Lecture 8 - Hardness of Bandits

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

Bandit lower bounds

The regret lower bound: given any fixed bandit algorithm, what is the regret that this algorithm will suffer on some bandit instance?

Select two bandit problem instances under the following conditions:

1. Competition: An action, or, more generally, a sequence of actions that is good for one bandit is not good for the other.
2. Similarity: The instances are 'close' enough that the policy interacting with either of the two instances cannot statistically identify the true bandit with reasonable statistical accuracy.

It seems these two requirements are clearly conflicting.

Can they happen simultaneously?

Analysis

Preliminary 1: When a probability measure $\mathbb{P}$ is absolutely continuous with respect to a probability measure $\mathbb{P}'$ and λ is a common dominating σ-finite measure for $\mathbb{P}$ and $\mathbb{P}'$ (their distributions supported on the same space), denote

$$d_{\mathrm{KL}}(\mathbb{P}\|\mathbb{P}') = \int \mathbb{P} \log \frac{\mathbb{P}}{\mathbb{P}'} d\lambda$$

as the KL-divergence, which is also known as the relative entropy. For example, the KL-divergence between $\mathcal{N}(0, \sigma)$ and $\mathcal{N}(c, \sigma)$ is $\frac{c^2}{2\sigma^2}$.

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

Analysis

Preliminary 2: The discrepancy between probabilities of the same event can be bounded by the discrepancy between the measures, among which we utilize the Bretagnolle-Huber inequality.

Lemma (The Bretagnolle-Huber inequality)

Let $\mathbb{P}, \mathbb{P}'$ be probability measures defined on the same measurable space, then for an arbitrary event A,

$$\mathbb{P}(A) + \mathbb{P}'(\neg A) \geq \frac{1}{2}\exp(-d_{KL}(\mathbb{P}\|\mathbb{P}')) .$$

Analysis

This lemma can be moderately improved by La Cam's inequality. The lemma also trades off with Pinsker's inequality, which bounds the total variation distance

$$\mathbb{P}(A) - \mathbb{P}'(A) \leq \sqrt{\frac{1}{2} d_{\mathsf{KL}}(\mathbb{P} \| \mathbb{P}')}\,.$$

For small $d_{\mathsf{KL}}(\mathbb{P} \| \mathbb{P}')$ Pinsker's inequality is tighter, but for a large KL divergence the Bretagnolle-Huber inequality is more accurate.

Analysis

Lemma (Divergence decomposition)

Consider two bandit instances with reward distribution $\mathbb{P}_1, \ldots, \mathbb{P}_m$ *and* $\mathbb{P}'_1, \ldots, \mathbb{P}'_m$*. Given a fixed policy* π*, denote the distribution of the trajectories on these two instances as* $\mathbb{P}$ *and* $\mathbb{P}'$*. Then,*

$$d_{KL}(\mathbb{P}\|\mathbb{P}') = \sum_{i\in[m]} \mathbb{E}_{\mathbb{P}_\pi}[N_{i,T}]\, d_{KL}(\mathbb{P}_i\|\mathbb{P}'_i).$$

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

Analysis

Armed with the lemmas, we show that the regret of a bandit algorithm is at least $O(\sqrt{mT})$. This bound matches with the instance-independent regret upper bounds achieved by several algorithms that we have discussed.

Theorem

Let $T \geq m-1 \geq 1$. Then for any policy π, there exist $\mu_1, \ldots, \mu_m$, such that with stochastic rewards $\mathcal{N}(\mu_i, 1)$ for arm i, the regret of π on this bandit instance is at least

$$\overline{R}_T \geq \frac{1}{16\sqrt{e}} \sqrt{(m-1)T}\,.$$

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

Analysis

Proof: Let π be a fixed algorithm and write $\mathbb{P}_\pi$ as the probability measure of over the trajectories under executing π on unit-variance Gaussian arms with mean μ. Let $\Delta = \sqrt{\frac{m-1}{4T}}$. Consider two bandit instances $\mu = (\mu_1, \ldots, \mu_m)$ and $\mu' = (\mu'_1, \ldots, \mu'_m)$ where

$$\mu_i = \begin{cases} \Delta, & \text{for } i = 1\,, \\ 0 & \text{otherwise}\,, \end{cases}$$

and

$$\mu'_i = \begin{cases} \Delta, & \text{for } i = 1\,, \\ 2\Delta, & \text{for } i = \arg\min_{j \neq 1} \mathbb{E}_{\mathbb{P}_\pi}\left[N_{j,T}\right]\,, \\ 0 & \text{otherwise}\,, \end{cases}$$

where arg min breaks ties arbitrarily.

