Advanced Algorithms VII (Fall 2020)

Instructor: Chihao Zhang
Scribed by: Haichen Dong, Yi Gu

Last modified on Nov. 1, 2020

Concentration inequalities are ubiquitous in the analysis of randomized algorithms. Today we will see that they also help while designing algorithms.

We will study the Multi-Armed Bandit (MAB) problem. MAB is an important model for online decision making that has been widely studied. Today we will study the simplest setting of the problem and two classic algorithms.

1 The Problem Setting

Suppose there is a k-arm bandit, and the reward of each arm is some distribution $f_i \in [0,1]$ with $\mathbf{E}[f_i] = \mu_i$. W.l.o.g, suppose $\mu_1 > \mu_2 \geq \cdots \geq \mu_k$. Now suppose you can pull the bandit for T rounds. If we learn $\mu_1, \dots, \mu_k$ well, the optimal strategy is to pull the arm 1 for T times, and the expected reward is $T\mu_1$. However, we currently have no idea about the k arms, and have to explore the bandit during the T rounds.

Denote by a_t the arm pulled at 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:

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

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 randomness from the strategy. For every $i \in [k]$, we denote $\Delta_i \triangleq \mu_1 - \mu_i \geq 0$. A naive algorithm pulls each arm for equal times, resulting in $R(T) = \frac{\sum_{i=1}^{k} \Delta_i}{k} \cdot T$ 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 theorem is useful when analyzing randomized algorithms for MAB.

Theorem 1. $\forall t \in [T], n_i(t) = \sum_{s=1}^{t} 1_{a_s = i}$ *denotes the number of times arm i pulled in the first t rounds. Then*

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

Proof.

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

□

We also write $R_i(T) = \Delta_i \cdot \mathbf{E}[n_i(T)]$ 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 a small regret, the 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 compare their rewards. The Explore-then-Commit algorithm implements this idea: Pull every arm for L times (within total kL times), and calculate $\hat{\mu}_i$ (the average reward gained in that L times). Then always pull the arm with highest $\hat{\mu}_i$. We can write its regret as

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

When $i \neq 1$,

$$\mathbf{Pr}\left[\hat{\mu}_i > \max_{j\neq i}\hat{\mu}_j\right] \leq \mathbf{Pr}\left[\hat{\mu}_i > \hat{\mu}_1\right]. \tag{2}$$

We bound (2) by concentration inequalities. To this end, let X_j be the jth reward of f_i, Y_j be the jth 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 $\mathbf{E}[Z] = -L\Delta_i$.

By Hoeffding's Inequality,

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

Combining this with (1), we have

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

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

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

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

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

We have $\dfrac{\partial g(L,\Delta_i)}{\partial \Delta_i} > 0$ when $0 \leq \Delta_i < \dfrac{1}{\sqrt{L}}$, and $\dfrac{\partial g(L,\Delta_i)}{\partial \Delta_i} < 0$ when $1 \geq \Delta_i > \dfrac{1}{\sqrt{L}}$.

Thus, for all $L > 1$,

$$g(L, \Delta_i) \leq g(L, \frac{1}{\sqrt{L}}) = L + \frac{T\mathrm{e}^{-1/2}}{\sqrt{L}}.$$

Finally,

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

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

The Explore-then-Commit algorithm works better than the naive one but still has some disadvantages. It treats all arms equally in the exploration step and pulls each of them L times regardless of the rewards already obtained. The regret bound $O\left(T^{\frac{2}{3}}\right)$ is suboptimal.

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 $i \in [k]$, on 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 a parameter $\delta_i(t)$ to be chosen later. Let $a_i(t) := \hat{\mu}_i(t) - c_i(t)$ and $b_i(t) := \hat{\mu}_i(t) + c_i(t)$. Let us see how to pick $c_i(t)$. In the below, when t is clear from the context, we may drop it.

By Hoeffding's inequality, c_i should meet

$$\mathbf{Pr}\left[|\mu_i - \hat{\mu}_i| > c_i\right] \leq 2\exp(-3n_i c_i^2) \leq \delta_i \implies c_i \geq \sqrt{\frac{\log(2/\delta_i)}{2n_i}}.$$

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 $\hat{\mu}_i$ is large, or n_i is small, which means we have not explored it enough.

3.1 Bounding the regret $R(T)$

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

$$R_i(T) = \Delta_i \mathbf{E}\left[n_i(T)\right] \leq \Delta_i \sum_{t=1}^{T} \mathbf{Pr}\left[\hat{\mu}_i(t) + c_i(t) \geq \hat{\mu}_1(t) + c_1(t)\right].$$

We define events

$$\begin{aligned}\mathcal{A} &: \text{ Every one is in its interval at any time;}\\ \mathcal{B}_i(t) &: \hat{\mu}_i(t) + c_i(t) \geq \hat{\mu}_1(t) + c_1(t).\end{aligned}$$

Thus

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

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

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

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

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

- 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

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

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

  $$\begin{aligned}B_i(t) \text{ not happening conditioned on } A &\impliedby \mu_i + 2c_i(t) \leq \mu_1\\ &\iff \sqrt{\frac{\log(2/s_i(t))}{2n_i(t)}} \leq \frac{\Delta_i}{2}\\ &\iff n_i(t) \geq \frac{4\log(\sqrt{2}T)}{\Delta_i^2}\\ &\impliedby n_i(t) \geq \frac{6\log T}{\Delta_i^2}.\end{aligned}$$

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

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

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 very small. However, remember that if Δ_i is small, pulling i will only cause little regret, thus we fix some Δ and divide the k arms into two groups of $\Delta \leq \Delta_i$ and $\Delta > \Delta_i$, and calculate the correspond regret separately as follows:

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

by letting $\Delta = \sqrt{\frac{6k\log T}{T}}$.