[CS3958: Lecture 5] Multi-Armed Bandit, Explore-then-Commit, Upper-Confidence-Bound

Instructor: Chihao Zhang, Scribed by Yulin Wang

November 9, 2022

Today we will start a new topic – online optimization. We begin with a classic problem called *Multi-Armed Bandit (MAB)*.

1 The Problem Setting

Suppose there is a k-arm bandit, and the reward of each arm follows some distribution $f_i \in [0,1]$ with mean μ_i. We assume without loss of generality that $\mu_1 \ge \mu_2 \ge \cdots \ge \mu_k$. Now suppose you can pull the bandit for T rounds and the goal is to obtain maximum reward in expectation. If we know $\mu_1, \dots, \mu_k$ well, the optimal strategy is to pull the arm 1 for T times, and the expected reward is $T\mu_1$. However, in case that we do not know the distribution, we have to design some strategy to explore the bandit first.

Denote by a_t the arm pulled at the round t, and thus we have the reward in the t-th round $X_t \sim f_{a_t}$. The *regret* of a strategy is defined as the gap between $T\mu_1$ and expected rewards of the strategy in T rounds, namely the regret of not always choosing the first arm:

$$R(T) \triangleq T\mu_1 - \mathbf{E}\left[\sum_{t=1}^{T} X_t\right] \ge 0.$$

Note that in the expression above, the randomness of the expectation $\mathbf{E}[\cdot]$ usually comes from two parts: the randomness of the distributions f_i and the (possible) randomness of the strategy.

For every $i \in [k]$, we denote $\Delta_i \triangleq \mu_1 - \mu_i \ge 0$ as the gap between the reward of the i-th arm and the optimal arm. The naive strategy that pulls each arm for equal times is bad, since its regret $R(T) = \frac{\sum_{i=1}^k \Delta_i}{k} \cdot T$, which is linear in T. We consider a strategy/algorithm good if it holds that $\lim_{T\to\infty} R(T)/T = 0$ or equivalently $R(T) = o(T)$.

The following simple observation is useful when analyzing randomized algorithms for MAB.

Proposition 1 *For every $t \in [T]$, let $n_i(t) \triangleq \sum_{s=1}^{t} \mathbf{1}[a_s = i]$ denote the number of times that the arm i is pulled in the first t rounds. Then*

$$R(T) = \sum_{i=2}^{k} \Delta_i \cdot \mathbf{E}\left[n_i(T)\right].$$

Proof.

$$\begin{aligned} R(T) &= T\mu_1 - \mathbf{E}\left[\sum_{t=1}^{T} X_t\right] \\ &= T\mu_1 - \sum_{t=1}^{T} \mathbf{E}_{a_t}\left[\mu_{a_t}\right] \\ &= \sum_{t=1}^{T}\sum_{i=1}^{k} \Delta_i \cdot \mathbf{E}\left[\mathbf{1}[a_t = i]\right] \\ &= \sum_{i=1}^{k} \Delta_i \cdot \mathbf{E}\left[\sum_{t=1}^{T} \mathbf{1}[a_t = i]\right] \\ &= \sum_{i=1}^{k} \Delta_i \cdot \mathbf{E}\left[n_i(T)\right]. \end{aligned}$$

We also write $R_i(T) \triangleq \Delta_i \cdot \mathbf{E}\left[n_i(T)\right]$ for every $i \in [k]$, and then $R(T) = \sum_{i=1}^{k} R_i(T)$. □

2 *The Explore-then-Commit (ETC) Algorithm*

To get small regret, our strategy should identify the best arm as soon as possible. The most straightforward way to find the best arm is to try each arm a few times and pick the one with best *empirical reward*. The Explore-then-Commit algorithm implements this idea: Pull every arm i for L times (so kL times in total for exploration), and calculate $\hat{\mu}_i$ (the average reward gained in that L times). After this, always pull the arm with greatest $\hat{\mu}_i$. We can write its regret as

$$\begin{aligned} R(T) &= L\sum_{i=1}^{k}\Delta_i + \sum_{i=2}^{k}\Delta_i \cdot \sum_{t=kL+1}^{T} \mathbf{Pr}\left[\hat{\mu}_i > \max_{j\neq i}\hat{\mu}_j\right] \\ &= L\sum_{i=1}^{k}\Delta_i + \sum_{i=2}^{k}\Delta_i \cdot (T-kL)\mathbf{Pr}\left[\hat{\mu}_i > \max_{j\neq i}\hat{\mu}_j\right]. \end{aligned}$$