Analysis

By the Bretagnolle-Huber inequality, for $A = \{N_{1,T} \leq \frac{T}{2}\}$,

$$\mathbb{P}_\mu(A) + \mathbb{P}_{\mu'}(\neg A) \geq \frac{1}{2}\exp(-d_{\mathrm{KL}}(\mathbb{P}_\mu \| \mathbb{P}_{\mu'})) .$$

By the divergence decomposition,

$$\begin{aligned} d_{\mathrm{KL}}(\mathbb{P}_\mu \| \mathbb{P}_{\mu'}) &= \sum_{i \in [m]} \mathbb{E}_{\mathbb{P}_\pi}[N_{i,T}]\, d_{\mathrm{KL}}(\mathbb{P}_{i,\mu} \| \mathbb{P}_{i,\mu'}) \\ &= \sum_{i \in [m]} \mathbb{1}\{i = \arg\min \mathbb{E}_{\mathbb{P}_\pi}[N_{i,T}]\} \mathbb{E}_{\mathbb{P}_\pi}[N_{i,T}]\, d_{\mathrm{KL}}(\mathbb{P}_{i,\mu} \| \mathbb{P}_{i,\mu'}) \\ &= \min \mathbb{E}_{\mathbb{P}_\pi}[N_{i,T}]\, d_{\mathrm{KL}}(\mathcal{N}(0,1) \| \mathcal{N}(2\Delta, 1)) \\ &\leq \frac{T}{m-1} \cdot 2\Delta^2 . \end{aligned}$$

Analysis

Then, the regret $\overline{R}_T$ and $\overline{R}'_T$ of π on μ and μ' satisfy

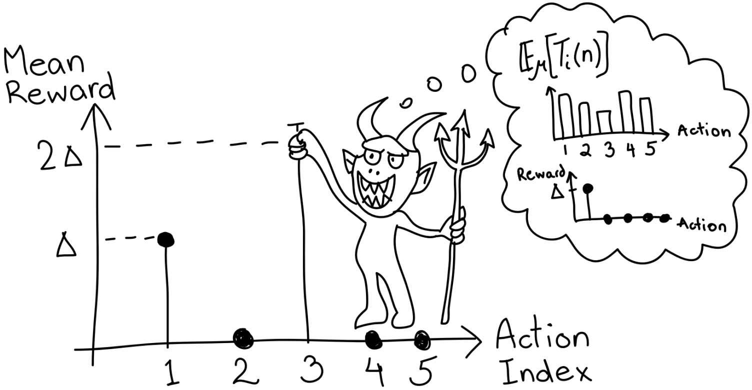
$$\begin{aligned}
\overline{R}_T + \overline{R}'_T &\geq \mathbb{P}_\mu(N_{1,T} \leq \frac{T}{2})\frac{T}{2}\Delta + \mathbb{P}_{\mu'}(N_{1,T} > \frac{T}{2})\frac{T}{2}\Delta \\
&= \frac{T\Delta}{2}(\mathbb{P}_\mu(A) + \mathbb{P}_{\mu'}(\neg A)) \\
&\geq \frac{T\Delta}{2}\frac{1}{2}\exp(-\frac{2T\Delta^2}{m-1}) \\
&= \frac{1}{8\sqrt{e}}\sqrt{(m-1)T}.
\end{aligned}$$

This indicates that the arbitrary bandit algorithm π obtains a combined regret of at least $\frac{1}{8\sqrt{e}}\sqrt{(m-1)T}$ in bandit instances μ and μ'.

It shows that the regret is at least $\frac{1}{16\sqrt{e}}\sqrt{(m-1)T}$ on at least one of the instances.

Analysis

An antagonist who picks μ' to produce a large regret.



Instance-dependent lower bounds

For fixed Δ_i, $i \in [m]$, the regret lower bound is $O(\log T)$, which matches the instance-dependent regret bound of several algorithms that we have discussed.

Theorem

For Gaussian bandit arms with unit variance, the regret of a bandit algorithm is at least

$$\overline{R}_T \geq \sum_{i \in [m]} \frac{2}{\Delta_i} \log T + o(\log T).$$

Question and Answering (Q&A)