DDA4230 Reinforcement learning	Bandit lower bounds
Lecture 9	
Lecturer: Baoxiang Wang	*Scribe: Baoxiang Wang*

1 Goal of this lecture

To understand hardness of bandit problems and the fact that no algorithms guarantee a regret better than $O(\log T)$.
Suggested reading: Chapter 14, 15, 16, and 17 of *Bandit algorithms*;

2 Recap: ε-greedy, ETC, and UCB

For ε-greedy, by choosing $\varepsilon_t = \min\{1, Ct^{-1}\Delta_{\min}^{-2} m\}$ for some absolute constant C, the regret satisfies

$$\overline{R}_T \leq C' \sum_{i \geq 2} \left(\Delta_i + \frac{\Delta_i}{\Delta_{\min}^2} \log \max \left\{ e, \frac{T\Delta_{\min}^2}{m} \right\} \right), \tag{1}$$

where C' is an absolute constant.

For ETC under 2-armed bandits, when $T \geq 4\sqrt{2\pi e}/\Delta^2$, By choosing $k = \lceil \frac{2}{\Delta^2} W(\frac{T^2\Delta^4}{32\pi}) \rceil$, the regret satisfies

$$\overline{R}_T \leq O(\frac{1}{\Delta} \log T\Delta^2) + o(\log T) + \Delta, \tag{2}$$

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

For UCB, by setting $\delta = T^{-2}$, the regret satisfies

$$\overline{R}_T \leq 3 \sum_{i=1}^{m} \Delta_i + \sum_{i:\Delta_i > 0} \frac{16 \log T}{\Delta_i}.$$

This result is followed by a series of improvements.

For TS (Bernoulli bandits) on Bernoulli bandits, the regret satisfies

$$\overline{R}_T \leq \sum_{i:\Delta_i > 0} \frac{\mu_1 - \mu_i}{d_{\mathrm{KL}}(\mu_1 \parallel \mu_i)} \log T + o(\log T)\,,$$

where d_{KL} is the Kullback-Leibler divergence.

3 Bandit lower bounds

In this lecture, we desire to show some negative results, which describe the hardness of bandit problems. We are particularly interested in the regret lower bound, that is, given

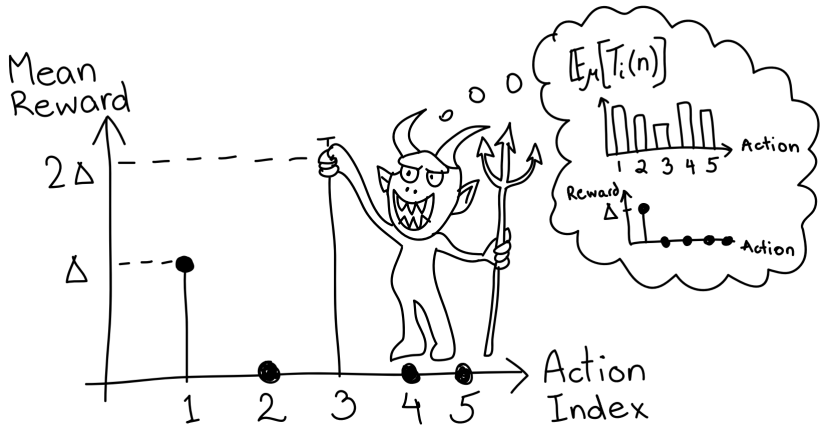

Figure 1: An antagonist who picks μ' to produce a large regret.

any fixed bandit algorithm, what is the regret that this algorithm will suffer on some bandit instance. This gives us some ideas on the performance of our algorithms.

The high-level idea to show a regret lower bound is to select two bandit problem instances in such a way that the following two conditions hold simultaneously:

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.

The two requirements are clearly conflicting. The first makes us want to choose instances with means that are far from each other, while the second requirement makes us want to choose them to be close to each other. The lower bound will follow by optimizing this trade-off.

3.1 Analysis

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 $\mathbb{P}'$, denote

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

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 $\frac{c^2}{2\sigma^2}$. The discrepancy between probabilities of the same event can be bounded by the discrepancy between the discrepancy between the measures, among which we utilize the Bretagnolle–Huber inequality.

Lemma 1 (The Bretagnolle–Huber inequality) *Let $\mathbb{P}, \mathbb{P}'$ be probability measures defined on the same measurable space, then for an arbitrary event A,*

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

This inequality takes equality $\mathbb{P}(A)+\mathbb{P}'(\neg A) = \frac{1}{2}\exp(-d_{\mathrm{KL}}(\mathbb{P} \parallel \mathbb{P}'))$ when $A = \{x \leq x'\}$ for $x \sim \mathbb{P}$, $x' \sim \mathbb{P}'$. When $\mathbb{P}$ and $\mathbb{P}'$ correspond to the probability space of the null hypothesis and the alternative hypothesis, this agrees with the Neyman-Pearson lemma.

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

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

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

Lemma 2 (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, denote the distribution of the trajectories on these two instances as $\mathbb{P}$ and $\mathbb{P}'$. Then,*

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

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

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

Proof: Let π be a fixed algorithm and write $\mathbb{P}_\mu$ 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

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

where arg min breaks ties arbitrarily.

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

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

By divergence decomposition,

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

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

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

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, μ', which indicates that the regret is at least $\frac{1}{16\sqrt{e}}\sqrt{(m-1)T}$ on at least one of the instances, as desired. □

When $T < m-1$, a lower bound of $\frac{(2m-T-1)}{2m}T$ can be shown.

3.2 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 4 *For Gaussian bandit arms with unit variance, the regret of a bandit algorithm is at least*

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

3.3 Alternative analyses

Our proof elaborates two Gaussian bandit instances to lower bound the regret. If we desire regret lower bounds on some specific set of bandit problems, we can replace $d_{\mathrm{KL}}(\mathbb{P}_\mu \parallel \mathbb{P}_{\mu'})$

with the KL-divergence on those distributions as well. We can also use variants of the Bretagnolle-Huber inequality to bound $\mathbb{P}(A) + \mathbb{P}'(\neg A)$, when necessary.

A similar regret lower bound can be proved by Pinsker's inequality. See Exercise 15.2 from the book for more details. For probabilistic lower bounds, we refer to Chapter 17 of the book. We refer readers who are interested in lower bound research to complete these exercise and investigate the lower bounds of other bandit settings.

Acknowledgement

This lecture notes partially use material from *Bandit algorithms*. Figure 1 is from *Bandit algorithms*.