EE675A - IITK, 2022-23-II

Lecture 3: Adaptive Exploration Algorithms

16th January 2023

Lecturer: Subrahmanya Swamy Peruru *Scribe: N Bhuvan, Lakshay Jangu*

This is the third lecture of EE675A - Introduction to Reinforcement Learning. In the previous lectures, we have seen the formulation of the 'Bandit Problem' and a few algorithms to solve it. In this lecture, we will see the shortcomings of the previous approaches and will explore an algorithm with a better upper bound on the regret.

1 Recap

The Bandit problem that we are trying to solve is as follows - Given K arms with unknown rewards r_i and T rounds, we wish to maximise our reward in T rounds. As we have no prior information, we must do some exploration to figure out the rewards. Following this line of thought, we proposed some algorithms in previous classes -

1.1 Explore Then Commit

Algorithm 1 : Explore Then Commit

1. Explore each arm N times
2. Select arm $\hat{a}$ with highest sample average reward r_i
3. Play arm $\hat{a}$ in all remaining trials

The upper bound on expected regret of this approach is $T^{2/3} * O(K \log(T))^{1/3}$. A major flaw in this algorithm is that if the difference in rewards of different arms is big, then we are making a very bad choice in exploration. To work on the shortcomings of this algorithm, we suggested a different algorithm named ϵ-greedy.

1.2 ϵ-greedy

The upper bound on expected regret for this approach is $t^{2/3} * O(K \log(t))^{1/3}$ if the exploration probability $\epsilon_t = t^{-1/3} * (K \log(t))^{1/3}$. The flaw in the ϵ-greedy algorithm is that we are not taking into account all the exploration we did to decide which random arm to pull. An arm with less sample average reward is as equally likely to be pulled as the best sample average reward arm. This 'non-adaptive' property of ϵ-greedy is its major flaw.

Algorithm 2 : Epsilon-greedy

for t = 1 to T **do**
  Toss a coin which lands on heads with probability ϵ_t
  if Heads **then**
    Choose a random arm
  else if Tails **then**
    Choose the arm with the highest sample average reward
  end if
  $t \leftarrow t + 1$
end for

2 Adaptive Exploration

In the previous algorithms, we were not eliminating bad arms soon enough. In this section, we introduce a new algorithm named 'Successive Elimination', which phases out bad arms much more quickly and has a better upper bound on regret. We use upper and lower confidence bounds on sample averages to decide which arm to eliminate.

Algorithm 3 : Successive Elimination

Declare all arms as active
while t $\neq$ T **do**
  Sample all the active arms once ($t \leftarrow t + 1$ for every active arm)
  Compute UCB and LCB of all arms
  For all arms a if there exists an arm b such that UCB(a) $\leq$ LCB(b) then deactivate arm a
end while

Intuitively this makes sense because we are eliminating an arm as long as there is an arm which has better sample average rewards according to our confidence bounds. In Figure 1, we can see that UCB(a_2) is less than LCB(a_3). Therefore arm a_2 will be deactivated according to the Successive Elimination Algorithm's deactivation rule.

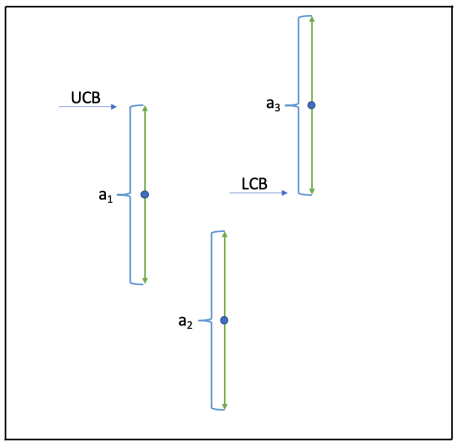

Figure 1: Confidence bounds of rewards of three arms

2.1 Computing Confidence Interval

We have to compute the UCB and LCB of all the active arms during each loop in order to deactivate arms which have UCB less than the LCB of any other active arm. The confidence interval of an arm **a** at time **t** derived using Hoeffding Inequality is given by:

Confidence interval: $[\bar{\mu}_t(a) - \epsilon_t, \bar{\mu}_t(a) + \epsilon_t]$

$\bar{\mu}_t(a)$: average reward attained by sampling arm **a**

ϵ_t : confidence radius

$$\epsilon_t = \sqrt{\frac{2.\log T}{n_t(a)}}$$

$n_t(a)$: no of times arm **a** has been sampled till time **t**

We can observe that the confidence interval bounds come closer as $n_t(a)$ increases.

