EE 290 Theory of Multi-armed Bandits and Reinforcement Learning Lecture 6 - 02/04/2021

Lecture 6: Minimax Lower Bound and Thompson Sampling

Lecturer: Jiantao Jiao *Scribe: Mariel Werner, Fangchen Liu*

In this lecture, we complete last lecture's objective of proving a minimax lower bound for expected regret in the finite-arm bandit setting, and we introduce Thompson sampling.

1 Minimax Lower Bound for Finite-Arm Bandits cont.

Recall the minimax lower bound for expected regret in finite-arm bandit algorithms stated at the end of last lecture:

Theorem 1. *For $T \geq K - 1$, π a policy, and ν from a family of Gaussian bandit instances*

$$\inf_{\pi} \sup_{\nu} \mathbb{E}_{\nu}[R(T)] \gtrsim \sqrt{KT} \tag{1}$$

Before proving this theorem, we remind ourselves of its setting. In Theorem 1, T is the number of rounds in the bandit problem, K the number of arms, and $\nu = \{P_a \sim \mathcal{N}(\mu_a, 1) : a \in \mathcal{A},\ \mu_a \in [0,1]\}$ a bandit instance, consisting of Gaussian reward distributions with [0,1]-bounded mean and variance 1, over which we are maximizing expected regret.

The following information-theoretic results are useful for the proof of Theorem 1.

Definition 2 (Total Variation (TV) Distance)**.** *Let P and Q be two probability measures defined on (Ω, F), with density functions p and q respectively. The total variation distance between P and Q is defined*

$$TV(P,Q) = \sup_{A \in F}(P(A) - Q(A)) = \sup_{A \in F} \int_{x \in A} (p(x) - q(x))dx \tag{2}$$

Clearly, for any P and Q, $TV(P,Q) \in [0,1]$.

Next, we introduce two useful upper-bounds for TV-distance based on the KL-divergence. The first guarantees that whenever KL-divergence is finite, so is TV-distance.

Lemma 3 (Pinsker's Inequality)**.** *Let P and Q be two probability measures. Then*

$$TV(P,Q) \leq \sqrt{\frac{1}{2} D_{KL}(P||Q)} \tag{3}$$

where $D_{KL}(P||Q)$ is the KL-divergence between P and Q.

As TV-distance is always bounded by 1, Pinsker's inequality becomes meaningless for very large D_{KL}. The following inequality on the other hand provides a non-trivial upper bound for TV-distance when D_{KL} is large.

Lemma 4. *Let P and Q be two probability measures. Then*

$$1 - TV(P,Q) \geq \frac{1}{2} e^{-D_{KL}(P||Q)} \tag{4}$$

In particular, even for extremely large finite $D_{KL}(P||Q)$, this inequality guarantees that $TV(P,Q) < 1$, i.e. there is still uncertainty about P and Q (when $TV(P,Q) = 1$, P and Q have disjoint support and thus are easily distinguishable with no samples).

We now have the tools to prove Theorem 1.

Proof Fix policy π and let $\Delta \in [0, \frac{1}{2}]$ be a constant which we will set later. Let

$$\nu = (p_1, ..., p_K) : p_{j\in[K]} \sim \mathcal{N}(\mu_{j\in[K]}, 1), \mu = (\Delta, 0, ..., 0) \tag{5}$$

$$\nu' = (p_1, ..., p_K) : p_{j\in[K]} \sim \mathcal{N}(\mu'_{j\in[K]}, 1), \mu' = (\Delta, 0, .., 0, 2\Delta, 0, ..., 0) \tag{6}$$

be two instances of Gaussian reward distributions, with 2Δ the mean reward of the ith arm under instance ν'. Let P_ν and P'_ν be the joint distribution of actions and rewards under π and ν and ν' respectively. Finally, assume that $i = \arg\min_{j>1} \mathbb{E}_\nu n_T(j)$. As shown last time, in this setting the following inequalities hold:

$$\mathcal{R}_\nu \triangleq \mathbb{E}_\nu[R(T)] \geq P_\nu \left(n_T(1) \leq \frac{T}{2} \right) \frac{T\Delta}{2} \tag{7}$$

$$\mathcal{R}_{\nu'} \triangleq \mathbb{E}_{\nu'}[R(T)] > P_{\nu'} \left(n_T(1) > \frac{T}{2} \right) \frac{T\Delta}{2} \tag{8}$$