When $i \neq 1$,

$$\mathbf{Pr}\left[\hat{\mu}_i > \max_{j\neq i}\hat{\mu}_j\right] \leq \mathbf{Pr}\left[\hat{\mu}_i > \hat{\mu}_1\right].$$

We bound above probability by concentration inequalities. To this end, let X_j be the j-th reward of f_i, Y_j be the j-th reward of f_1. Let $Z_j = X_j - Y_j \in [-1,1]$, then $\mathbf{E}\left[Z_j\right] = -\Delta_i \leq 0$. Let $Z = \sum_{j=1}^{L} Z_j$, then $\mathbf{E}\left[Z\right] = -L\Delta_i$.

By Hoeffding's Inequality,

$$\mathbf{Pr}\left[\hat{\mu}_i > \hat{\mu}_1\right] = \mathbf{Pr}\left[Z > 0\right] = \mathbf{Pr}\left[Z - \mathbf{E}\left[Z\right] \geq L\Delta_i\right] \leq \exp\left(-\frac{2(L\Delta_i)^2}{\sum_{j=1}^{L} 2^2}\right) = \exp\left(-\frac{L\Delta_i^2}{2}\right).$$

Therefore we have

$$\begin{aligned} R(T) &\le L \sum_{i=1}^{k} \Delta_i + (T - kL) \sum_{i=2}^{k} \Delta_i \exp\left(-\frac{L\Delta_i^2}{2}\right) \\ &\le \sum_{i=1}^{k} \left(L\Delta_i + T\Delta_i \exp\left(-\frac{L\Delta_i^2}{2}\right)\right) \le \sum_{i=1}^{k} \left(L + T\Delta_i \exp\left(-\frac{L\Delta_i^2}{2}\right)\right). \end{aligned}$$

To further upper bound $R(T)$, we define

$$g(L, \Delta_i) \triangleq L + T\Delta_i \exp\left(-\frac{L\Delta_i^2}{2}\right).$$

We would like to determine L minimizing the upper bound of $R(T)$ among all possible Δ_i, i.e., $\min_L \max_{\Delta_i} R(T)$. First we calculate $\max_{\Delta_i} g(L, \Delta_i)$:

$$\frac{\partial g(L, \Delta_i)}{\partial \Delta_i} = T(1 - L\Delta_i^2) \exp\left(-\frac{L\Delta_i^2}{2}\right).$$

We have $\frac{\partial g(L,\Delta_i)}{\partial \Delta_i} > 0$ when $0 \le \Delta_i < \frac{1}{\sqrt{L}}$, and $\frac{\partial g(L,\Delta_i)}{\partial \Delta_i} < 0$ when $1 \ge \Delta_i > \frac{1}{\sqrt{L}}$. Thus, for all $L > 1$,

$$g(L, \Delta_i) \le g(L, \frac{1}{\sqrt{L}}) = L + \frac{Te^{-1/2}}{\sqrt{L}}.$$

Finally,

$$R(T) \le \sum_{i=1}^{k} (L + \frac{Te^{-1/2}}{\sqrt{L}}) = \Theta(kT^{\frac{2}{3}}),$$

by setting $L = \Theta(T^{\frac{2}{3}})$.

The Explore-then-Commit algorithm enjoys sublinear reget, which is good, but still suboptimal. The main disadvantage is that it treats all arms equally in the exploration step and pulls each of them for fixed L times regardless of the rewards already obtained.

3 The Upper Confidence Bounds (UCB) Algorithm

Therefore in order to overcome the weakness of ETC, during the exploration phase, the algorithm should *adaptively* make use of the information obtained so far. The brilliant idea of the UCB algorithm is to adaptively maintain an interval $[a_i, b_i]$ for each arm i so that the mean μ_i is within the interval with high probability based on the current knowledge on μ_i.

