CS 7545: Machine Learning Theory **Fall 2019**

Lecture 18: Stochastic Bandits & UCB

Lecturer: Jacob Abernethy *Scribes: V. R. Makkapati & H. Sarabu*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

18.1 Notation

- $\mathbb{E}[.]$: Expectation
- $\Pr[B]$: Probability of event B
- $\mathbb{1}\{B\}$: Indicator function for the event B

18.2 Stochastic Bandits

Setup

Consider the situation where there exists n probability distributions $D_1, \ldots, D_n$, and let μ_i be the mean of D_i, for all $i = 1, \ldots, n$. In the multi-arm bandit setting with n arms, in round t, arm i pays $X_i^t \overset{\text{iid}}{\sim} D_i$ (note: iid in rounds). Typically, it is assumed that a distribution D_i is 1-sub-Gaussian. The assumption is to ensure that the empirical and the true mean of the distributions are within certain bound with some probability.

Now, an algorithm for the multi-arm bandit problem selects $i_t \in \{1, \ldots, n\}$ at round t, while receiving a reward $X_{i_t}^t$, and can only observe that reward ($X_{i_t}^t$) alone. For the sake of analysis, without loss of generality, assume

$$\mu_1 = \max_{i \in \{1,\ldots,n\}} \mu_i, \tag{18.1}$$

and let $\Delta_i = \mu_1 - \mu_i$, $i = 2, \ldots, n$. Finally, assume that the value of $\Delta_* = \min_{i \in \{2,\ldots,n\}} \Delta_i$ is known.

A simple algorithm to minimize expected regret over T rounds, which is defined as

$$\mathbb{E}[\text{regret}_T] = \sum_{i=2}^{n} \Delta_i \mathbb{E}\left[N_i^{T+1}\right], \tag{18.2}$$

where $N_i^t = \sum_{s=1}^{t-1} \mathbb{1}\{i_s = i\}$, is discussed below. Note that N_i^t denotes the number of times arm i is pulled up till the t^{th} round.

Simple Algorithm

if $t \in [(i-1)k + 1,\ k]$ **then**
  $i_t = i$
end if
if $t > nk$ **then**
  $i_t = \arg\max_{i \in \{1,\ldots,n\}} \hat{\mu}_i$
end if

Here,

$$k = \left\lceil \frac{4\log(nT)}{\Delta_*^2} \right\rceil, \tag{18.3}$$

and

$$\hat{\mu}_i = \frac{1}{k} \sum_{t=(i-1)k+1}^{ik} X_i^t.$$

Theorem 18.1 *The bound on the expected regret for the simple algorithm over T rounds is given by*

$$\mathbb{E}\left[\text{regret}_T\right] \le \sum_{i=2}^{n} \left[\frac{4\Delta_i \log(nT)}{\Delta_*^2} + O(\Delta_i) \right]. \tag{18.4}$$

Proof: We prove the above statement by first finding a bound to $\mathbb{E}\left[N_i^{T+1}\right]$, for $i \neq 1$. Note that

$$\mathbb{E}\left[N_i^{T+1}\right] = \mathbb{E}\left[N_i^{T+1}\mathbb{1}\{B\} + N_i^{T+1}\mathbb{1}\{\bar{B}\}\right], \tag{18.5}$$

where $B = \{\hat{\mu}_1 > \hat{\mu}_i,\ \forall\, i = 2, \ldots, n\}$, and $\bar{B}$ represents the nonoccurence of event B. If event B takes places, then arm 1 will be chosen after kn rounds, then all the other arms would have been chosen atmost k times. Subsequently $\mathbb{E}\left[N_i^{T+1}\mathbb{1}\{B\}\right] \le k$. Clearly, the upper bound for the latter term can be $\mathbb{E}\left[N_i^{T+1}\mathbb{1}\{\bar{B}\}\right] \le T\Pr\left[\bar{B}\right]$.

