CSE599s, Spring 2014, Online Learning Lecture 15 - 2/19/2014

Stochastic bandits: Explore-First and UCB

Lecturer: Brendan McMahan or Ofer Dekel *Scribe: Javad Hosseini*

In this lecture, we like to answer this question: what happens when everything is stochastic. We answer this question for the bandit setting. In some cases, stochastic world provides better performance. For examples, in online advertisement in MSN, we want to decide which news to show and may assume the world is stochastic.

1 The Bandit Game with Rewards

For historical reasons, we assume the stochastic bandit game with rewards, instead of loss. This is because in stochastic settings, people are more interested in rewards rather than loss. The bandit game with rewards is defined as:

for $t = 1, \ldots, T$

- player chooses arm $I_t \in 1, \ldots, d$
- player receives and observes a reward $R_{t,I_t} \in [0, 1]$, but does not observe $R_{t,i}$ for $i \neq I_t$

If the problem was naturally specified with losses $L_{t,i}$, we can define $R_{t,i} = 1 - L_{t,i}$

1.1 Stochastic Rewards

We assume there exists unknown distributions $\nu_1, \ldots, \nu_d$, each supported on $[0, 1]$, such that $R_{t,i} \sim \nu_i$ and $R_{1,i}, \ldots, R_{T,i}$ are independent. We also define the player's expected regret as:

$$\max_{i \in \{1 \ldots d\}} \mathbb{E}\Big[\sum_{t=1}^{T} R_{t,i}\Big] - \mathbb{E}\Big[\sum_{t=1}^{T} R_{t,I_t}\Big]. \tag{1}$$

We define some notations to use in this lecture:

- arm i's expected reward: $\mu_i = \mathbb{E}[R_{1,i}]$
- the best expected reward: $\mu^* = \max_{i \in \{1 \ldots d\}} \mu_i$
- the gap between arm i and the best arm: $\Delta_i = \mu^* - \mu_i$ [1]
- the number of times arm i is pulled up until round t:

$$\tau(t) = \sum_{s=1}^{t} \mathbf{1}_{I_s = i}. \tag{2}$$

Lemma 1. *The player's regret can be rewritten as*

$$T\mu^* - \sum_{i=1}^{d} \Delta_i \mathbb{E}[\tau_i(T)]. \tag{3}$$

[1] how much is the loss if we pull arm i

Proof.

$$\max_{i\in\{1\dots d\}} \mathbb{E}\Big[\sum_{t=1}^{T} R_{t,i}\Big] - \mathbb{E}\Big[\sum_{t=1}^{T} R_{t,I_t}\Big]$$

$$= T\mu^* - \sum_{t=1}^{T} \mathbb{E}\Big[\mathbb{E}[R_{t,I_t}|I_t]\Big] \text{(by def. of } \mu^*\text{; total expectation)}$$

$$= \sum_{t=1}^{T} \mathbb{E}[\mu^* - \mu I_t] \text{(by def. of } \mu_i\text{; linearity of expectation)}$$

$$= \sum_{i=1}^{d} \Delta_i \sum_{t=1}^{T} Pr(I_t = i) = \sum_{i=1}^{d} \Delta_i \mathbb{E}[\tau_i(T)] \tag{4}$$

□

Therefore, the only thing that matters is the number of times each arm was pulled. The order in which the arms were pulled does not matter. As a result, our goal is to upper-bound $\mathbb{E}[\tau_i(T)]$ for each suboptimal arm.

2 Explore First

Assume the player knows a lower bound on positive Δ_i's. Namely, he knows

$$\Delta = min\{\Delta_i : \Delta_i > 0\}. \tag{5}$$

The *Explore First* algorithm is:

- choose confidence level $\delta \in (0, 1]$
- sample each arm $C_{\Delta,\delta}$ times (in any order)
- compute the empirical mean of each arm's rewards $\hat{\mu}_i$
- find the maximizer $\hat{i} = \text{argmax}_{i\in\{1\dots d\}} \hat{\mu}_i$
- pull $\hat{i}$ for the remaining $T - dC$ rounds

2.1 Analysis of the Explore First Algorithm

We state a theorem without proof.

