Lower Bounds of Stochastic Bandits

MA5249 Presentation II

Dick Jessen William

NUS

November 2021

Outline

Stochastic Bandits Revisited

- We will have a look on Stochastic Bandits in a different way
- Instead on looking on one algorithm, we will look at all possible algorithm and prove that it cannot achieves some level of regret rate.

The $\Omega\sqrt{KT}$ Lower Bound

Theorem

For any time horizon T and total number of arms K. For any bandit algorithm, there exist a problem instance such that $\mathbb{E}[R(T)] \geq \Omega\sqrt{KT}$.

Proof.

Consider 0-1 rewards and the following family of problem instances, with $\epsilon > 0$ to be adjusted. For $j = \{1, 2, \cdots, K\}$, define I_j as below.

$$I_j = \begin{cases} \mu_i = \frac{1+\epsilon}{2} & \text{if i = j} \\ \mu_i = \frac{1}{2} & \text{otherwise} \end{cases}$$

$\square$

(continued).

By noting that Successive Elimination would sample every suboptimal arm at most $O(\epsilon^{-2}\log^k \epsilon^{-2})$ times, we note that sampling each arm $O(\epsilon^{-2}\log^k \epsilon^{-2})$ times suffices for our regret bound. Our goal is to prove that sampling each arm $\Theta(\epsilon^{-2})$ is necessary to check whether the arm is good or not. Hence, the regret is $\Theta(K/\epsilon)$. Choosing $\epsilon = \Omega(\sqrt{K/T})$ finishes our proof. The next sections will explore the technical details of this computation. $\square$

The rest of the section will explain the steps here.

KL-Divergence

Definition (KL-Divergence)

Consider a finite sample space Ω and p, q be two probability distribution on Ω. Then, define the KL-divergence as

$$KL(p, q) = \sum_{x \in \Omega} p(x) \ln \frac{p(x)}{q(x)} = \mathbb{E}_p \left[\ln \frac{p(x)}{q(x)} \right].$$

Some Useful Facts

Theorem (Gibbs)

For any distribution p, q, we have $KL(p, q) \geq 0$. Equality holds iff $p = q$.

Theorem (Chain Rule)

Let the sample space be a product $\Omega = \Omega_1 \times \Omega_2 \times \cdots \times \Omega_n$. Let p, q be two distributions of Ω such that $p = p_1 \times p_2 \cdots \times p_n$ and $q = q_1 \times q_2 \times \cdots \times q_n$, with p_i, q_i are distributions on Ω_j for $j \in \{1, 2, \cdots, n\}$. Then, $KL(p, q) = \sum_{i=1}^{n} KL(p_i, q_i)$.

Theorem (Pinsker)

For any event $A \in \Omega$, we have $2(p(A) - q(A))^2 \leq KL(p, q)$.

Theorem (Random Coins)

Let RC_ϵ denote a biased random coin with bias $\epsilon/2$ for a positive ϵ. Then, $KL(RC_\epsilon, RC_0) \leq 2\epsilon^2$ and $KL(RC_0, RC_\epsilon) \leq \epsilon^2$ for all $\epsilon \in (0, \frac{1}{2})$.

Using these theorems, we can prove this lemma.

Lemma

Consider sample space $\Omega = \{0,1\}^n$ and two distributions on Ω, $p = RC_\epsilon^n$ and $q = RC_0^n$. Then, there exists $\epsilon > 0$ such that for all $A \in \Omega$, $|p(A) - q(A)| \leq \epsilon\sqrt{n}$.

Flipping One Coin

- Define $\Omega = \{0,1\}^T$ as the sample space for the T coin tosses.
- We want to have a decision rule $Rule : \Omega \to \{HIGH, LOW\}$ that satisfies

$$P(Rule(Observations) = HIGH | \mu = \mu_1) \geq 0.99,$$

$$P(Rule(Observations) = LOW | \mu = \mu_2) \geq 0.99.$$

- We aim to find T such that $Rule$ exists.

Special Cases

Using the previous lemma, we can prove the following.

Lemma (Special Case when near 0.5)

Let $\mu_1 = \frac{1+\epsilon}{2}$ *and* $\mu_2 = \frac{1}{2}$. *With a decision rule like above, we have* $T > \frac{1}{4\epsilon^2}$.

- Consider the Best Arm Identification problem : Given a bandit problem, predict the most optimal arm.
- We will not consider the regret on the algorithm.

Definition (Good Algorithm for Best Arm Identification)

An algorithm is called good for best-arm identification if for all problem instances I, $P(y_T \text{ is the best arm} | T) \geq 0.99$.

We will the family of problem instances discussed earlier with parameter ϵ to argue that $T \geq \Omega(\frac{K}{\epsilon^2})$ for any working algorithm.

In fact, the following are true for two arms.

