IE 498: Online Learning and Decision Making, Fall 2019 Sep 3rd, 2019

Lecture 03: Multi-Armed Bandit Problem

Lecturer: Yuan Zhou *Scribe: Menglong Li, Reza Yousefi Maragheh*

1 Warm-up: The better arm

Setting: Given 2 arms with unknown reward distribution D_i, $i = 1, 2$. Assume that D_i is supported on $[0, 1]$ and denote $\mu_i = \mathbb{E}[D_i]$, $i = 1, 2$.
Goal: identify $\arg\max_{i=1,2}\{\mu_i\}$.
A natural way to detect the arm with higher expected value (higher μ_i) is to play each arm equal times, say $T/2$ times for a given time horizon of T, and output the arm with the better empirical mean. The following algorithm uses this method:

ALG1

Step 1: Play arm 1 $T/2$ times and then play arm 2 $T/2$ times.
Step 2: Calculate empirical means $\hat{\mu}_1, \hat{\mu}_2$ and output $\arg\min_{i=1,2}\{\hat{\mu}_i\}$.

When performing algorithm one, its success in detecting the better arm is dependent on the distance between the values of μ_1 and μ_2. In other words, if the distance is large enough, then our confidence about the result of *ALG1* is more than the case when μ_1 and μ_2 are close to each other. For instance, when $\mu_1 = 0.1$ and $\mu_2 = 0.8$ we can be sure about the correctness of output in *ALG1* more than the situation where $\mu_1 = 0.49$ and $\mu_2 = 0.51$.

Denote $\Delta = |\mu_1 - \mu_2|$. The following result shows that ALG1 outputs the correct arm (the better arm) with high probability and this probability is a function of Δ.

Theorem 1. *ALG1 outputs the better arm with probability at least* $1 - 4\exp(-\Delta^2 T/4)$.

Proof. WLOG, assume $\mu_1 > \mu_2$. Denote $\Delta = |\mu_1 - \mu_2|$. By Hoeffding's inequality,

$$\Pr[|\mu_i - \hat{\mu}_i| \leq \frac{\Delta}{2}] \geq 1 - 2\exp(-\Delta^2 T/4), \qquad i = 1, 2.$$

Define the events A and B as the cases when $\{|\mu_1 - \hat{\mu_1}| \leq \frac{\Delta}{2}\}$ and $\{|\mu_2 - \hat{\mu_2}| \leq \frac{\Delta}{2}\}$ respectively. In this way, the event $A \cap B$ can be denoted by $\{\max_{i=1,2}|\mu_i - \hat{\mu_i}| \leq \frac{\Delta}{2}\}$. By using the union bound: $\Pr(A \cap B) \geq 1 - \Pr(A^c) - \Pr(B^c)$ we get:

$$\Pr[\max_{i=1,2}\{|\mu_i - \hat{\mu}_i|\} < \frac{\Delta}{2}] \geq 1 - 4\exp(-\Delta^2 T/4).$$

Observe that if $\max_{i=1,2}\{|\mu_i - \hat{\mu}_i|\} < \frac{\Delta}{2}$, then $\hat{\mu}_1 > \mu_1 - \frac{\Delta}{2} > \mu_2 - \frac{\Delta}{2} > \hat{\mu}_2$. Hence, ALG1 outputs the correct arm with probability at least $1 - 4\exp(-\Delta^2 T/4)$. □

Remark 1. *If ALG1 wants to succeed with probability* $1-\delta$*,* T *should be at least* $\frac{4}{\Delta^2}\log(\frac{4}{\delta})$.

2 Regret minimization

In some scenarios, the objective may be maximizing the rewards collected instead of identifying the better arm. In these cases, we want to select one arm in each period to maximize the expected overall rewards $\mathbb{E}[\sum_{t=1}^{T} r_t]$, where r_t is the corresponding reward in period t, i.e., r_t follows D_i if selecting arm i in period t, $i = 1, 2$. This problem of maximizing the collected revenue is equivalent to the following problem:

$$\min R_T,$$

where R_T is the expected regret defined as $R_T = T\max\{\mu_1, \mu_2\} - \mathbb{E}[\sum_{t=1}^{T} r_t]$. Note that $T\max\{\mu_1, \mu_2\}$ is the maximal expected reward by playing the best arm all the time. The following result shows that the regret of *ALG1* is linear in T.

Theorem 2. $R_T^{ALG1} = \frac{T}{2}\Delta$.

Proof. Suppose $\mu_1 > \mu_2$. Because ALG1 play arm 2 in $\frac{T}{2}$ times. It follows that the regret is $\frac{T}{2}\Delta$. □

As the regret of *ALG1* is of order $\Omega(T)$, it is a relatively bad regret. Thus, it is natural to think about othe algorithms.