Theorem 2. *(Hoeffding-Azuma): Let $X_1, \dots, X_m$ be i.i.d. random variables supported on $[0, 1]$. Define $\mu = \mathbb{E}[X_1]$ and $\hat{\mu} = \frac{1}{m}\sum_{i=1}^{m} X_i$. Then, for any $\delta \in (0, 1]$, with probability at least $1 - \delta$, it holds that*

$$|\hat{\mu} - \mu| < \sqrt{\frac{\log(2/\delta)}{2m}}. \tag{6}$$

We apply the Hoeffding-Azuma theorem to the empirical mean of each arm. Using the union bound, the guarantee holds simultaneously for all arms. Namely, for any $\delta \in (0, 1]$, with probability at least $1 - d\delta$,

$$\forall i : |\hat{\mu}_i - \mu_i| < \sqrt{\frac{\log(2/\delta)}{2C_\Delta}}. \tag{7}$$

To find an arm with expected reward μ^*, we need

$$\forall i : |\hat{\mu}_i - \mu_i| < \frac{\Delta}{2}. \tag{8}$$

In Figure 1, you can see why this is true. The μ_i and $\hat{\mu}_i$ are actual and observed means. Since the actual and observed values differ at most $\frac{\Delta}{2}$, even the arm with second highest actual mean will not have higher observed mean than the the best arm. If we choose $C_{\Delta,\delta} = \frac{2\log(\delta/2)}{\Delta^2}$, then the difference between μ_i and $\hat{\mu}_i$ is at most $\frac{\Delta}{2}$. Therefore, with probability at least $1 - d\delta$, $\hat{\mu}_i = \mu^*$, and the regret is only due to the $dC_{\Delta,\delta}$ exploration rounds. As a result, we have

$$\text{Regret} \leq (1 - d\delta)\frac{2d\log(\frac{2}{\delta})}{\Delta^2} + d\delta T, \tag{9}$$

where the first term is for exploration cost and the second term will be applied if our estimates are off. Setting $\delta = \theta(1/T)$, gives a bound of $O(\frac{d\log T}{\Delta^2})$

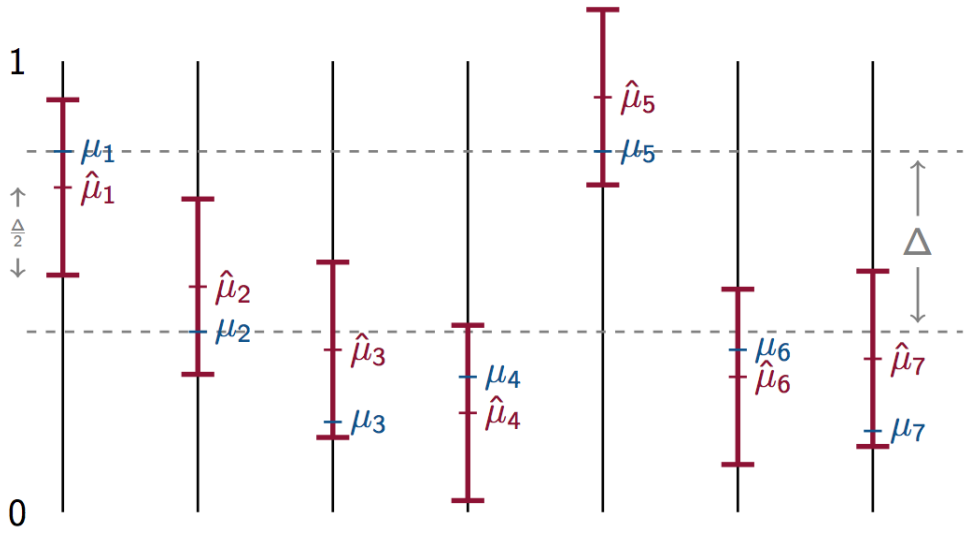

Figure 1: Google Adwords ads

2.2 Uncertainty and Optimism and Upper Confidence Bound Algorithm (UCB)

In the previous section, we assumed that we know $\Delta = min\{\Delta_i : \Delta_i > 0\}$. However, in general, this is not a realistic assumption. In order to deal with the problem in general, we use a *General Principle*: **"Optimism in the face of uncertainty"**.

Given the number of pulls so far, $\tau_1(t), ..., \tau_d(t)$, we compute a confidence interval for each arm and pretend that the true value is the best value in the interval.