2.2 Regret Analysis

Let's first do the regret analysis for K = 2 arms and then follow up for general K.

2.2.1 Regret Analysis for K = 2

Algorithm 4 : Successive Elimination for K = 2

Sample two arms until there exist an arm **a** s.t $\text{UCB}_t(\text{a}) \leq \text{LCB}_t(\text{b})$ after some even round **t**.
Sample only arm **b** for the rest of the time **T-t**.

Here we use the successive elimination algorithm, which reduces the regret by eliminating the arms with UCB less than any other arm's LCB.

Let t be the round where the confidence interval of arms a and b overlap on each other for the last time. At t+1 round, we are going to eliminate arm a according to the deactivation rule because for arm a, there exists arm b where $\text{UCB}_{t+1}(\text{a}) < \text{LCB}_{t+1}(\text{b})$. A detailed representation of deactivation is shown in figure 2. Because we were sampling both arms until the deactivation of arm a, $n_t(a) \approx t/2$ and $n_t(b) \approx t/2$.

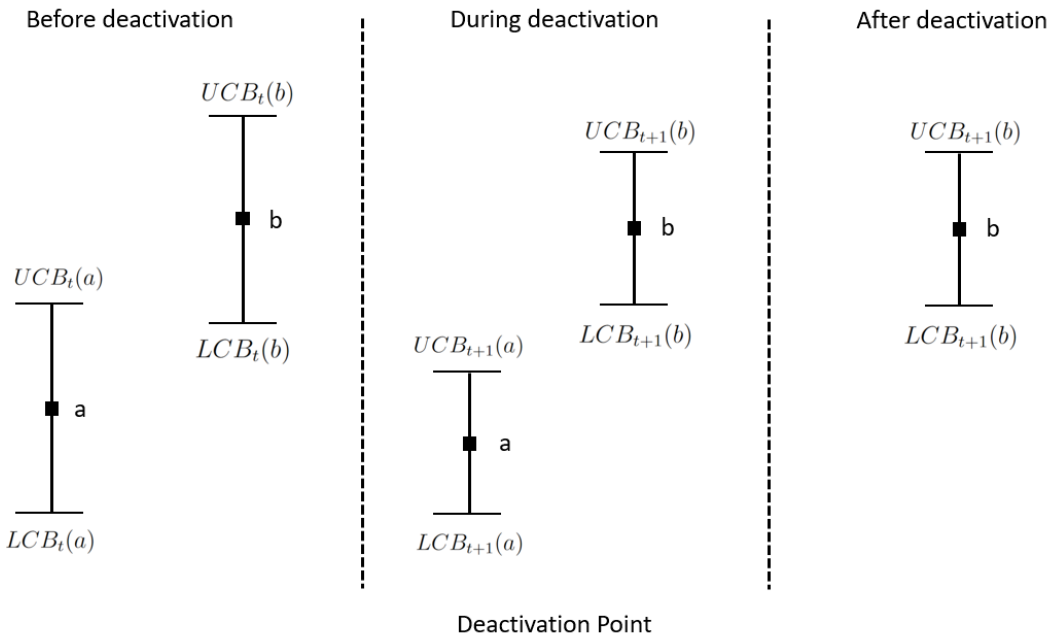

Figure 2: Confidence intervals of two arms during deactivation

We now derive the total regret accumulated because of sampling arm a using confidence intervals of both arms and use the big-O notation of attained regret to show its asymptotic dependence on the parameters of interest. Just before the elimination, the confidence interval of both arms overlap with each other; therefore, $UCB_t(a) > LCB_t(b)$. Using this equation, we can get the upper bound on the difference between $\bar{\mu}_t(b)$ and $\bar{\mu}_t(a)$.