Now suppose you already know $\mu_i \in [a_i, b_i]$ for each arm $i \in [k]$ with high probability after some exploration, which arm will you pull now? The name *upper confidence bound* means that we always choose the one with the highest upper bound b_i.

This sounds like you are walking on a snack street in a country that you have never been to. There is a Chinese canteen, which you are very familiar with. The food there is at least not bad, but can never be surprisingly wonderful. Besides, there is also a local canteen. As you have little idea about the local food, it may taste horrible, but also has a possibility to have a heavenly good taste. The upper confidence tells you to walk into the local canteen, even with more risk to take an extremely bad dinner.

In order to implement the idea, we have to specify how to maintain the an interval for each arm. Formally, for all $t \in [T]$ and $i \in [k]$, at round t we not only track $\hat{\mu}_i(t)$, but also maintain an interval $[a_i(t), b_i(t)]$ so that $\mu_i \in [a_i(t), b_i(t)]$ with probability no less than $1 - \delta_i(t)$ for some parameter $\delta_i(t)$ to be chosen later. Let $a_i(t) \triangleq \hat{\mu}_i(t) - c_i(t)$ and $b_i(t) \triangleq \hat{\mu}_i(t) + c_i(t)$. Let us see how to pick $c_i(t)$. In the discussion below, we may drop t if it is clear from the context.

By Hoeffding's inequality, c_i should meet

$$\mathbf{Pr}\left[|\mu_i - \hat{\mu}_i| > c_i\right] \le 2\exp\left(-2n_i c_i^2\right) \le \delta_i,$$

so we choose $c_i = \sqrt{\frac{\log(2/\delta_i)}{2n_i}}$.

Note that the upper bound $b_i = \hat{\mu}_i + c_i$ can be large (which means that we are more likely to explore the arm i) if either μ_i is large (so the i-th arm is good), or n_i is small (so the i-th arm is not well-explored).

Bounding the regret $R(T)$

For all $i \in [k]$, the regret contributed by the arm i is

$$R_i(T) = \Delta_i \mathbf{E}\left[n_i(T)\right] \le \Delta_i \sum_{t=1}^{T} \mathbf{Pr}\left[\hat{\mu}_i(t) + c_i(t) \ge \max_{j \ne i} \hat{\mu}_j(t) + c_j(t)\right].$$

The probability $\mathbf{Pr}\left[\hat{\mu}_i(t) + c_i(t) \ge \max_{j \ne i} \hat{\mu}_j(t) + c_j(t)\right]$ can be controlled only when each μ_i is in $[a_i(t), b_i(t)]$. Therefore, it is a common trick to decompose the probability with this desired good event. We define events

$$\begin{aligned} \mathcal{A} &: \text{ Every } \mu_i \text{ is in its interval } [a_i(t), b_i(t)] \text{ at any time;} \\ \mathcal{B}_i(t) &: \hat{\mu}_i(t) + c_i(t) \ge \max_{j \ne i} \hat{\mu}_j(t) + c_j(t). \end{aligned}$$

Thus,

$$\begin{aligned} R_i(T) &= \Delta_i \cdot \sum_{t=1}^{T} \mathbf{Pr}\left[\mathcal{B}_i(t)\right] \\ &= \Delta_i \cdot \sum_{t=1}^{T} \left(\mathbf{Pr}\left[\mathcal{B}_i(t) \mid \mathcal{A}\right] \mathbf{Pr}\left[\mathcal{A}\right] + \mathbf{Pr}\left[\mathcal{B}_i(t) \,\middle|\, \overline{\mathcal{A}}\right] \mathbf{Pr}\left[\overline{\mathcal{A}}\right]\right) \\ &\le \Delta_i \cdot \left(\sum_{t=1}^{T} \mathbf{Pr}\left[\mathcal{B}_i(t) \mid \mathcal{A}\right] + \sum_{t=1}^{T} \mathbf{Pr}\left[\overline{\mathcal{A}}\right]\right). \end{aligned}$$

We then bound $\sum_{t=1}^{T} \mathbf{Pr}\left[\overline{\mathcal{A}}\right]$ and $\sum_{t=1}^{T} \mathbf{Pr}\left[\mathcal{B}_i(t) \mid \mathcal{A}\right]$ respectively.

