EE675A : Introduction to Reinforcement Learning

Lecture 03: Adaptive exploration-based algorithms

16-01-2023

Lecturer: Subrahmanya Swamy Peruru *Scribes: Amritanshu Manu, Anupam Kumar Yadav*

In the previous lecture, we derived regret bounds for *explore-first* and *epsilon-greedy* algorithms. These algorithms do not adapt their exploration schedule to the history of the observed rewards. We refer to this property as non-adaptive exploration. In this lecture, we will obtain much better regret bounds by adapting exploration to the observations.

1 Successive Elimination algorithm

We need to eliminate the under-performing arms sooner to achieve better regret bounds. The successive elimination algorithm attempts to sample each arm a minimal number of times and eliminate the arms one after the other. The eliminated arms are never sampled again.

1.1 Outline of the algorithm

Let there be four arms a, b, c, and d (see Fig. 1). The true mean reward and the sample mean reward of an arm a is given by $\mu(a)$ and $\bar{\mu}(a)$, respectively. Let $n_t(a)$ be the number of rounds before t in which arm a is chosen, and $\mu_t(a)$ be the average reward in these rounds. Using Hoeffding inequality, we can write the following:

$$\mu(a) \in [\bar{\mu_t}(a) - \epsilon, \bar{\mu_t}(a) + \epsilon], \tag{1}$$

where $\epsilon = \sqrt{\frac{2 \log T}{n_t(a)}}$. For each arm a at round t, we define *upper* and *lower confidence bounds*,

$$\begin{aligned} \texttt{UCB}_t(a) &= \bar{\mu_t}(a) + \epsilon, \\ \texttt{LCB}_t(a) &= \bar{\mu_t}(a) - \epsilon. \end{aligned} \tag{2}$$

The idea is to eliminate an arm if its best possible reward is less than the worst reward of any other arm. For instance, in Fig. 1, arms a and c will be eliminated as $\texttt{UCB}_t(c)$ and $\texttt{UCB}_t(a)$ are less than $\texttt{LCB}_t(d)$. So, the successive elimination algorithm can be given as follows:

1. Declare all arms as active (arms not yet eliminated)
2. Sample all the active arms once
3. Compute lower and upper confidence bounds
4. For all active arms, eliminate an arm a if $\texttt{UCB}_t(a) \leq \texttt{LCB}_t(a')$ for some arm a'
5. Go to Step 2 and repeat. Stop when only one arm survives.

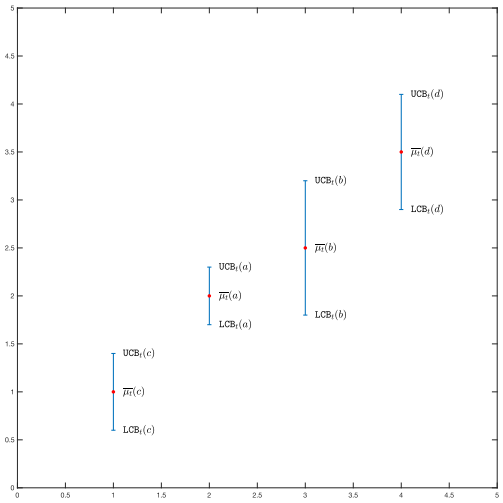

Figure 1: Elimination of arms based on upper and lower confidence bounds

1.2 Regret analysis

We will derive the regret bound for the case of $K = 2$ arms, namely a and b. We keep alternating between the arms until $\texttt{UCB}_t(a) \leq \texttt{LCB}_t(b)$. This is the exploration phase. Once arm a is eliminated, we keep using arm b. This is the exploitation phase. The regret in the exploitation phase is zero since we are using the best arm. We have to find the cumulative regret for the exploration phase. Let t be the last round when we did not invoke the stopping rule, i.e., when the confidence intervals of the two arms still overlap. Since the algorithm has been alternating the two arms before time t, we have $n_t(a) = n_t(b) = \frac{t}{2}$, and $\epsilon_t = \sqrt{2 \log T / \lfloor t/2 \rfloor}$. Then

$$\bar{\mu}_t(b) - \bar{\mu}_t(a) < \epsilon_t(a) + \epsilon_t(b) = 2\epsilon_t. \tag{3}$$

We have to bound $\mu(b) - \mu(a)$ as it is the difference between the optimal and the sub-optimal arm.

$$\begin{aligned} \mu(b) - \mu(a) &\leq [\bar{\mu}_t(b) + \epsilon_t] - [\bar{\mu}_t(a) - \epsilon_t] \\ &= \bar{\mu}_t(b) - \bar{\mu}_t(a) + 2\epsilon_t \\ &\leq 2\epsilon_t + 2\epsilon_t \\ &= 4\sqrt{2 \log T / \lfloor t/2 \rfloor} \\ &= \mathcal{O}(\sqrt{\log T / t}). \end{aligned} \tag{4}$$

Then the total regret accumulated till round t is

$$R(t) \leq (\mu(b) - \mu(a)) \cdot t \leq \mathcal{O}\left(t \cdot \sqrt{\frac{\log T}{t}} \right) = \mathcal{O}(\sqrt{t \log T}). \tag{5}$$

The $\sqrt{t}$ dependence in this regret bound is an improvement over the $T^{2/3}$ dependence for explore-first. This is possible due to adaptive exploration.

