IE 498: Online Learning and Decision Making, Fall 2019 Sep. 5th

Lecture 04: Pure Exploration Algorithms for MAB Problem

Lecturer: Yuan Zhou *Scribe: Juan Xu, Rahul Swamy*

In the last lecture, we discussed Multi-Armed Bandit (MAB) problem and analyzed the performance of multiple algorithms (e.g., the Upper Confidence Bound (UCB) method) in achieving the goal of identifying the best arm or the goal of minimizing regret. By assuming that $\mu_1 > \mu_2 \geq \mu_3 \geq \cdots \geq \mu_n$ and defining $\Delta_i = \mu_1 - \mu_i$ for $i \in \{2, 3, \ldots, n\}$, the regret of UCB method can be bounded as $R_T^{UCB} \lesssim \min\left\{(\log T)\sum_{i=2}^n \frac{1}{\Delta_i}, \sqrt{nT\log T}\right\}$. If the gap between the best arm and each of the remaining arms is large, i.e., Δ_i is large, the regret is bounded by the first term $(\log T)\sum_{i=2}^n \frac{1}{\Delta_i}$ which is small. If some gaps between the best arm and another arms are small, the regret is bounded by the second term $\sqrt{nT\log T}$ which is a fixed number. In this lecture, we mainly discuss some pure exploration algorithms for MAB problem with the assumption $\mu_1 > \mu_2 \geq \mu_3 \geq \cdots \geq \mu_n$ and a given "confidence parameter" $\delta \in (0, \frac{1}{2})$.

1 Successive Reject (SR) Algorithm

The idea of SR algorithm lies in that in each round t, we play each arm in the active set S_{t-1} for a certain number of times, and then use the empirical means to update the active set S_t by eliminating some bad arms in S_{t-1}. In the end, there is only one arm returned by the final round.

1. $S_0 \leftarrow [n], t \leftarrow 0$
2. WHILE $|S_t| > 1$ Do
   - 2a. $t \leftarrow t+1, \epsilon_t = 2^{-t}$
   - 2b. Play each arm in S_{t-1} by $\frac{C\cdot\log(nt^2/\delta)}{\epsilon_t^2}$ times
   - 2c. Set $S_t \leftarrow \{i \in S_{t-1} : \hat{\mu}_{it} \geq \max_{j\in S_{t-1}} \hat{\mu}_{jt} - \epsilon_t\}$
3. RETURN S_t.

Theorem 1. *With probability $1-\delta$,*
1) The SR algorithm returns the best arm 1.
2) The sample complexity $\lesssim \sum_{i=2}^n \Delta_i^{-2}\left(\log n + \log\delta^{-1} + \log\log\Delta_i^{-1}\right)$.

Proof. Define an event at round t as $E_t = \{1 \in S_t \text{ and } \forall j, \mu_j < \mu_1 - 2\epsilon_t \Rightarrow j \notin S_t\}$.

First, we have $\Pr[E_0] = 1$. At round $t = 0$, all arms are in the active set S_0, so arm 1 is naturally in S_0. Also, recall that in Lecture 03, we assume the mean of each arm μ_i lies in $[0, 1]$ for $i \in [n]$. Then by setting $\epsilon_0 = 2^{-0} = 1$, there won't exist an arm j such that $\mu_j < \mu_1 - 2\epsilon_0 = \mu_1 - 2$, so all arms are in S_0 for sure.

Before moving to the next step, for large enough C, by Hoeffding's Inequality, we have

$$\Pr\left[|\mu_i - \hat{\mu}_{it}| > \frac{\epsilon_t}{2}\right] \leq \frac{\delta}{3nt^2}, \quad \forall i \in [n], \forall t > 0.$$

By the result above and union bound, for every $t > 0$, we have

$$\Pr[E_t|E_{t-1}] \geq \Pr\left[\forall i \in S_{t-1}, \hat{\mu}_{it} \in \mu_i \pm \frac{\epsilon_t}{2}|E_{t-1}\right] \geq 1 - \frac{\delta}{3nt^2}n = 1 - \frac{\delta}{3t^2}.$$