Define event $A = \{n_T(1) \leq \frac{T}{2}\}$. Then

$$\begin{aligned}
\mathcal{R}_\nu + \mathcal{R}_{\nu'} &\geq \frac{T\Delta}{2} \left[P_\nu \left(n_T(1) \leq \frac{T}{2} \right) + P_{\nu'} \left(n_T(1) > \frac{T}{2} \right) \right] && (9)\\
&= \frac{T\Delta}{2} [P_\nu(A) + P_{\nu'}(A^c)] && (10)\\
&= \frac{T\Delta}{2} [P_\nu(A) + 1 - P_{\nu'}(A)] && (11)\\
&\geq \frac{T\Delta}{2} [1 + \inf_B (P_\nu(B) - P_{\nu'}(B))] && (12)\\
&= \frac{T\Delta}{2} [1 - \sup_B (P_{\nu'}(B) - P_\nu(B))] && (13)\\
&= \frac{T\Delta}{2} [1 - TV(P_\nu, P_{\nu'})] && (14)\\
&\geq \frac{T\Delta}{4} e^{-D(P_\nu||P_{\nu'})} && (15)
\end{aligned}$$

where (14) follows from Definition 2 and (15) follows from Lemma 4.

By the Divergence Decomposition Lemma (Lecture 5, Lemma 3),

$$\begin{aligned}
D_{KL}(P_\nu||P_{\nu'}) &= \sum_{j=1}^{k} \mathbb{E}_\nu[n_T(j)] D_{KL}(p_j||p'_j) && (16)\\
&= \mathbb{E}_\nu[n_T(i)] D_{KL}(\mathcal{N}(0,1)||\mathcal{N}(2\Delta, 1)) && (17)\\
&\leq \frac{2T\Delta^2}{K-1} && (18)
\end{aligned}$$

where the second equality follows from the fact that the reward distributions in environments ν and ν' differ only on the ith arm with $p_i \sim \mathcal{N}(0,1)$ and $p'_i \sim \mathcal{N}(2\Delta, 1)$; and the last inequality follows from the fact that, by assumption that $i = \arg\min_{j>1} \mathbb{E}_\nu n_T(j)$, $\mathbb{E}_\nu[n_T(i)] \leq \frac{T}{K-1}$, and also

$$D(\mathcal{N}(0,1)||\mathcal{N}(2\Delta,1)) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \log\left(e^{-\frac{1}{2}[x^2-(x-2\Delta)^2]}\right) dx \tag{19}$$

$$= 2\Delta^2 \tag{20}$$

Combining (15) and (18), we have that

$$\mathcal{R}_\nu + \mathcal{R}_{\nu'} \geq \frac{T\Delta}{4} e^{-\frac{2T\Delta^2}{K-1}} \tag{21}$$

Since $\Delta \in [0, \frac{1}{2}]$ was arbitrarily chosen, we can maximize the lower bound in (21) by setting

$$\Delta = \sqrt{\frac{K-1}{4T}} \tag{22}$$

which is the solution to

$$\frac{d}{d\Delta}\left(\frac{T\Delta}{4} e^{-\frac{2T\Delta^2}{K-1}}\right) = 0 \tag{23}$$

Substituting (22) into (21), we have

$$\max(\mathcal{R}_\nu, \mathcal{R}_{\nu'}) \geq \frac{\mathcal{R}_\nu + \mathcal{R}_{\nu'}}{2} \tag{24}$$

$$\geq \frac{\sqrt{T(K-1)}}{16} e^{(-\frac{1}{2})} \tag{25}$$

$$\gtrsim \sqrt{KT} \tag{26}$$

□

Since π was chosen arbitrarily, the bound in (26) holds for an infimum over all π, thus proving the desired result.