We consider another algorithm that first run *ALG1* for some time and then play the best arm identified by *ALG1* for the remaining time.

Exploration and commit Algorithm

Step 1: Call ALG1 with parameter T_0, i.e., play each arm $\frac{T_0}{2}$ times.
Step 2: play the best arm returned by ALG1 for the remaining $T - T_0$ periods.

Theorem 3. *By choosing* $T_0 = T^{2/3}\log^{1/3}(T)$*, the regret of exploration and commit algorithm is* $O(T^{2/3}\log^{1/3}(T))$.

Proof. The regret incurred for the first T_0 periods is $\frac{T_0}{2}\Delta$. The regret incurred for the second step is

$$\Pr[\text{ALG1 fails}](T - T_0)\Delta \leq 4\exp(-\frac{T_0\Delta^2}{4})T\Delta.$$

Hence,

$$R_T \lesssim T_0\Delta + 4\exp(-\frac{T_0\Delta^2}{4})T\Delta.$$

Here, for two numbers a, b, $a \lesssim b$ means that $a \leq bc$ for some constant $c > 0$. Denote $S = T_0\Delta + 4\exp(-\frac{T_0\Delta^2}{4})T\Delta$. We consider two cases:
(1) If $\Delta^2 \leq \frac{4\log(T)}{T_0}$, then $S \leq T\Delta + 4T\Delta \lesssim T\sqrt{\frac{\log(T)}{T_0}}$
(2) If $\Delta^2 > \frac{4\log(T)}{T_0}$, then $\exp(-\frac{T_0\Delta^2}{4})T < 1$, and thus $S \leq (T_0+1)\Delta \leq T_0 + 1 \lesssim T_0$.
(The threshold of Δ^2 is choose by the following intuition: when Δ is large, $T_0\Delta$ dominants $4\exp(-\frac{T_0\Delta^2}{4})T\Delta$, while when Δ is small, $4\exp(-\frac{T_0\Delta^2}{4})T\Delta$ dominants $T_0\Delta$. So it is reasonable to choose a threshold that makes $T_0\Delta \approx 4\exp(-\frac{T_0\Delta^2}{4})T\Delta$.)
Hence, $R_T \leq T\sqrt{\frac{\log(T)}{T_0}} + T_0$. Let $T_0 = T^{2/3}\log^{1/3}(T)$ (this threshold is choose with the same reason). Then $R_T \leq 2T^{2/3}\log^{1/3}(T)$. □

Remark 2. *This regret bound $O(T^{2/3}\log^{1/3}(T))$ is tight for the exploration and commit algorithm.*

The exploration and commit algorithm is not always reliable as shown by the following example. Consider the Bernoulli bandit problem, where D_1, D_2 are Bernoulli distributions and $\mu_1 = 0.9, \mu_2 = 0.1$. We first explore each arm once, and commit the best arm for the remaining periods. Suppose the realized reward for arm 1 is 0 and the realized reward for arm 2 is 1. Then the algorithm will play arm 2 afterwards, which is bad! Playing the wrong arm for ever is bad!

A natural way to avoid this is to try the other arm with some probability in each period. More specifically, we play the arm with the best empirical mean with probability $1-\epsilon$ and choose a random arm with probability ϵ.

ϵ-Greedy:

In each time t, play $\begin{cases} \arg\max_i\{\hat{\mu}_i\} & w.p.\ 1-\epsilon \\ \text{random arm} & w.p.\ \epsilon \end{cases}$

3 The upper-confidence bound (UCB) method

We propose another algorithm that beats the regret bound $O(T^{2/3}\log^{1/3}(T))$ of the exploration and commit algorithm. We now work with n arms. Without loss of generality, we assume $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n$. Denote $\Delta_i = \mu_1 - \mu_i$. Denote $T_i(t)$ as the number of plays of arm i before time $t+1$, $\hat{\mu}_{i,t}$ as the empirical mean for μ_i based on the first t samples for all arms. Note that $\hat{\mu}_{i,t}$ is the average of the $T_i(t)$ realized rewards of arm i. By Hoeffding's inequality,

$$\Pr[|\mu_i - \hat{\mu}_{i,t}| \leq \sqrt{\frac{\log(2Tn)}{T_i(t)}}] \geq 1 - \frac{1}{2T^2n^2}.$$

That is, with high probability $1 - \frac{1}{2T^2n^2}$, the up-to-date empirical mean $\hat{\mu}_{i,t}$ is in the interval $[\mu_i - \sqrt{\frac{\log(2Tn)}{T_i(t)}}, \mu_i + \sqrt{\frac{\log(2Tn)}{T_i(t)}}]$.

One can see that when t becomes large, this interval will concentrate more on μ_i, i.e., the empirical mean is more accurate. The upper-confidence bound of each arm at the beginning of time t is $\hat{\mu}_{i,t-1} + \sqrt{\frac{\log(2Tn)}{T_i(t-1)}}$. The UCB algorithm play the arm with the highest upper-confidence bound in each time.

