IE 498: Online Learning and Decision Making, Fall 2019 Sept. 24th

Lecture 09: Parameter-dependent Lower Bound and Adversarial Bandit

Lecturer: Yuan Zhou *Scribe: Seiyun Shin, Summer Xia*

1 Instance-dependent Lower Bound for Multi-armed Bandit

In the last lecture, we proved the minimax lower bound for n-armed bandit with T horizon: $R_T \gtrsim \sqrt{nT}$. Also, a similar approach can prove that we need the sample complexity of at least $n\epsilon^{-2}\ln(\delta^{-1})$. Here one can see that these two results do not depend on the instances (recall the meaning of minimax).

Today we will prove an instance-dependent lower bound. Before going through it, let us first take a look at the following simple case of the Bernoulli instance:

- Case of the Bernoulli instance: Consider an instance where $I = (\mu_1, \mu_2, \ldots, \mu_n) = (\mathcal{B}_{\mu_1}, \mathcal{B}_{\mu_2}, \ldots, \mathcal{B}_{\mu_n})$. Specifically, let us define $\mathcal{I} := \{(\mu_1, \mu_2, \ldots, \mu_n) : \mu_i \in [0.05, 0.95],\ \forall i\}$. In fact, the cases of interest are to keep μ_i's far away from the bounds 0 and 1. The reason why we only consider this case will be clearer with the following proof.

As in the previous lectures, one can readily expect a key parameter that characterizes the hardness of the bandit problem. The parameter is called "complexity parameter" and it turns out to be defined as:

$$c^*(I) := \sum_{i:\Delta_i>0} \frac{1}{\Delta_i}.$$

One can see that as Δ_i ($:= \mu^* - \mu_i$) gets smaller, it is harder to differentiate between the optimal arm and arm i. Hence, it is natural that the hardness is reciprocal with Δ_i for all i.

With this parameter, our goal is to derive the regret lower bound to be $\Omega(c^*(I)\log T)\ \forall \pi, \forall I \in \mathcal{I}$. However, this is indeed not possible for all policies. Here is a counter example:

- Counter example: Consider a trivial regret where a player always plays the second best arm. Then we get a linear regret.

This motivates us to consider a reasonable policy as follows:

Definition 1 (Reasonable Policy)**.** *We define π to be (C,p)-reasonable if $R_T^{\pi,I} \le C \cdot T^p\ \forall I, T$, where $p < 1$.*

As an example, consider a UCB policy where $R_T \le \sqrt{nT\log T}$. One can see that the UCB is $(\mathcal{O}(\sqrt{n}), 0.51)$-reasonable. Now we are ready to derive the following theorem about an instance-dependent lower bound.

Theorem 2. *$\forall (C,p)$-reasonable π and $\forall I \in \mathcal{I}$,*

$$R_T^{\pi,I} \gtrsim \sum_{i:\Delta_i>0} \frac{1}{\Delta_i}\left[(1-p)\ln T + \ln\frac{\min\{\Delta_i, 0.05\}}{8C}\right]^+,$$

where $(x)^+ := \max(0, x)$.

Proof. With Regret Decomposition Lemma, we get:

$$\begin{aligned} R_T^{\pi,I} &= \sum_i \Delta_i \mathbb{E}[T_i] \\ &\overset{(a)}{\gtrsim} \sum_{i:\Delta_i>0} \frac{1}{\Delta_i} \left[(1-p)\ln T + \ln \frac{\min\{\Delta_i, 0.05\}}{8C} \right]^+, \end{aligned}$$

where (a) follows from the following lemma. □

Lemma 3. *$\forall I = (\mu_1, \mu_2, \ldots, \mu_n)$ and $\forall i$ with $\Delta_i > 0$,*

$$\mathbb{E}[T_i] \gtrsim \frac{1}{\Delta_i^2} \left[(1-p)\ln T + \ln \frac{\min\{\Delta_i, 0.05\}}{8C} \right]^+.$$

Proof. The idea is to use Divergence Decomposition Lemma (see Lemma 1, Lecture 08), which reveals the relationship between $\mathbb{E}[T_i]$ and KL-divergence. To deal with KL-divergence, we consider another instance different from the previous I, where $I' = (\mu_1, \mu_2, \ldots, \mu_{i-1}, \mu_i + \lambda, \mu_{i+1}, \ldots, \mu_n)$. Here we define $\lambda = \Delta_i + \epsilon$, where $\epsilon := \min\{\Delta_i, 0.05\}$. Then one can see that the i^{th} arm becomes the best arm in the case of I', unlike the case of I where μ_1 is the best.