We now obtain an upper bound for $\Pr\left[\bar{B}\right]$ by first realizing

$$\begin{aligned}
\Pr\left[\bar{B}\right] &= \Pr\left[\exists\, i \in \{2, \ldots, n\} : \hat{\mu}_i \ge \hat{\mu}_1\right] \\
&\le \sum_{i=2}^{n} \Pr\left[\hat{\mu}_i \ge \hat{\mu}_1\right] \\
&\le \sum_{i=2}^{n} \Pr\left[\hat{\mu}_i - \mu_i \ge \frac{\Delta_i}{2} \text{ or } \hat{\mu}_1 - \mu_1 \ge \frac{\Delta_i}{2}\right] \text{ (It can be understood from simple inspection)} \\
&\le \sum_{i=2}^{n} \left(\Pr\left[\hat{\mu}_i - \mu_i \ge \frac{\Delta_i}{2}\right] + \Pr\left[\hat{\mu}_1 - \mu_1 \ge \frac{\Delta_i}{2}\right] \right) \\
&\le \sum_{i=2}^{n} \left(\Pr\left[\hat{\mu}_i - \mu_i \ge \frac{\Delta_*}{2}\right] + \Pr\left[\hat{\mu}_1 - \mu_1 \ge \frac{\Delta_*}{2}\right] \right) \text{ (Because } \Delta_* \text{ is the minimum among } \Delta_i\text{s)} \\
&\le 2(n-1)\exp\left(-2k\frac{\Delta_*^2}{4}\right) \\
&\le 2(n-1)\exp\left(-2\frac{4\log(nT)}{\Delta_*^2}\frac{\Delta_*^2}{4}\right) \text{ (From (18.3))} \\
&= 2(n-1)\frac{1}{n^2T^2} \le \frac{1}{T}.
\end{aligned} \tag{18.6}$$

Therefore,

$$\mathbb{E}\left[N_i^{T+1}\right] \le \frac{4\log(nT)}{\Delta_*^2} + 1. \tag{18.7}$$

Finally,

$$\mathbb{E}\left[\text{regret}_T\right] \le \sum_{i=2}^{n} \Delta_i \mathbb{E}\left[N_i^{T+1}\right]. \tag{18.8}$$

Hence,

$$\mathbb{E}\left[\text{regret}_T\right] \leq \sum_{i=2}^{n} \left[\frac{4\Delta_i \log(nT)}{\Delta_*^2} + O(\Delta_i)\right]. \tag{18.9}$$

■

18.3 Upper Confidence Bound (UCB)

Consider the previous definition for $N_i^t = \sum_{s=1}^{t-1} \mathbb{1}\{i_s = i\}$, and the estimate of the mean for arm i up till the t^{th} round can be given by

$$\hat{\mu}_i = \sum_{s=1}^{t-1} \frac{X_i^t \mathbb{1}\{i_s = i\}}{N_i^t}. \tag{18.10}$$

The UCB algorithm chooses an arm at round t according to the relation

$$i_t = \underset{i \in \{1,\dots,n\}}{\arg\max}\; UCB_i^t, \tag{18.11}$$

where

$$UCB_i^t = \hat{\mu}_i + \sqrt{\frac{2\log(1/\delta)}{N_i^T}}. \tag{18.12}$$

Here $\sqrt{\frac{2\log(1/\delta)}{N_i^T}}$ is called the exploration bonus (optimism term). The term essentially increases the mean for an arm i if it has not been explored much, thus incentivizing exploration.

Theorem 18.2 *For $\delta = 1/t^2$, the bound on the expected regret for the UCB algorithm over T rounds is given by*

$$\mathbb{E}\left[\text{regret}_T\right] \leq \sum_{i=2}^{n} \left[\frac{16\log T}{\Delta_i} + O(\sum_{i=2}^{n} \Delta_i)\right]. \tag{18.13}$$

The proof for the above theorem requires showing that bad arms are not chosen that often. Essentially, it has to be shown that $N_i^t > k_i$, where $k_i = \left\lceil \frac{8\log(1/\delta)}{\Delta_i^2} \right\rceil$. Now, let $\mu_i^{(\hat{k}_i)} = \hat{\mu}_i^t$ when $N_i^t = k_i$ i.e., sample arm i enough times such that $\hat{\mu}_i^k$ is an empirical of k samples. We can define define a good scenario as

$$G_i = \{\mu_1 < UCB_1^t,\ \forall\ t = 1,\dots,T\} \cap \left\{ \mu_i^{(\hat{k}_i)} + \sqrt{\frac{2\log(1/\delta)}{k_i}} < \mu_1 \right\}. \tag{18.14}$$

Lemma 18.3 *If G_i is true, then it is guaranteed that $N_i^{T+1} < k_i$.*