Lemma

Consider a best arm identification problem with $T \leq \frac{cK}{\epsilon^2}$ for a small positive constant. For any fixed deterministic algorithm, there exists at least $\lceil K/3 \rceil$ arms such that for the problem instance in the earlier page I_a, we have $P(y_t = a | I_a) < 0.75$.

Also, lemma implies this fact.

Corollary

Assuming T as above, if we fix any algorithm for best arm identification and we choose an arm a uniformly at random then running the algorithm on instance I_a, then $P(y_T \neq a) > \frac{1}{12}$.

Finally, we conclude the lower bound as follows.

Theorem ($\sqrt{KT}$ bound)

Fix time horizon T, number of arms K and a bandit algorithm. Run the algorithm on an instance I_a. Then, $\mathbb{E}[R(T)] \geq \Omega(\sqrt{KT})$.

- Using the proof for the case $K = 2$ only works if $T \leq c/\epsilon^2$.
- Consider an additional problem instance $I_0 = \{\mu_i = \frac{1}{2} \text{ for all arms } i\}$.
- Denote $\mathbb{E}[\cdot]$ be the expectation given this problem instance and T_a be the total number of times arm a is played.
- The following are true:
  - There are at least $2K/3$ arms j such that $\mathbb{E}_0(T_j) \leq 3T/K$.
  - There are at least $2K/3$ arms j such that $P_0(y_T = j) \leq 3/K$.
- Using Markov's inequality, we find out that we conclude that there are at least $K/3$ arms j such that $P(T_j \leq 24T/K) \geq 7/8$ and $P_0(y_T = j) \leq 3/K$.
- Fix an arm j satisfying the inequality above.
- The crux move : Prove that $P_j[Y_T = j] \leq 1/2$.

- Consider the sample space which j is played only $\min(T, 24T/K)$ times $\Omega^* = \Omega_j^m \times \prod_{a \neq j} \Omega_a^T$.
- Define the distribution P_l^* on Ω^* as $P_l^*(A) = P(A|I_l)$ $\forall A \subset \Omega^*$.
- Using KL-divergence argument, if $T \leq \frac{cK}{\epsilon^2}$ with small c, we have that $|P_0^*(A) - P_j^*(A)| \leq \epsilon\sqrt{m} < \frac{1}{8}$ for all $A \subset \Omega^*$.
- Using this, we can do some manipulations to conclude that $P_j(Y_T = j) \leq \frac{1}{2}$. Hence, our bound is proven.

The information theoretic approach implies stronger bounds for non-adaptive exploration.

Theorem

For any non-adaptive exploration, if we fix T and K with $K < T$. Then, there exists a problem instance such that $\mathbb{E}[R(T)] \geq \Omega(T^{2/3}K^{1/3})$.

The following version imposes a rule that the algorithm must not perform terribly in worst case.

Theorem

Keep the setup from above. If $\mathbb{E}[R(T)] \leq CT^{\gamma}$ for all problem instances, with $2/3 \geq \gamma < 1$. Then, for any problem instance, a random arms satisfies that $\mathbb{E}[R(T)] \geq \Omega(C^{-2}T^{2-2\gamma}\sum_a \Delta(a))$.

- The other fundamental lower bounds states that $\Omega(\log T)$ regret with an instance-dependent constant and applies to every problem instance.
- The lower bound can be used to combine the $\log T$ upper bound in UCB1 and Successive Elimination algorithms.

Theorem

No algorithm can have regret $\mathbb{E}[R(t)] = o(c_I \log t)$ for all problem instance I, for some constant c_I which depends on I but not t.

Hence, we have a guarantee that there is a problem instance which an algorithm has a high regret

Next, we see the case where we require an algorithm to perform good enough across every problem instance.

Theorem

For a fixed K, consider an algorithm such that $\mathbb{E}[R(t)] \leq O(C_{I,\alpha} t^{\alpha})$, $\forall I, \alpha > 0$. Here, $C_{I,\alpha}$ depends on I, α, but not on t. Now, fix a problem instance I. For this I, there exists t_0 such that $\forall t \geq t_0, \mathbb{E}[R(t)] \geq C_I \ln t$, with C_I depends on I but not t.

We can make this stronger.

Theorem

Keep the setup from the previous theorem. For any I and algorithm that satisfies the previous theorem,

1. *The bound works with $C_I = \sum_{\Delta(a)>0} \frac{\mu^*(1-\mu^*)}{\Delta(a)}$.*
2. *For each $\epsilon > 0$, the bound holds with $C_I = \sum_{\Delta(a)>0} \frac{\Delta(a)}{KL(\mu(a),\mu^*)} - \epsilon$.*

Some notable extensions in lower bounds are listed below.

- Lower bounds in dynamic pricing and Lipschitz bandits, researched by Kleinberg (2003).
- Linear Bandits, researched by Shamir (2015).
- For pay-per-click ad auctions, parametized by click probabilities learned over time, researched by Babaioff (2014).
- For dynamic pricing with limited supply and bandits with resource constraints, researched by Badanidiyuru (2018) and Besbes (2009).