UCB algorithm

Step 1: Play each arm once.
Step 2: At time $t \in \{n+1, ..., T\}$, select arm $i_t = \arg\max_i\{\hat{\mu}_{i,t-1} + \sqrt{\frac{\log(2Tn)}{T_i(t-1)}}\}$.

Theorem 4 (parameter dependent bound)**.** $R_T^{UCB} \lesssim \log(T)\sum_{i=1}^n \frac{1}{\Delta_i}$.

Proof. Let event $E_i = \{|\mu_i - \hat{\mu}_{i,t}| \leq \sqrt{\frac{\log(2Tn)}{T_i(t)}},\ \forall t\}$ and $E = \cap_{i=1}^n E_i$. For any event B, denote B^c as the

complement of even B. Again,

$$\Pr(E) = 1 - \Pr(\cup_{i=1}^n E_i^c) \geq 1 - \sum_{i=1}^n \Pr(E_i^c) \geq 1 - n\frac{1}{2T^2n^2} = 1 - \frac{1}{2T^2n}.$$

$\Pr(E^c) \leq \frac{1}{2T^2n}$. If even E holds, we show that the number of arm i played in the T periods is $\lesssim \frac{\log(Tn)}{\Delta_i^2}$. Let t_i be the period that arm i is last played. Then

$$\mu_i + 2\sqrt{\frac{\log(2Tn)}{T_i(t_i-1)}} \geq \mu_{i,t_i-1} + \sqrt{\frac{\log(2Tn)}{T_i(t_i-1)}} \geq \mu_{1,t_i-1} + \sqrt{\frac{\log(2Tn)}{T_1(t_i-1)}} \geq \mu_1.$$

The first inequality is from even E_i, the second inequality is from the fact that arm i is played at time t_i and thus $i = \arg\max_j\{\hat{\mu}_{j,t_i-1} + \sqrt{\frac{\log(2Tn)}{T_j(t_i-1)}}\}$, the third inequality is from even E_1. Hence, $\mu_i + 2\sqrt{\frac{\log(2Tn)}{T_i(t_i-1)}} \geq \mu_1$, then $2\sqrt{\frac{\log(2Tn)}{T_i(t_i-1)}} \geq \Delta_i$ and $T_i(t_i-1) \lesssim \frac{\log(Tn)}{\Delta_i^2}$. Hence, the number of armed i played is $\lesssim \frac{\log(Tn)}{\Delta_i^2}$. With these facts,

$$R_T^{UCB} \lesssim \Pr(E)\sum_{i=2}^n \frac{\log(Tn)}{\Delta_i^2} + \Pr(E^c)T \lesssim \sum_{i=2}^n \frac{\log(Tn)}{\Delta_i^2} \lesssim \sum_{i=2}^n \frac{\log(T)}{\Delta_i^2}.$$

The last inequality is from $T \geq n$. □

Theorem 5 (parameter independent bound)**.** $R_T^{UCB} \lesssim \sqrt{nT\log(T)}$.

Proof. When event E happens, the number of arm i played $T_i(T) \lesssim \frac{\log(Tn)}{\Delta_i^2}$. Hence, $\Delta_i \lesssim \sqrt{\frac{\log(nT)}{T_i(T)}}$. Then

$$\begin{aligned}
R_T^{UCB} &\leq \mathbb{E}[\sum_{i=2}^n T_i(T)\Delta_i \mid \text{even } E \text{ holds}] + \Pr(E^c)T \\
&\lesssim \mathbb{E}[\sum_{i=2}^n T_i(T)\Delta_i \mid \text{even } E \text{ holds}] \\
&= \mathbb{E}[\sum_{i=2}^n \sqrt{\log(nT)T_i(T)} \mid \text{even } E \text{ holds}] \\
&\leq \mathbb{E}[(n-1)\sqrt{\log(nT)\sum_{i=2}^n \frac{T_i(T)}{n-1}} \mid \text{even } E \text{ holds}] && \text{(Jesson's inequality)} \\
&\leq \sqrt{(n-1)\log(nT)T} && \text{(total number of arms played is at most } T\text{)} \\
&\lesssim \sqrt{nT\log(T)} && (T \geq n)
\end{aligned}$$

□

Remark 3 (Lower bound)**.** *Even for Bernoulli bandits with* $\mu_i \in [0.1, 0.9]$,

$$\liminf_{T\to\infty} R_T^\Pi \gtrsim \sum_{i=1}^n \frac{1}{\Delta_i}$$

for any policy Π*. Moreover, there exists an instance such that* $R_T^\Pi \gtrsim \sqrt{nT}$ *for any policy* Π*.*