Now let $A := \{I \in \mathcal{I} : T_i > \frac{T}{2}\}$. Note that this is the desired event for I' but not for I. As a result, we get:

$$\begin{aligned} R_T^{\pi,I} &\geq P_{\pi,I}(A) \cdot \frac{T\Delta_i}{2}; \\ R_T^{\pi,I'} &\geq P_{\pi,I'}(\bar{A}) \cdot \frac{T\epsilon}{2}. \end{aligned}$$

The last inequality holds because in the complement of event A, the difference between the best arm and any other arm is at least ϵ ($:= \min\{\Delta_i, 0.05\}$). Hence,

$$\begin{aligned} R_T^{\pi,I} + R_T^{\pi,I'} &\geq \frac{T\epsilon}{2} \left(P_{\pi,I}(A) + P_{\pi,I'}(\bar{A}) \right) \\ &\overset{(a)}{\geq} \frac{T\epsilon}{2} \left(1 - \left| P_{\pi,I}(A) - P_{\pi,I'}(A) \right| \right) \\ &= \frac{T\epsilon}{2} \left(1 - \Delta(\mathcal{P}_{\pi,I}, \mathcal{P}_{\pi,I'}) \right) \\ &\overset{(b)}{\geq} \frac{T\epsilon}{4} \exp\left(-D_{KL}(\mathcal{P}_{\pi,I} || \mathcal{P}_{\pi,I'}) \right) \\ &\overset{(c)}{\geq} \frac{T\epsilon}{4} \exp\left(-\mathbb{E}_{\pi,I}[T_i] D_{KL}(\mathcal{B}_{\mu_i} || \mathcal{B}_{\mu_i+\lambda}) \right), \end{aligned}$$

where (a) follows from the fact that $P_{\pi,I}(A) + P_{\pi,I'}(\bar{A}) = P_{\pi,I}(A) + (1 - P_{\pi,I'}(A)) \geq 1 - |P_{\pi,I}(A) - P_{\pi,I'}(A)|$; (b) follows from Pinsker's inequality (see Lemma 8 in Lecture 07); and (c) follows from Divergence Decomposition Lemma (see Lemma 1 in Lecture 08) and the fact the difference of the two probability distribution sets occurs only on the i^{th} arm.

The key observation here is that $D_{KL}(\mathcal{B}_{\mu_i} || \mathcal{B}_{\mu_i+\lambda}) \leq \mathcal{O}(\lambda^2)$ if $0 < \mu_i,\ \mu_i + \lambda < 1$. This explains why we assumed that $\mu_i,\ \mu_i + \lambda \in [0.05, 0.95]$ earlier. With this observation, we can further proceed as follows:

$$\begin{aligned} \mathbb{E}_{\pi,I}[T_i] &\geq \frac{\ln \frac{T\epsilon}{4(R_T^{\pi,I} + R_T^{\pi,I'})}}{D_{KL}(\mathcal{B}_{\mu_i} || \mathcal{B}_{\mu_i+\lambda})} \\ &\gtrsim \frac{1}{\lambda^2} \ln \frac{T\epsilon}{4(R_T^{\pi,I} + R_T^{\pi,I'})} \end{aligned}$$

$$\begin{aligned}
&\overset{(d)}{\geq} \frac{1}{\lambda^2} \ln \frac{T\epsilon}{8C \cdot T^p} \\
&= \frac{1}{\lambda^2} \left[(1-p) \ln T + \ln \frac{\epsilon}{8C} \right] \\
&\overset{(e)}{\gtrsim} \frac{1}{\Delta_i^2} \left[(1-p) \ln T + \ln \frac{\epsilon}{8C} \right]
\end{aligned}$$

where (d) follows from the assumption that π is (C, p)-reasonable, $R_T^{\pi,I} \leq C \cdot T^p$; and (e) follows since we defined $\lambda := \Delta_i + \epsilon$ and $\epsilon \leq \Delta_i$. Since $\mathbb{E}[T_i] \geq 0$ is just the number of pulls of i^{th} arm, we have the desired result:

$$\mathbb{E}[T_i] \gtrsim \frac{1}{\Delta_i^2} \left[(1-p) \ln T + \ln \frac{\min\{\Delta_i, 0.05\}}{8C} \right]^+,$$

This completes the proof of Lemma 3. □

We omitted the details about a lower bound of sample complexity since a similar but tedious approach yields the desired result. This completes the lower bound part of Multi-armed Bandit (MAB).