- Note that $\overline{\mathcal{A}}$ is the event "$\exists t \in [T], \exists i \in [k], \mu_i \notin [a_i(t), b_i(t)]$", by union bound, we have

  $$\mathbf{Pr}\left[\overline{\mathcal{A}}\right] \le \sum_{i=1}^{k}\sum_{t=1}^{T} \delta_i(t).$$

  Therefore if we choose $\delta_i(t) = 1/T^2$ for all $i \in [k]$ and $t \in [T]$, then

  $$\sum_{t=1}^{T} \mathbf{Pr}\left[\overline{\mathcal{A}}\right] \le T \cdot kT \cdot \frac{1}{T^2} = k.$$

- Since conditioned on $\mathcal{A}$, $\mu_i \in [a_i(t), b_i(t)]$ for all $i \in [k]$ and $t \in [T]$, we have

  $$\begin{cases} \hat{\mu}_i(t) + c_i(t) \le (\mu_i + c_i(t)) + c_i(t) = \mu_i + 2c_i(t) \\ \hat{\mu}_1(t) + c_1(t) \ge (\mu_1 - c_1(t)) + c_1(t) = \mu_1 \end{cases}.$$

  Therefore $\mu_i + 2c_i(t) \le \mu_1$ is a sufficient condition of $\mathcal{B}_i(t)$ not happening conditioned on $\mathcal{A}$. With $\delta = 1/T^2$, we have

  $$\begin{aligned} B_i(t) \text{ not happening conditioned on } \mathcal{A} &\Longleftarrow \mu_i + 2c_i(t) \le \mu_1 \\ &\Longleftrightarrow \sqrt{\frac{\log(2/\delta_i(t))}{2n_i(t)}} \le \frac{\Delta_i}{2} \\ &\Longleftrightarrow n_i(t) \ge \frac{4\log\left(\sqrt{2}T\right)}{\Delta_i^2} \\ &\Longleftarrow n_i(t) \ge \frac{6\log T}{\Delta_i^2}. \end{aligned}$$

  This indicates that if $n_i(t) \ge \frac{6\log T}{\Delta_i^2}$, $\mathcal{B}_i(t)$ will never happen conditioned on $\mathcal{A}$. The fact implies that

  $$\sum_{t=1}^{T} \mathbf{Pr}\left[\mathcal{B}_i(t) \mid \mathcal{A}\right] = \sum_{t=1}^{T} \mathbf{E}\left[\mathbf{1}[\mathcal{B}_i(t)] \mid \mathcal{A}\right] \le \frac{6\log T}{\Delta_i^2}.$$

So far we have found an upper bound for $\mathbf{Pr}\left[\mathcal{B}_i(t) \mid \mathcal{A}\right]$, but it may be very large if Δ_i is close to zero. However, remember that if Δ_i is small, pulling the i-th arm only causes little regret, and thus we can divide the k arms into two groups of $\Delta_i \le \Delta$ and $\Delta_i > \Delta$ for some threshold Δ. Then we can calculate the correspond regret separately as follows:

$$\begin{aligned}
R(T) &= \sum_{i=1}^{k} \Delta_i \mathbf{E}\left[n_i(T)\right] \\
&= \sum_{i:\Delta_i \le \Delta} \Delta_i \mathbf{E}\left[n_i(T)\right] + \sum_{i:\Delta_i \ge \Delta} \Delta_i \mathbf{E}\left[n_i(T)\right] \\
&\le T\Delta + \sum_{i:\Delta_i > \Delta} \Delta_i \left(\frac{6 \log T}{\Delta_i^2} + k \right) \\
&\le T\Delta + \frac{6k \log T}{\Delta} + k^2 \\
&= \Theta(\sqrt{kT \log T}),
\end{aligned}$$

by setting $\Delta = \sqrt{\frac{6k \log T}{T}}$. This regret bound is optimal up to a $\sqrt{\log T}$ factor.