Bandit Problem and UCB

Subhojyoti Mukherjee

IIT Madras

February 2, 2017

Overview

Bandit Problem

- In stochastic multi-armed bandit problem we are presented with a set of arms or choices.
- The rewards for each of the arms is drawn from identical and independent distributions.
- The learner does not know the mean of the distributions, denoted by μ_i.
- The learner has to find the optimal arm the mean of whose distribution is denoted by μ^* such that $\mu^* > \mu_i \forall i \in A$

Exploration Exploitation Dilemma

- The bandit problem is a sequential decision making process where at each timestep we have to choose one arm from a set of arms.
- After say pulling each arm once we are presented with an exploitation-exploration problem, that is whether to continue to pull the arm for which we have observed the highest estimated reward till now(exploitation) or to explore a new arm(exploration).
- If we become too greedy and always exploit we may miss the chance of actually finding the optimal arm and get stuck with a sub-optimal arm.

Bandit Algorithm

- Goal: To minimize Regret
- Average reward of best action is μ^* and any other action i as μ_i. There are K total actions. $n_i(t)$ is number of times tried action i is executed till t-timesteps.
- Cumulative Regret: The loss we suffer because of not pulling the optimal arm till the total number of timesteps T.

$$R_T = r^* T - \sum_{i \in A} r_i n_i(T),$$

- The expected regret of an algorithm after T rounds can be written as

$$\mathbb{E}[R_T] = \sum_{i=1}^{K} \mathbb{E}[n_i(T)]\Delta_i,$$

- $\Delta_i = r^* - r_i$ denotes the gap between the means of the optimal arm and of the i-th arm.

Concentration Bounds

- The issue of coin tossing.
- Chernoff-Hoeffding Bounds and its applications.
- Let $X_1, ..., X_n$ be random variables with common range [0, 1] and such that $E[X_t | X_1, ..., X_{t-1}] = \mu$. Let $Sn = X_1+, ..., +X_n$. Then for all $a \geq 0$
  $P\{S_n \geq n\mu + a\} \leq e^{-2a^2/n}$ and $P\{S_n \leq n\mu - a\} \leq e^{-2a^2/n}$

Algorithm 1 UCB1

1: Pull each arm once

2: **for** $t = K + 1, ..., T$ **do**

3: Pull the arm such that $\max_{i \in A} \left\{ \hat{\mu} + \sqrt{\frac{2 \log t}{n_i}} \right\}$

4: **end for**

Auer et al. [2002]

UCB 1 Theorem on Regret Bound

Theorem

For all $K > 1$, if policy UCB1 is run on K machines having arbitrary reward distributions $P_1, ..., P_K$ with support in [0, 1], then its expected regret after any number n of plays is at most

$$\left[8 \sum_{i:\mu_i<\mu^*} \left(\frac{\ln n}{\Delta_i}\right)\right] + \left(1 + \frac{\pi^2}{3}\right)\left(\sum_{j=1}^{K} \Delta_j\right)$$

where $\mu_1, ..., \mu_K$ are the expected values of $P_1, ..., P_K$.

UCB 1 Proof

$$T_i(n) = 1 + \sum_{t=K+1}^{n} \{I_t = i\}$$

Initially each arm has been pulled once. So from K+1 th attempt till n we are trying to bound $T_i(n)$

$$\leq \ell + \sum_{t=K+1}^{n} \{I_t = i,\ T_i(t-1) \geq \ell\}$$

We have played arm i at least ℓ number of times till t-1 th time indicated by $T_i(t-1)$. This means that the arm i pulled till time n will be less than equal to some arbitrary positive integer ℓ

$$\leq \ell + \sum_{t=K+1}^{n} \Big\{\bar{X}^*_{T^*(t-1)} + c_{t-1,T^*(t-1)} \leq \bar{X}_{i,T_i(t-1)} + c_{t-1,T_i(t-1)},\ T_i(t-1) \geq \ell\Big\}$$

Since we are pulling this arm i again this means that the UC of optimal arm(l.h.s) is less than the UC of the i th arm(r.h.s) given that the number of times the arm i is pulled till time t-1 is atleast greater than ℓ.

The Upper confidence bound is given by the mean of the reward and the confidence interval term of that arm. This confidence interval term c_t we will derive later by the Chernoff-Hoeffding bound.

$$\leq \ell + \sum_{t=K+1}^{n} \left\{ \min_{0<s<t} \bar{X}_s^* + c_{t-1,s} \leq \max_{\ell \leq s_i < t} \bar{X}_{i,s_i} + c_{t-1,s_i} \right\}$$

For atleast once the minimum of the UC of the optimal arm from 0 - t th time is less than the maximum of the UC of the i th arm from ℓ to t th time. Here the number of times the optimal arm is pulled is denoted by s and for the ith arm denoted by s_i

$$\leq \ell + \sum_{t=1}^{\infty} \sum_{s=1}^{t-1} \sum_{s_i=\ell}^{t-1} \{ \bar{X}_s^* + c_{t,s} \leq \bar{X}_{i,s_i} + c_{t,s_i} \}. \tag{6}$$

Summing over all pulls from the start.

