Lecture 5 - Explore-then-commit algorithms

Guiliang Liu

The Chinese University of Hong Kong, Shenzhen

DDA4230: Reinforcement Learning
Course Page: [Click]

DDA 4230 Resources

Join our Wechat discussion group.

Check our course page.

Course Page Link (all the course relevant materials will be posted here):

https://guiliang.github.io/courses/cuhk-dda-4230/dda_4230.html

The Explore-then-commit (ETC) Algorithm

Explore-then-commit Algorithm: 1) In the first km rounds, the algorithm pulls each arm for k times. 2) The algorithm then calculates the empirical mean of the rewards of each arm. 3) The arm with the best mean will be selected **for the rest of the horizon**.

Algorithm 1: The explore-then-commit algorithm

Input: k: number of exploration pulls on each arm

Output: $\pi(t), t \in \{0, 1, \ldots, T\}$

while $0 \le t \le km - 1$ **do**

$$a_t = (t \bmod m) + 1$$

while $km \le t \le T - 1$ **do**

$$a_t = \arg\max_{i \in [m]} \frac{1}{k} \sum_{t'=0}^{mk-1} r_{t'} \mathbb{1}\{a_{t'} = i\}$$

The Regret of the Explore-then-commit (ETC) Algorithm

We now show a general regret bound of ETC.

Theorem

Assume that $r(i)$ is 1-sub-Gaussian for each i. The regret under ETC satisfies

$$\overline{R}_T \le k \sum_{i \in [m]} \Delta_i + (T - mk) \sum_{i \in [m]} \Delta_i \exp\left(-\frac{k\Delta_i^2}{4} \right). \tag{1}$$

For two-armed bandits ($m = 2$), taking $k = \lceil \max\{1, 4\Delta_2^{-2} \log(T\Delta_2^2/4)\} \rceil$ yields

$$\overline{R}_T \le \Delta_2 + \frac{4}{\Delta_2} + \frac{4}{\Delta_2} \log\left(\frac{T\Delta_2^2}{4} \right). \tag{2}$$

香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen

The Regret of the Explore-then-commit (ETC) Algorithm

Important properties of ETC:

- The regret bound depends on the suboptimality gaps Δ_2 and the horizon T.
- The dependency on $\frac{1}{\Delta_2}$ could be removed at a cost of a larger order of T, e.g., $\overline{R}_t \leq (\Delta_2 + e^{-2})\sqrt{T}$ when $m = 2$.
- The dependence of Δ_2 could be removed with a regret bound of $O(T^{2/3})$,
- The dependence on T can be resolved by a doubling trick without increasing the regret by too much.

The Regret of the Explore-then-commit (ETC) Algorithm

In fact, if the rewards are Gaussian with variance at most 1, the gap-dependent regret bound under $m = 2$ can be further improved by $O(\log\log T)$ by a more careful choice of k. Denote $\Delta = \Delta_2$ and π as the Archimedes' constant.

Theorem

Assume that $r(i)$ is Gaussian with variance at most 1 for each i and $T \geq 4\sqrt{2\pi e}/\Delta^2$. By choosing $k = \lceil \frac{2}{\Delta^2} W(\frac{T^2\Delta^4}{32\pi}) \rceil$, the regret of ETC satisfies

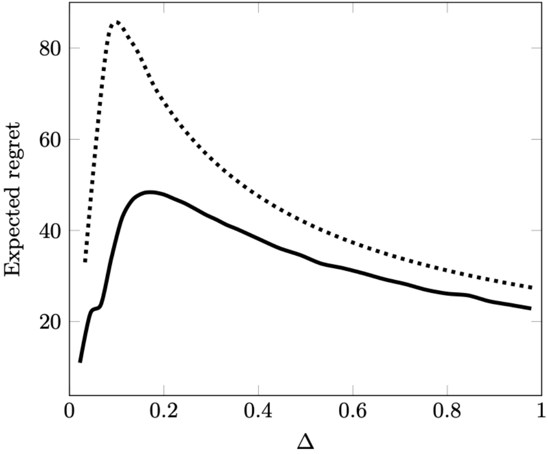
$$\overline{R}_T \leq \Delta + \frac{2}{\Delta}(\log\frac{T^2\Delta^4}{32\pi} - \log\log\frac{T^2\Delta^4}{32\pi} + \log(1+\frac{1}{e}) + 2), \tag{1}$$

where $W(y)\exp(W(y)) = y$ denotes the Lambert function.

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

The Regret of the Explore-then-commit (ETC) Algorithm

Some empirical results. In the following figure we shall see that our upper bound is indeed not bad when the suboptimality gap Δ is large.



Regret (solid line) and regret upper bound (dashed line) of ETC with 2-armed bandit with underlying distribution being Gaussian.

Question and Answering (Q&A)