2 Thompson Sampling

In this section, we will talk about Thompson Sampling (TS) in the bounded reward case, and also analyze its worst-case expected regret. We start from the Bernoulli Bandit Thompson Sampling as shown in algorithm 1, and then talk about how to reduce the bounded reward case to Bernoulli bandit case.

Here are some notations used in the algorithm:

- $N_i(t)$: number of pulls of arm i up to time $t-1$
- $S_i(t)$: number of success of arm i up to time $t-1$
- $F_i(t)$: number of failures of arm i up to time $t-1$

For Bernoulli bandits, the result is success or failure (with reward 1 or 0, respectively) for each pull, so we have $F_i(t) + S_i(t) = N_i(t)$.

Note that in the bounded reward case, $r \in [0, 1]$. We can simulate the Bernoulli reward by introducing another r' sampled from $Bern(r)$, and performing TS using $r' \sim Bern(r)$. In this way we can reduce the bounded reward case to Bernoulli reward case.

We use beta distribution because it's the conjugate prior of Bernoulli distribution. For Bernoulli sampling, if the prior is beta distribution, the posterior will be another beta distribution with different parameters. Otherwise, the posterior computation is hard.

Algorithm 1 Bernoulli Bandit Thompson Sampling

Require: each arm $i \in [k]$, set $S_i(1) = 0, F_i(1) = 0$
  for $t = 1, 2, \ldots, T$ **do**
    for each arm i **do**
      sample $\theta_i(t) \sim Beta(S_i(t) + 1, F_i(t) + 1)$
    end for
    Play arm $i(t) \triangleq \arg\max_j \theta_j(t)$, observe reward r
    if $r = 1$ **then**
      Increment $S_i(t)$
    else
      Increment $F_i(t)$
    end if
    Increment $N_i(t)$
  end for

Theorem 5. *(Agrawal-Goyal'12) For the TS algorithm described above, the expected regret*

$$\mathbb{E}_\nu[R(T)] \lesssim \sqrt{KT \log T} \tag{27}$$

for any bounded reward family ν.

Proof

The performance guarantee is the same as what we proposed in the past lecture, but the proof is different from UCB. To complete the proof, we first introduce some notations.

For each arm i, introduce two numbers x_i, y_i, such that $\mu_i < x_i < y_i < \mu_1$ $(i \neq 1)$. Assume arm 1 is the unique optimal arm.

Let $L_i(T) = \frac{\log T}{D(x_i \| y_i)}$, where $D(x_i \parallel y_i) \triangleq D(Bern(x_i) \parallel Bern(y_i))$ is the KL-Divergence between two Bernoulli distributions. We also define $\hat{\mu}_i(t) \triangleq \frac{S_i(t)}{N_i(t)}$ which is the empirical frequency of success for arm i.

Recall $\theta_i(t)$ is the sampled reward for arm i at time t. We then define two events, both of them are "good":

- $E_i^\mu(t) = \{\hat{\mu}_i(t) \leq x_i\}$
- $E_i^\theta(t) = \{\theta_i(t) \leq y_i\}$

Let's understand why those events are "good". Since $\hat{\mu}_i(t)$ is the empirical rewards at time t of arm i. Then if t is large enough, $\hat{\mu}_i(t)$ will get close to μ_i, which is upper-bounded by x_i. Similarly, $\theta_i(t)$ is the reward sampled from the posterior beta distribution which reflects the true reward distribution. Given sufficiently large t, $\theta_i(t)$ should also be close to μ_i, thus bounded by y_i with high probability.

We also introduce another notation to denote the history of information up to time $i-1$: $\mathcal{F}_{i-1} \triangleq \{(i(t), r_{i(t)}(t)) : 1 \leq t \leq i\}$, where $i(t)$ is the arm pulled at time t, $r_{i(t)}(t)$ is the corresponding reward.

The complete proof will be finished in next lecture. (cont'd) $\square$