2 Instance-dependent bound

We can get even better bounds on the expected regret in the successive elimination algorithm in case the minimal gap (the amount by which the true mean of the best arm leads the true mean of a sub-optimal arm) among all arms is known. We can try to eliminate it more efficiently by exploiting this information. Such a bound on the regret is called the *instance-dependent bound*.

2.1 Deriving the instance-dependent bound

Let a^* be an optimal arm, and it cannot be deactivated. Choose any arm a such that $\mu(a) < \mu(a^*)$. Consider the last round $t \leq T$ when arm a remained active. As in the argument for $K = 2$ arms, the confidence intervals of a and a^* must overlap at round t. Therefore, using (4)

$$\Delta(a) \coloneqq \mu(a^*) - \mu(a) \leq 4\epsilon_t. \tag{6}$$

By the choice of t, arm a can be played at most once afterwards: $n_T(a) \leq 1 + n_t(a)$. Thus, we arrive at the following property:

$$\Delta(a) \leq \mathcal{O}(\epsilon_T) = \mathcal{O}(\sqrt{\log T / n_T(a)}) \quad \text{for each arm } a \text{ with } \mu(a) < \mu(a^*). \tag{7}$$

Equation (7) signifies that if an arm is played many times, it is not bad. By rewriting (7) as

$$n_T(a) \leq \mathcal{O}(\log T / [\Delta(a)]^2), \tag{8}$$

we can restructure our claim: if an arm is bad, it won't be played too often. Hence we'll reach closer to the optimal reward. The expected total regret contributed by the arm a, $R(T; a)$, is

$$\begin{aligned} R(T; a) &= \Delta(a) \cdot n_T(a) \\ &\leq \Delta(a) \cdot \mathcal{O}(\log T / [\Delta(a)]^2) \\ &\leq \mathcal{O}\left(\frac{\log T}{\Delta(a)}\right). \end{aligned} \tag{9}$$

The total regret, $R(T)$, can be calculated as the sum of regrets obtained for the set of all arms, $\mathcal{A}$.

$$\begin{aligned} R(T) &= \sum_{a \in \mathcal{A}} R(T; a) \\ &= \sum_{a \in \mathcal{A}} \mathcal{O}(\log T / \Delta(a)) \\ &\leq \mathcal{O}(\log T) \sum_{a \in \mathcal{A}} \frac{1}{\Delta(a)} \\ &\leq \mathcal{O}\left(\frac{K \log T}{\Delta}\right), \end{aligned} \tag{10}$$

where, $\Delta := \min_a \Delta(a)$ is the minimal gap. This bound is significant only when Δ is large enough. For smaller gaps, the bound turns out to be quite large and is useless. For such cases, *instance-independent bound* is a better choice. We can derive the instance-independent bound for the general case, i.e., for any number of arms, using the instance-dependent bound.

2.2 Deriving the general (instance-independent) bound

From (9) and (10), we get

$$\begin{aligned} R(T;a) &\leq \mathcal{O}\left(\frac{\log T}{\Delta(a)}\right) \\ R(T) &= \sum_{a \in \mathcal{A}} R(T;a). \end{aligned} \tag{11}$$

We'll bound the gap, $\Delta(a)$, by some confidence interval ϵ and analyze the regret for two cases.

$Case\ (i):$ if $\Delta(a) < \epsilon$

$$R(T;a) \leq \epsilon\, n_T(a)$$

$Case\ (ii):$ if $\Delta(a) > \epsilon$

$$R(T;a) \leq \mathcal{O}\left(\frac{\log T}{\Delta(a)}\right) \leq \mathcal{O}\left(\frac{\log T}{\epsilon}\right)$$

Consider,

$$\begin{aligned} R(T) &= \sum_{a \in A} R(T;a) \\ &= \sum_{\{a|\Delta(a)<\epsilon\}} R(T;a) \; + \sum_{\{a|\Delta(a)>\epsilon\}} R(T;a) \\ &\leq \mathcal{O}(\epsilon \cdot T + \frac{K}{\epsilon} \cdot \log T). \end{aligned} \tag{12}$$

Since this holds for any $\epsilon > 0$, we can choose the value of ϵ that minimizes the right-hand side. Ensuring that $\epsilon T = \frac{K}{\epsilon} \log T$ yields $\epsilon = \sqrt{\frac{K}{T} \log T}$. Using this value of ϵ in (12) gives us

$$R(T) \leq \mathcal{O}\sqrt{KT \log T}. \tag{13}$$

For more details on these topics refer to [1] and [2].

References

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

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