$$\begin{aligned} UCB_t(a) &> LCB_t(b) \\ \bar{\mu}_t(a) + \epsilon_t &> \bar{\mu}_t(b) - \epsilon_t \\ \bar{\mu}_t(b) - \bar{\mu}_t(a) &< 2\epsilon_t \end{aligned}$$

Regret accumulated because of sampling arm a before its deactivation is R(t).

$$\begin{aligned} R(t;a) &= n_t(a) * (\mu(b) - \mu(a)) \\ n_t(a) &= t/2 \end{aligned}$$

We can get the upper bound of $\mu(b) - \mu(a)$ using the confidence interval we derived at time t.

$$\begin{aligned}
\mu(b) - \mu(a) &\leq (\bar{\mu}_t(b) + \epsilon_t) - (\bar{\mu}_t(a) - \epsilon_t) \\
\mu(b) - \mu(a) &\leq \bar{\mu}_t(b) - \bar{\mu}_t(a) + 2\epsilon_t \\
\mu(b) - \mu(a) &\leq 4\epsilon_t \\
\mu(b) - \mu(a) &\leq 4\sqrt{\frac{2.\log T}{n_t(a)}}
\end{aligned}$$

We used the notation of $\Delta(a)$ for defining the difference between the mean of optimal and sub-optimal arms. Here, the optimal arm is b, and the sub-optimal arm is a. Therefore,

$$\begin{aligned}
\Delta(a) &= \mu(b) - \mu(a) \\
\Delta(a) &\leq 4\epsilon_t
\end{aligned}$$

Total regret accumulated till round t is

$$\begin{aligned}
R(t;a) &= n_t(a) * (\mu(b) - \mu(a)) \\
R(t;a) &\leq n_t(a) * 4 * \epsilon_t \\
R(t;a) &\leq n_t(a) * 4 * \sqrt{\frac{2.\log T}{n_t(a)}} \\
R(t;a) &\leq O\big(\sqrt{n_t(a).logT}\big) \\
R(t;a) &\leq O\big(\sqrt{t.logT}\big)
\end{aligned}$$

$R(t;a) \leq O\big(\sqrt{n_t(a).logT}\big)$ is a general equation which can be derived for any number of K arms.

$$\begin{aligned}
R(t) &= \sum_{a \epsilon A} R(t;a) \\
R(t) &\leq \sum_{a \epsilon A} O\big(\sqrt{n_t(a) logT}\big) \\
R(t) &\leq O\big(\sqrt{logT}\big) \sum_{a \epsilon A} \sqrt{n_t(a)}
\end{aligned}$$

Using Jensen's inequality on $\sum_{a \epsilon A} \sqrt{n_t(a)}$ we get,

$$\frac{1}{K} \sum_{a \epsilon A} \sqrt{n_t(a)} \leq \sqrt{\frac{1}{k} \sum_{a \epsilon A} n_t(a)} \leq \sqrt{\frac{t}{K}}$$

We can now use this inequality to get the upper bound on R(t).

$$R(t) \leq O\big(\sqrt{logT}\big) \sum_{a \epsilon A} \sqrt{n_t(a)}$$
$$R(t) \leq O\big(\sqrt{logT}\big)\sqrt{Kt}$$
$$R(t) \leq O\big(\sqrt{Kt \log T}\big)$$

Successive Elimination Algorithm achieves regret

$$E(R(t)) = O\big(\sqrt{Kt \log T}\big) \quad \text{for all rounds of } \; t \leq T$$

2.3 Derivation of Instance Dependent bound

We will now derive a bound on the total regret accumulated using the bound we found for $\Delta(a)$ and also rearranging it to get a bound on $n_t(a)$.