For the first inequality above, we need to show event $\{\forall i \in S_{t-1}, \hat{\mu}_{it} \in \mu_i \pm \frac{\epsilon_t}{2}|E_{t-1}\}$ is a subset of event $E_t|E_{t-1}$. Given $\hat{\mu}_{it} \in \mu_i \pm \frac{\epsilon_t}{2}, \forall i \in S_{t-1}$, we have $\max_{j\in S_{t-1}} \hat{\mu}_{jt} - \epsilon_t \leq \max_{j\in S_{t-1}} \mu_j - \frac{\epsilon_t}{2} < \mu_1 - \frac{\epsilon_t}{2}$. Hence, $\hat{\mu}_{1t} \in \mu_1 \pm \frac{\epsilon_t}{2}$, implies $1 \in S_t$. When $\mu_i < \mu_1 - 2\epsilon_t$, the maximum value of $\hat{\mu}_{it} \leq \mu_i + \frac{\epsilon_t}{2} < \mu_1 - \frac{3\epsilon_t}{2}$. However, the minimum value of $\max_{j\in S_{t-1}} \hat{\mu}_{jt} - \epsilon_t \geq \max_{j\in S_{t-1}} \mu_j - \frac{3\epsilon_t}{2} = \mu_1 - \frac{3\epsilon_t}{2}$, which implies $j \notin S_t$ since $\hat{\mu}_{it}$ cannot be no less than $\max_{j\in S_{t-1}} \hat{\mu}_{jt} - \epsilon_t$.

Let $E = E_0 \wedge E_1 \wedge E_2 \wedge E_3 \wedge \ldots$. Event E refers to the event where only arm $1 \in S_t$.

$$\Pr[E] = \prod_{t=1}^{+\infty} \Pr[E_t|E_{t-1} \wedge E_{t-2} \wedge \cdots \wedge E_0] \geq \prod_{t=1}^{+\infty} \left(1 - \frac{\delta}{3t^2}\right) \geq 1 - \sum_{t=1}^{+\infty} \frac{\delta}{3t^2} \geq 1 - \frac{\delta}{3}\frac{\pi^2}{6} \geq 1 - \delta.$$

Therefore, we proved the first part of the theorem, i.e., $\Pr[E] \geq 1 - \delta$.

We now prove the sample complexity. For arm 1, the number of pulls is no more than the number of pulls to arm 2. For arm $i \geq 2$, the number of samples is $\sum_{t=1}^{\left\lceil \log_2 \frac{1}{\Delta_i} \right\rceil} \frac{C\cdot\log(nt^2/\delta)}{2^{-2t}} \lesssim \Delta_i^{-2}\left(\log\frac{n}{\delta} + \log\log\frac{1}{\Delta_i}\right)$. Note that for arm $i \geq 2$, on average, the maximum number of needed rounds is t such that $\mu_i < \max_j \mu_j - \epsilon_t = \mu_1 - \frac{1}{2^t}$, which implies that $\Delta_i > \frac{1}{2^t}$, i.e., $t > \log_2 \frac{1}{\Delta_i}$. Therefore, it is proved that the sample complexity $\lesssim \sum_{i=2}^{n} \Delta_i^{-2}\left(\log n + \log\delta^{-1} + \log\log\Delta_i^{-1}\right)$. □

Remark 1. *Theorem 1 tells us with probability $1-\delta$, what the sample complexity is. In fact, the expectation of the sample complexity has the sample bound as what we shown in Theorem 1. Lecture 05 will give a rigorous statement and proof.*

2 Uniform Sampling

For a given $\epsilon, \delta \in (0, 1)$, we say that an algorithm is $(\epsilon - \delta)-$*probably approximately correct* if

$$Pr[\mu_i \geq \mu_1 - \epsilon] \geq 1 - \delta,$$

where i is the returned arm.

