_z \vec{y}_z \quad \text{where} \quad u^s_t = I + \sum_{z \in \Psi^s_t} \vec{y}_z \vec{y}_z^\top.$$

In addition, for a specific partition and time, we define a confidence interval parameter $\omega^s_{t,i}$ and estimated mean reward $\hat{r}^s_{t,i}$ where

$$\omega^s_{t,i} = c \left\| (u^s_t)^{-\frac{1}{2}} \vec{x}_{t,i} \right\|_2 \sqrt{\ln(T^2 n)} \quad \text{and} \quad \hat{r}^s_{t,i} = \vec{x}^\top_{t,i} \hat{\vec{\theta}}^s_t$$

We now can state the SupLinUCB algorithm, which was initially proposed by Chu et al. [CLRS11]. The pseudocode is given in Figure 1. Note that the algorithm is only stated for time step t for ease of notation.

Step 1 of the algorithm sets A_1 to be the set of all possible actions $[n]$. In the second step, there are two cases in which the algorithm will select an action. If the conditions of these two cases are not satisfied, then the algorithm deletes some non-optimal actions and repeats the process. This process is repeated until an action is found in case 1 or case 2. (We are guaranteed to find an action because s ranges from 1 to $\log T$, so eventually the condition for case 1 is satisfied.) Finally, in step 3, the algorithm updates the partition for the next time period. In order to be able to apply Lemma 1, we need to make sure that by using SupLinUCB, the noises and actions are independent. We first observe that the action chosen at time t does not depend on any of the sets in the partition with a smaller index.

Observation 1. *For every $t \in [T]$, let s_t be the s at step 3. Then i_t only depends on $\epsilon_{t'}$ where $t' \in \Psi^1_t \cup \cdots \cup \Psi^{s_t - 1}_t$.*

In order to confirm the observation, we first observe that i_t is only decided if we end up in case 1 or case 2. If i_t is chosen in case 1, the arm is chosen based on the rewards from round $s-1$. If i_t is chosen in case 2, we notice that the confidence interval $\omega^s_{t,i}$ just depends on actions not the rewards, and the choice of i_t only depends $\omega^s_{t,i}$. This confirms the above observation.

Fix $t \in [T]$. With the above observation, we have that in each "layer" Ψ^s_t, all of the noises and the chosen actions are independent. This is stated in the following lemma.

Lemma 3. *For all $t \in [T]$ and all $s \in [\log T]$, $\{\epsilon_\tau\}_{\tau \in \Psi^s_t}$ is independent from $\{y_t\}_{\tau \in \Psi^s_t}$.*

As a matter of notation, let s_t denote the value of s at the end of step 3 in the algorithm for step t and let $i_t^* = \text{argmax}_{j\in[n]} \vec{x}_{t,j}^\top \vec{\theta}$. Furthermore, let $A_{s,t}$ be the set A_s of the algorithm at the t^{th} step. We define the following events for convenience.

- E is the event that for all $t \in [T]$, $s \in [\log T]$, and $i \in [n]$, $\left|\vec{x}_{t,i}^\top \vec{\theta} - \hat{r}_{t,i}^s\right| \le \omega_{t,i}^s$.
- F_1 is the event that for all $t \in [T]$ and $s \le s_t$, for all $i \in A_{s,t}$, $\left|\vec{x}_{t,i}^\top \vec{\theta} - \max_{j\in A_{s,t}} \vec{x}_{t,j}^\top \vec{\theta}\right| \le 2^{2-s}$.
- F_2 is the event that for all $t \in [T]$ and for all $s \le s_t$, $i_t^* \in A_{s,t}$.

We can show that the event E holds with high probability easily via a union bound and the error bound in Lemma 1. Note that we can only apply Lemma 1 as Lemma 3 implies that actions and noises are independent across layers. Therefore, for the rest of this section, we will simply assume that E occurs, as we know it happens with high probability.

We can show that if E holds, then F_1 and F_2 holds. We can prove this in a straightforward manner via induction. The following lemma gives an upper bound for the regret the algorithm incurs at time t. Recall that s_t is the value of s at the end of step 3 of the algorithm at time step t. We use the notation $A \lesssim B$ if $A \le \alpha B$ for some constant $\alpha > 0$.

Lemma 4. *Given events E, F_1, and F_2 all occur simultaneously, for all $t \in [T]$,*

$$\left|\vec{x}_{t,i_t}^\top \vec{\theta} - \vec{x}_{t,i_t^*}^\top \vec{\theta}\right| \lesssim \max\left\{\frac{1}{T}, \omega_{t,i_t}^{s_t}\right\}$$

Proof. If i_t is chosen after entering case 1 of the algorithm, we have

$$\left|\theta^T x_{t,i_t} - \theta^T x_{t,i_t^*}\right| \le 2^{2-(s_t-1)} \le \frac{8}{\sqrt{T}}.$$

If i_t is chosen after entering case 2 of the algorithm, as event F_1 occurred, we have

$$\left|\theta^T x_{t,i_t} - \theta^T x_{t,i_t^*}\right| \le 2^{2-s_t} \le 4 \cdot \omega_{t,i_t}^{s_t}.$$

This concludes the proof. □

Now that we have bounded the regret for each time step, we can bound the regret for the entire algorithm.