$$\Delta(a) \leq O\left(\sqrt{\frac{2. \log T}{n_t(a)}}\right)$$

This inequality on $\Delta(a)$ can be interpreted to explain the relation between $\Delta(a)$ and $n_t(a)$. An arm played too many times is not a bad arm because the limit on $\Delta(a)$ decreases as $n_t(a)$ increases. Rearranging this equation, we will get,

$$n_t(a) \leq O\left(\frac{\log T}{(\Delta(a))^2}\right)$$

This inequality on $n_a(a)$ can be interpreted to explain that an arm won't be played if it is a bad arm. The bound on total regret ($R(T)$)can be derived using the above inequality,

$$R(T; a) \leq \Delta(a) * n_t(a)$$
$$R(T; a) \leq \Delta(a) * O\left(\frac{\log T}{(\Delta(a))^2}\right)$$
$$R(T; a) \leq O\left(\frac{\log T}{\Delta(a)}\right)$$

Regret attained is as follows,

$$R(T) \leq O(\log T)\left[\sum_{\text{arms a with } \mu(a) < \mu(a^*)} \frac{1}{\mu(a^*) - \mu(a)}\right]$$

This regret bound's constant (in square brackets) depends on the problem instance, has a maximum limit of $O(\frac{K}{\Delta})$, where

$$\Delta = \min_{\text{sub optimal arms a}} \mu(a^*) - \mu(a) = \Delta(a)$$

A bound which has a format of $C * f(T)$ is said to be instance dependent if C depends on the mean reward μ, whereas it is instance independent if C doesn't depend on the mean reward μ.

The Regret attained using Δ as follows,

$$R(T) \leq O\left(\frac{K \log T}{\Delta}\right)$$

The above derivation of $R(T)$ is known as the **instance dependent bound** derivation.

2.4 Derivation of Instance Independent bound

We can derive the instance independent bound from the instance dependent bound, the derivation is as follows,

$$R(T, a) \leq O\left(\frac{\log T}{\Delta(a)}\right)$$
$$R(T) = \sum_{a \epsilon A} R(T, a)$$

Let ϵ be an arbitrary number greater than zero, we are going to use this ϵ to derive an instance independent bound. The regret can be divided into two parts

- For arms a with $\Delta(a) < \epsilon$, the arm's regret contribution is ϵT.
- For arms a with $\Delta(a) > \epsilon$, the arm's regret contribution is $O\left(\frac{\log T}{\epsilon}\right)$. We substituted $\Delta(a)$ with ϵ to get a better upper bound.

The regret when we combine both parts is,

$$R(T) = \sum_{a \epsilon A} R(T, a)$$
$$R(T) = \sum_{a|\Delta(a)<\epsilon} R(T, a) + \sum_{a|\Delta(a)>\epsilon} R(T, a)$$
$$R(T) \leq \epsilon * T + O\left(\frac{K * \log T}{\epsilon}\right)$$
$$R(T) \leq O\left(\epsilon * T + \frac{K * \log T}{\epsilon}\right)$$

As ϵ is an arbitrary positive number, we can minimize the RHS of the limit. By minimizing the RHS, we get

$$\epsilon = \epsilon^{'} = \sqrt{\frac{K \log T}{T}}$$

Substituting the $\epsilon^{'}$ into the RHS we get,

$$R(T) \leq O\left(\sqrt{KT \log T}\right)$$

This derivation of the bound has a constant which doesn't depend on the μ, therefore, this is an **instance independent bound** derivation.

Explore-first	ϵ-greedy	Successive elimination
$T^{2/3} * O(K \log(T))^{1/3}$	$t^{2/3} * O(K \log(t))^{1/3}$	$O(\sqrt{KT \log T})$

Table 1: Regret bounds for three algorithms

References:

[1] A. Slivkins. Introduction to multi-armed bandits, Sec 1.3, 2019.

[2] R. S. Sutton and A. G. Barto. Reinforcement Learning: An Introduction. The MIT Press, second edition, 2018.