Uniform sampling algorithm:

1. Play each arm $\frac{4ln(\frac{n}{\delta})}{\epsilon^2}$ times.
2. Return $\arg\max_i\{\hat{\mu}_i\}$.

Correctness: From Hoeffding's inequality,

$$Pr[\forall i : \hat{\mu}_i \in \mu_i \pm \frac{\epsilon}{2}] \geq 1 - \frac{\delta^2}{n^2} \cdot n = 1 - \frac{\delta^2}{n}.$$

When this happens, for the returned arm j,

$$\mu_j \geq \hat{\mu}_j - \frac{\epsilon}{2} \geq \hat{\mu}_1 - \frac{\epsilon}{2} \geq \mu_1 - \frac{\epsilon}{2} - \frac{\epsilon}{2}.$$

This proves that the Uniform sampling algorithm is $(\epsilon, \delta)-$probably approximately correct.

Sample complexity: Given by $\frac{4ln(\frac{n}{\delta})}{\epsilon^2} \cdot n$

3 Median Elimination

Median Elimination algorithm:

1. Let $S_0 \leftarrow [n]$, $r \leftarrow 0$
2. WHILE $|S_r| > 1$ Do
   (a) $r \leftarrow r + 1$
   (b) Sample each arm in S_{r-1} for $\frac{cr^4 log(\frac{r^2}{\delta})}{\epsilon^2}$ times
   (c) Sort arms from S_{r-1} according to $\hat{\mu}_{i,r}$; let S_r be among the top $\frac{|S_{r-1}|}{2}$ arms
3. Return S_r

Correctness: Note that in 2 (b) of the algorithm, there is no "n" in the log term. Therefore, we can expect that the top arms will still contain some good arms.

Let $i_r = \arg\max_{i \in S_r}\{\mu_i\}$ (best arm in the survived set at round r).

Let event $E_r = \{\mu_{i_r} \geq \mu_{i_{r-1}} - \frac{\epsilon}{3r^2}\}$.

Claim 1. $\forall\ r, Pr[E_r] \geq 1 - \frac{\delta}{2r^2}$.

Proof. For large enough c,

$$Pr\big[|\mu_{i_{r-1}} - \hat{\mu}_{i_{r-1}}| < \frac{\epsilon}{6r^2}\big] \geq 1 - \frac{\delta}{6r^2}.$$

Also,

$$\forall\ i \in S_{r-1},\ \mu_i < \mu_{i_{r-1}} - \frac{\epsilon}{3r^2},\ Pr\big[|\mu_i - \hat{\mu}_{i,r}| < \frac{\epsilon}{6r^2}\big] \geq 1 - \frac{\delta}{6r^2}.$$

Define a "bad i" for $i \in S_{r-1}$ if $\mu_i < \mu_{i_{r-1}} - \frac{\epsilon}{3r^2}$. Define a "terrible i" for $i \in S_{r-1}$ if i is "bad" and $\hat{\mu}_{i_r} > \mu_i + \frac{\epsilon}{6r^2}$. We are interested in: How many bad i's become terrible i's?

$$\mathbb{E}[\text{Number of terrible } i's] \leq \frac{\delta}{6r^2}|S_{r-1}|.$$

From Markov's inequality,

$$Pr[\text{Number of terrible arms} \geq \frac{|S_{r-1}|}{2}] \leq \frac{\delta}{3r^2}.$$

Hence,

$$Pr\big[|\mu_{i_{r-1}} - \hat{\mu}_{i_{r-1},r}| < \frac{\epsilon}{6r^2} \text{AND the number of terrible } i's \in S_{r-1} < \frac{|S_{r-1}|}{2}\big] \leq \frac{\delta}{2r^2}.$$

When the above event happens, we want the bad i's to be ranked before i_{r-1}. For that to happen, it has to be terrible. Hence, the number of bad i's ranked before $i_{r-1} \leq$ the number of terrible $i's < \frac{|S_{r-1}|}{2}$. Therefore, at least one not-bad i is in the top $\frac{|S_{r-1}|}{2}$ arms. This implies that $\mu_{i,r} \geq \mu_{i,r-1} - \frac{\epsilon}{3r^2}$. The rest of the proof is left an exercise. □

Sample complexity: $\lesssim \sum_{r=1}^{\infty} \frac{r^4 log(\frac{r^2}{\delta})}{\epsilon^2} \cdot \frac{n}{2^r} \lesssim \frac{n}{\epsilon^2} log(\frac{1}{\delta})$.