Theorem 5. *We have*

$$\mathcal{R}_t \lesssim \sqrt{dT \log^2 T \log(nT)}.$$

Proof. Applying Lemma 4 for each $t \in [T]$, we have

$$\mathcal{R}_t \lesssim \sqrt{T} + \sum_{t=1}^{T} \omega_{t,i_t}^{s_t}.$$

Then by Lemma 2,

$$\sum_{t=1}^{T} \omega_{t,i_t}^{s_t} = \sum_{s=1}^{\log_2 T} \sum_{t\in\Psi_T^s} \omega_{t,i_t}^{s_t} \lesssim \sum_{s=1}^{\log_2 T} \sqrt{\log(nT)} \cdot \sqrt{dT \log T} \le \sqrt{dT \log^2 T \log(nT)}.$$

□

3 Infinitely-many arms

In the previous section, we considered the case where there are only finitely-many arms. In this section, we address the following question: can we get finite regret bounds for the setting with infinitely-many arms? It is not obvious a priori that one can obtain finite regret in this setting. However, this is indeed possible via a simple discretization step via an ϵ-net.

Let $D_t \subseteq \{\vec{x} \in \mathbb{R}^d : \|x\|_2 \leq 1\}$ be the decision area at time t. Our goal is to construct a small, finite set $\tilde{D}_t \subseteq D_t$ for a given ϵ such that for all $\vec{u} \in D_t$, there exists $\vec{u}' \in \tilde{D}_t$ such that $\|\vec{u} - \vec{u}'\|_2 \leq \epsilon$. If we set $\epsilon = \frac{1}{T}$, this would imply there exists $\vec{u}' \in \tilde{D}_t$ such that $\left|\vec{\theta}^\top(\vec{u}^* - \vec{u}')\right| \leq \left\|\vec{\theta}\right\|_2 \cdot \|\vec{u}^* - \vec{u}'\|_2 \leq \frac{1}{T}$.

Lemma 6. *Let $\epsilon > 0$ and let $D_t \subseteq \{\vec{x} \in \mathbb{R}^d : \|x\|_2 \leq 1\}$. There exists a set $\tilde{D}_t \subseteq D_t$ such that $\left|\tilde{D}_t\right| \lesssim (\frac{2}{\epsilon}+1)^d$.*

Proof. We proceed via a greedy algorithm. Start with $\tilde{D}_t$ as the empty set. We repeatedly add $\vec{v} \in D_t$ to $\tilde{D}_t$ as long as for all $\vec{u} \in \tilde{D}_t$, we have $\|\vec{u} - \vec{v}\|_2 > \epsilon$. Once this algorithm terminates, for every $\vec{u}, \vec{v} \in \tilde{D}_t$ with $\vec{u} \neq \vec{v}$, we have $\|\vec{u} - \vec{v}\|_2 > \epsilon$.

Now consider placing balls of radius $\frac{\epsilon}{2}$ centered at each point in $\tilde{D}_t$. As each point is at least ϵ-distance apart, we have that these balls are not overlapping. However, we can cover the entire space with balls of radius $1 + \frac{\epsilon}{2}$. Therefore, letting $B^d(r)$ be a ball in d-dimensional space of radius r, we have

$$\left|\tilde{D}_t\right| \leq \frac{\text{vol}(B^d(1+\frac{\epsilon}{2}))}{\text{vol}(B^d(\frac{\epsilon}{2}))} = \left(\frac{1+\frac{\epsilon}{2}}{\epsilon/2}\right)^d = \left(\frac{2}{\epsilon}+1\right)^d.$$

□

Therefore, because we can construct such a set $\tilde{D}_t$ for $\epsilon = 1/T$, we have that the regret in this setting is at most

$$\mathcal{R}_T \lesssim \sqrt{dT \log^2 T \log\left(\left(\frac{2}{\epsilon}+1\right)^d T\right)} \lesssim d\sqrt{T \log^3 T}.$$

Remark 1. *We are able to simply "replace" the occurrence of n in the regret bound of Theorem 5 with the upper bound on $\left|\tilde{D}_t\right|$ in Lemma 6. Therefore, in the finite case, the worst case is when the number of arms is exponential in d. If the number of arms exceeds this value, we can simply use the method described in this section.*

References

[CLRS11] Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214, 2011.