The UCB algorithms works as follows: Define our estimate of arm i at time t as $\hat{\mu}_{t,i} = \frac{\sum_{s=1}^{t} R_{s,i}\mathbf{1}_{I_t=i}}{\tau_i(t)}$. Fix $\delta \in (0, 1]$ and define our upper confidence bound on arm i at time t as

$$U_{t,i} = \hat{\mu}_{t,i} + \sqrt{\frac{\alpha\log(t)}{\tau_i(t)}}. \tag{10}$$

α -UCB:

for $t = 1, \ldots, T$

- $I_t = \text{argmax}_{i \in \{1 \ldots d\}} U_{t-1,i}$.

The Intuition behind this algorithm is that every time we choose a suboptimal arm, we suffer regret, but our over-optimistic estimate decreases. This means we are learning. The UCB algorithm is deterministic, c.f. EXP3. In this problem, the world itself is random, so we do not want to add more randomization. In settings that the world is malicious, we may add randomization, though.

Note that, we have

$$\mathbb{E}[\Delta_{I_t} | I_t] = \mu^* - \mu_{I_t} \leq^{w.h.p.} U_{t,i^*} - \mu_{I_t} \leq^{argmax} U_{t,I_t} - \mu I_t. \tag{11}$$

Therefore, The regret on round t is bounded by the amount of optimism we apply to arm i.

Lemma 3. *Let i be such that $\Delta_i > 0$, then*

$$\mathbb{E}[\tau_i(T)] \leq \frac{2\alpha \log(T)}{\Delta_i^2} + \frac{2}{\alpha - 2}. \tag{12}$$

Proof Skecth: If $I_t = i$ then $U_{t,i} \geq U_{t,i^*}$. Therefore, at least one of the following events occur

$$A_1(t) = \{U_{t,i^*} \leq \mu^*\} \quad \text{or} \quad A_2(t) = \{U_{t,i} \geq \mu^* (= \mu_i + \Delta_i)\} \tag{13}$$

The event $A_1(t)$ can happen with probability at most δ.

Lemma 4.

$$\forall t : Pr(U_{t,i^*} < \mu^*) \leq t^{1-\alpha}. \tag{14}$$

Proof. Let $X_1, \ldots, X_t$ be i.i.d. random variables in $[0, 1]$, let $\mu = \mathbb{E}[X_1]$, and let $\hat{\mu}_\tau = \frac{1}{\tau} \sum_{j=1}^{\tau} X_j$.Then,

$$\begin{aligned} Pr\Big(\exists \tau \in \{1 \ldots t\} : |\mu - \hat{\mu}_\tau| > \sqrt{\frac{\alpha \log(t)}{\tau}}\Big) \\ \leq \sum_{\tau=1}^{t} Pr\Big(|\mu - \hat{\mu}_\tau| > \sqrt{\frac{\alpha \log(t)}{\tau}}\Big) \\ \leq \sum_{\tau=1}^{t} t^{-\alpha} = t^{1-\alpha}. \end{aligned} \tag{15}$$

□

Now we Define $u_i = \frac{2\alpha \log(T)}{\Delta_i^2}$ and prove $Pr\Big(A_2(t) | \tau_i(t) \geq \mu_i\Big) \leq t^{1-\alpha}$. Overall,

$$\begin{aligned} \mathbb{E}[\tau_i(T)] \leq u_i + \sum_{t : \tau_i(t) > u} Pr\Big(A_1(t) \vee A_2(t)\Big) \\ \leq u_i + 2 \sum_{t=1}^{\infty} t^{1-\alpha} \\ \leq u_i + \frac{2}{\alpha - 2}. \end{aligned} \tag{16}$$

We conclude that regret is upper-bounded by

$$\sum_{i : \Delta_i > 0} \Big(\frac{2\alpha \log(T)}{\Delta_i} + \frac{2\Delta_i}{\alpha - 2} \Big) = O\Big(\log(T) \sum_{i : \Delta_i > 0} \frac{1}{\Delta_i} \Big). \tag{17}$$

Therefore, as $\Delta_i \to 0$, the bound $\mathbb{E}[\tau_i(T)] = O(\frac{\log(T)}{\Delta_i^2})$ becomes vacuous. However, it always holds that $\mathbb{E}[\tau_i(T)] = O(T)$. Finally, we can use $min\{a, b\} \leq \sqrt{ab}$ to conclude that

$$min\Big\{\frac{\log(T)}{\Delta_i^2}, T\Big\} \leq \frac{1}{\Delta_i}\sqrt{T\log(T)}, \tag{18}$$

and get the *distribution free* bound

$$\text{Regret} = O(d)\sqrt{T\log(T)}. \tag{19}$$