CS 101.2: Notes for Lecture 2 (Bandit Problems)

Andreas Krause

January 9, 2009

In these notes we prove logarithmic regret for the UCB 1 algorithm (based on Auer et al, 2002).

1 Notation

- j: Index of slot machine arm (1 to k).
- n: Total number of plays we will make (known and specified in advance)
- t: Total number of plays we did so far
- $X_{j,t}$: Random variable for reward of arm j at time t. All $X_{j,t}$ are possibly continuous, but supported in the interval $[0, 1]$ (i.e., they do not take any values outside $[0, 1]$). All $X_{j,t}$ are independent.
- $T_j(t)$: Number of times arm j pulled during the first t plays. Note that $T_j(t)$ is a random quantity.
- $\mu_j = \mathbb{E}[X_{j,t}]$, and $\mu^* = \max_j \mu_j$
- $\Delta_j = \mu^* - \mu_j$, and $\Delta = \min_j \Delta_j$
- Expected regret after t plays:

$$R_t = \mathbb{E}\left[t\mu^* - \sum_j T_j(t)\mu_j\right] = \sum_j \mathbb{E}[T_j(t)]\Delta_j.$$

- $\bar{X}_j(t)$ is the sample average of all rewards obtained from arm j during the first t plays (i.e., if we've observed rewards $x_1, \ldots, x_m$ where $m = T_j(t)$, then $\bar{X}_j(t) = \frac{1}{m}(x_1 + \cdots + x_m)$).

2 The Upper Confidence Band algorithm (UCB1)

- Initially, play each arm once (hence $T_j(t) \geq 1$ for all $t \geq k$).
- Loop (for $t = k+1$ to n)
  - For each arm j compute "index"
    $$v_j = \bar{X}_j(t) + c_j(t),$$
    where $c_j(t) = \sqrt{\frac{\log n}{T_j(t)}}$.
  - Play the arm with $j^* = \operatorname{argmax}_j v_j$.

3 Analysis

Theorem 1. *If UCB_1 is run with input n, then its expected regret R_n is $O(\frac{K \log n}{\Delta})$.*

Proof. To prove Theorem 1, we will bound $\mathbb{E}[T_j(n)]$ for all arms j. Suppose, at some time t, UCB_1 pulls a suboptimal arm j. That means, that

$$\bar{X}_j(t) + c_j(t) \geq \bar{X}^*(t) + c^*(t).$$

Hence, in this case,

$$\begin{aligned} &\bar{X}_j(t) + 2c_j(t) - c_j(t) + (\mu_j - \mu_j) \geq \bar{X}^*(t) + c^*(t) + (\mu^* - \mu^*) \\ \Leftrightarrow\ &\underbrace{\bar{X}_j(t) - (\mu_j + c_j(t))}_{A} + \underbrace{(\mu_j - \mu^* + 2c_j(t))}_{B} \geq \underbrace{\bar{X}^*(t) - (\mu^* - c^*(t))}_{-C} \end{aligned}$$

We can see that at least one of A, B or C must be nonnegative, i.e., at least one of the following inequalities must hold:

$$\bar{X}_j(t) \geq \mu_j + c_j(t) \tag{1}$$
$$\bar{X}^*(t) \leq \mu^* - c^*(t) \tag{2}$$
$$\mu^* \geq \mu_j + 2c_j(t) \tag{3}$$

In order to bound the probability of (1) and (2), we use the Chernoff-Hoeffding inequality:

Fact 1 (Chernoff-Hoeffding inequality)**.** *Let $X_1, \ldots, X_n$ be independent random variables supported on $[0, 1]$, with $\mathbb{E}[X_i] = \mu$. Then, for every $a > 0$,*

$$P(\frac{1}{n}\sum_{i=1}^{n} X_i \geq \mu + a) \leq e^{-2a^2 n}$$

and

$$P(\frac{1}{n}\sum_{i=1}^{n} X_i < \mu - a) \leq e^{-2a^2 n}$$

□

Hence, we can bound the probability of (1) as

$$P(\bar{X}_j(t) \geq \mu_j + c_j(t)) \leq e^{-2c_j(t)^2 T_j(t)} = e^{-2\frac{\log n}{T_j(t)} T_j(t)} = e^{-2\log n} = n^{-2}.$$

Similarly,

$$P(\bar{X}^*(t) \leq \mu^* - c^*(t)) \leq n^{-2}.$$

Hence, (1) and (2) are very unlikely events. Now, note that whenever $T_j(t) \geq \ell = \lceil (4 \log n)/\Delta_j^2 \rceil$, (3) must be false, since

$$\mu_j + 2c_j(t) = \mu_j + 2\sqrt{\frac{\log n}{T_j(t)}} \leq \mu_j + 2\sqrt{\frac{\log n}{\frac{4\log n}{\Delta_j^2}}} \leq \mu_j + \Delta_j = \mu^*$$

Hence, if arm j has been played at least $\ell = O(\log_n / \Delta_j^2)$ times, then inequality (3) must be false, and hence arm j is pulled with probability at most $O(n^{-2})$.

Now we bound $\mathbb{E}[T_j(n)]$. By using conditional expectations, we have (writing T_j instead of $T_j(n)$ for short)

$$\mathbb{E}[T_j] = \underbrace{P(T_j \leq \ell)}_{\leq 1} \underbrace{\mathbb{E}[T_j \mid T_j \leq \ell]}_{\leq \ell} + \underbrace{P(T_j \geq \ell)}_{\leq 2n^{-2}} \underbrace{\mathbb{E}[T_j \mid T_j \geq \ell]}_{\leq n} \leq \ell + 2n^{-1}$$

since we have

$$P(T_j \geq \ell) \leq P(\text{inequality (1) or (2) violated }) \leq 2n^{-2}.$$

□

4 Some additional remarks

Note that as stated in Section 2, the total number of plays n needs to specified in advance. By setting

$$c_t = \sqrt{\frac{2 \log t}{T_j(t)}},$$

we can avoid this issue. A slightly more complex analysis (of Auer et al '02) shows that in this case after any number of t plays it holds that

$$R_t = O(\frac{k \log t}{\Delta}).$$