UCB 1 Proof

Now observe that $\bar{X}_s^* + c_{t,s} \leq \bar{X}_{i,s_i} + c_{t,s_i}$ implies that at least one of the following must hold

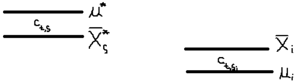
$$\bar{X}_s^* \leq \mu^* - c_{t,s} \tag{7}$$

$$\bar{X}_{i,s_i} \geq \mu_i + c_{t,s_i} \tag{8}$$

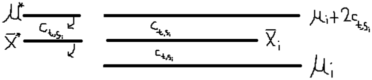
$$\mu^* < \mu_i + 2c_{t,s_i}. \tag{9}$$

UCB 1 Proof

These three conditions are shown below



This is the (7) and (8) condition. But this still does not make the upper confidence of the optimal term less than the i-th arm. Hence to make that possoble we have the condition (9).



The less than equal to sign has been shown by the downward arrows

The distance between the two levels(horizontal lines) is shown by the c term, which is the confidence interval

UCB 1 Proof

From the (9) condition we can definitely show that $\overline{X}^*_s$ is less than $\overline{X}_i$.

Next, we bound the probability of events (7) and (8) using Fact 1(Chernoff-Hoeffding bound)
For (7) $\mathbb{P}\{\overline{X}^*_s \leq \mu^* - c_{t,s}\} \leq e^{-2c^2_{t,s}/s} \leq e^{-4lnt} = t^{-4}$ by putting $c_{t,s} = \sqrt{2lnt/s}$
For (8) $\mathbb{P}\{\overline{X}_{i,s_i} \geq \mu_i + c_{t,s_i}\} \leq e^{-2c^2_{t,s_i}/s_i} \leq e^{-4lnt} = t^{-4}$ by putting $c_{t,s_i} = \sqrt{2lnt/s_i}$

Now, for $\ell = \lceil (8lnn)/\Delta_i^2 \rceil$ (9) is false. This can be proved by putting this ℓ value in
$\mu^* - \mu_i - 2c_{t,s_i} = \mu^* - \mu_i - 2\sqrt{2lnt/s_i} = \mu^* - \mu_i - 2\sqrt{2lnt\Delta_i^2/(8lnt)} = \mu^* - \mu_i - \Delta_i = \mu_* - \mu_i - \mu_* + \mu_i = 0$
by putting $\Delta_i = \mu_* - \mu_i$ above.
Next, in (6) we put the value of ℓ and the conditions which satisfies the necessary upper confidence bounds.

UCB 1 Proof

$$\mathbb{E}[T_i(n)] \le \left\lceil \frac{8 \ln n}{\Delta_i^2} \right\rceil + \sum_{t=1}^{\infty} \sum_{s=1}^{t-1} \sum_{s_i=\lceil (8\ln n)/\Delta_i^2 \rceil}^{t-1} \times \left(\mathbb{P}\{\bar{X}_s^* \le \mu^* - c_{t,s}\} + \mathbb{P}\left\{\bar{X}_{i,s_i} \ge \mu_i + c_{t,s_i}\right\}\right)$$

$$\le \left\lceil \frac{8 \ln n}{\Delta_i^2} \right\rceil + \sum_{t=1}^{\infty} \sum_{s=1}^{t} \sum_{s_i=1}^{t} 2t^{-4}$$

The inequality comes because we are summing over all all instances from s=1 to t-1 rather than only from ℓ

$\le \lceil \frac{8lnn}{\Delta_i^2} \rceil + \Sigma_{t=1}^{\infty} 2t^{-4} \Sigma_{s=1}^{t} \Sigma_{s_i=1}^{t}(1)$

$\le \frac{8lnn}{\Delta_i^2} + 1 + \Sigma_{t=1}^{\infty} 2t^{-4} t^{-2}$ [Removing the ceiling]

$\le \frac{8lnn}{\Delta_i^2} + 1 + \Sigma_{t=1}^{\infty} 2t^{-2}$

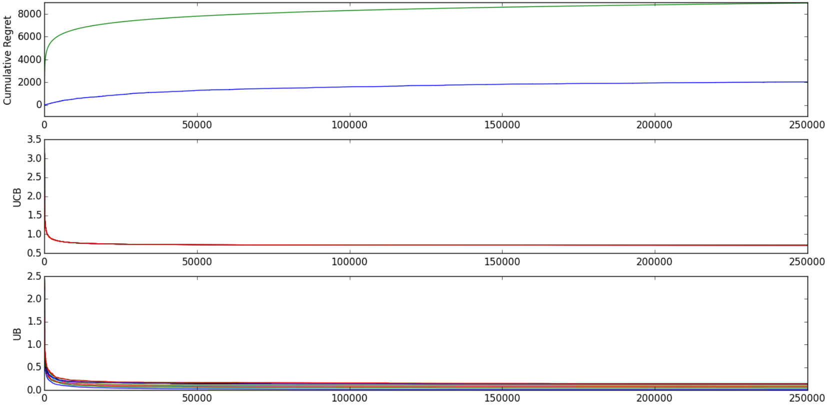
$\le \frac{8lnn}{\Delta_i^2} + 1 + 2 \times \frac{\pi^2}{6}$ from Basel's Equation

UCB 1 Proof

$$\leq \frac{8 \ln n}{\Delta_i^2} + 1 + \frac{\pi^2}{3}$$

This concludes the proof.

UCB 1 Experimental Run



Some Other Bandits and Applications

- Adversarial Bandits : Used in Investment in Stock Markets
- Contextual Bandits : Used in online Advertisement/news article selection
- Budgeted Bandits : Used in Clinical trials
- Distributed Bandits : Used in packet routing through a network

References

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

Thank You