2 Adversarial Bandit

So far, we have proved the performances of upper and lower bounds for Stochastic Multi-armed Bandit problems. Next we consider a different scenario called "Adversarial Bandit". Here is a setup:

- Let there be n actions and a horizon T.
- For $t = 1, 2, \ldots, T$,
  1. Adversary picks a reward vector $\overrightarrow{r_t} := (r_{t,1}, r_{t,2}, \ldots, r_{t,n}) \in [0,1]^n$.
  2. Player plays an action a_t.
  3. Player receives a reward $r_t = r_{t,a_t}$ (this is called a bandit feedback) and unless mentioned otherwise, player observes the whole $\overrightarrow{r_t}$ (this is called a full information setting).

The regret is defined as follows:

$$R_T := \mathbb{E}\left[\max_{a \in \mathcal{A}} \left\{ \sum_{t=1}^{T} r_{t,a} \right\} \right] - \mathbb{E}\left[\sum_{t=1}^{T} r_t \right].$$

Note that the algorithm could be randomized. By inspection, one can readily observe the following:

- Observation 1: Player has to be randomized to achieve a sublinear regret; otherwise the adversary can predict the player's actions (that are deterministic) and put 0 rewards for those actions.
- Observation 2: By definition, we see that Stochastic Bandit problems are easier than Adversary Bandit problems; hence, the lower bound of Stochastic Bandit implies that of Adversary Bandit, and thus $\forall \pi, \exists$ adversary such that $R_T^\pi \gtrsim \sqrt{nT}$.

2.1 Multiplicative Weights

One technique that enables us to get a regret upper bound for Adversarial Bandit problem is called "Multiplicative Weights." Before going into further details, we will consider the following warm-up problem:

- In an "Expert problem", let there be n experts and T horizons.
- At each time t,
  1. Each expert gives Yes or No (Y/N) advice.
  2. Player sees all advice and decides Y/N.
  3. Player sees outcome and suffers if the decision is wrong.

The expert problem leads to the following questions:

- Question 1: If there exists an expert that always gives the correct advice, what is the strategy to minimize the number of sufferings?
- Question 2: If there exists an expert that makes at most M mistakes, what is the strategy to minimize the number of sufferings?

For the first question, a naive strategy would be to follow the majority votes and fire the experts with minority votes at each time. In this case, the number of sufferings is at most $\lceil \log_2 n \rceil$.

For the second question, a simple answer to the minimum number of sufferings would be $\mathcal{O}((M+1)\log_2 n)$, which can be adapted from the answer to the first question. However, we can do even better with another strategy, Soft Penalty Algorithm. The idea is to let the player put weights on experts' votes, decide on the answer according to the weighted votes, and update the weights based on the correctness of the experts' last votes.

Algorithm 1 Soft Penalty Algorithm

1. $w_i^{(1)} = 1,\ \forall i \in [n]$;
for time $t \in \mathcal{T}$ $(:= \{1, 2, \ldots, T\})$ **do**
   2. Decide based on weighted majority votes w.r.t. $\{w_i^{(t)}\}_{i\in[n]}$, (i.e., the metric would be $\sum_i w_i^{(t)}$).
   3. Update weights as follows:

$$w_i^{(t+1)} = \begin{cases} w_i^{(t)}, & \text{if expert } i \text{ was correct;} \\ \frac{w_i^{(t)}}{2}, & \text{otherwise.} \end{cases}$$

end for

Analysis: As the algorithm makes a decision based on weighted majority votes, we first define the following metric which is called a potential function:

$$W^{(t)} := \sum_{i=1}^{n} w_i^{(t)}.$$

Note that $W^{(1)} = n$. We claim that if the player makes an incorrect decision, then $W^{(t+1)} \leq \frac{3}{4} W^{(t)}$. The claim is explained by the fact that the player decides based on the majority votes so the player's incorrect decision implies that at least half of the experts gave wrong advice, so at least half of the weights are to be halved.

Let R be the number of wrong decisions by the player. By the previous argument and induction on R, we get:

$$W^{(T+1)} \leq \left(\frac{3}{4}\right)^R W^{(1)} = \left(\frac{3}{4}\right)^R n.$$

If there exists an expert i which makes at most M mistakes, then $w_i^{(T+1)} \geq \left(\frac{1}{2}\right)^M$. Hence, we trivially get: $W^{(T+1)} \geq \left(\frac{1}{2}\right)^M$.

Combining the two inequalities that we derived above, one can readily see that

$$M \ln 2 \geq R \ln \frac{4}{3} - \ln n,$$

which in turn yields:

$$R \leq \log_{\frac{4}{3}} n + M \log_{\frac{4}{3}} 2.$$

Note that this is better than $\mathcal{O}((M+1)\log_2 n)$ that we earned naively before.