Lower bounds for Bandits and RL

Group Study and Seminar Series (Summer 20)

Yingru Li

The Chinese University of Hong Kong, Shenzhen, China

July 25, 2020

Lattimore, T., & Szepesvári, C. (2020). Bandit algorithms. Cambridge University Press
Osband, I., & Van Roy, B. (2016). On lower bounds for regret in reinforcement learning. arXiv:1608.02732.

Outline

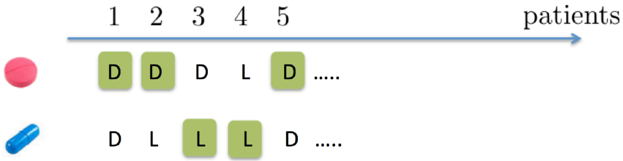
Bandit lower bounds

RL lower bounds

Discussions

Bandit learning and optimization: motivating example

- **First thought experiment:** Clinical trial, **Thompson 1933**



  - Patients with same diseases (e.g. COVID-19) arrive sequentially
  - Two available treatments with unknown rewards (e.g. 'Live' or 'Die')
  - Bandit feedback: after administrating the treatment to a patient, we observe whether she survives or dies. (only rewards of chosen treatment are observed)
  - Goal: design a treatment selection scheme π maximizing the number of patients cured after treatment

Stochastic Multi-armed bandit (MAB) problem

Modern definition follows [Robbins, 1952].

- Set of actions $\mathcal{A}$, say $\mathcal{A} = [k] := \{1, \cdots, k\}$
- A **stochastic bandit** instance is a collection of distributions $\nu = (P_a : a \in \mathcal{A})$
- In each round $t \in \{1, \ldots, n\}$,
  - the learner chooses an action $A_t \in \mathcal{A}$,
  - the environment samples a reward $X_t \in \mathbb{R}$ from distribution P_{A_t} and reveals X_t to the learner.
- The interaction between the learner (or policy) and environment (or instance) induces a probability measure on the sequence of outcomes

$$H_n = (A_1, X_1, A_2, X_2, \ldots, A_n, X_n).$$

Stochastic Multi-armed bandit (MAB) problem

- The interaction between the learner (or policy) and environment (or instance) induces a probability measure on the sequence of outcomes $H_n = (A_1, X_1, A_2, X_2, \dots, A_n, X_n)$.
- The sequence of outcomes should satisfy the following assumptions:

(a) The conditional distribution of reward X_t given H_{t-1}, A_t is P_{A_t},

- captures the intuition that the environment samples X_t from P_{A_t} in round t

(b) The conditional law of action A_t given H_{t-1} is $\pi_t(\cdot \mid H_{t-1})$, where $\pi_1, \pi_2, \dots$ is a sequence of probability kernels that characterize the learner. Define poliy $\pi = (\pi_t)_{t=1}^n$.

- captures the fact that the learner cannot use the future observations in current decisions.

- Density $p_{\nu\pi}(a_1, x_1, \dots, a_n, x_n) = \prod_{t=1}^n \pi(a_t \mid a_1, x_1, \dots, a_{t-1}, x_{t-1}) p_{a_t}(x_t)$

Performance measures of MAB problem - Regret

- Recall $\nu = (P_a : a \in \mathcal{A})$ is a stochastic bandit environment (**instance**)
- Define mean reward of each action for this instance ν: $\mu_a(\nu) = \int_{-\infty}^{\infty} x dP_a(x)$
- Then let $\mu^*(\nu) = \max_{a \in \mathcal{A}} \mu_a(\nu)$ be the largest mean of all the arms.
- The **regret** of policy π on bandit instance ν is

$$R_n(\pi, \nu) = n\mu^*(\nu) - \mathbb{E}\left[\sum_{t=1}^{n} X_t\right] \tag{1}$$

where the expectation is taken with respect to the probability measure on outcomes $H_n = (A_1, X_1, A_2, X_2, \ldots, A_n, X_n)$ induced by the **interaction of π and ν**.

Unstructured bandit (Our focus today)

- **Unstructured bandit**: Playing action a reveals no information about reward distribution on actions $b \neq a$
- **Example**: Class of k-arm bernoulli bandit $\mathcal{E}_{\mathcal{B}}^{k} := \left\{ \left(\mathcal{B}\left(\mu_i\right)\right)_i : \mu \in [0,1]^k \right\}$; Class of k-arm gaussian bandit with known var. $\mathcal{E}_{\mathcal{N}}^{k}\left(\sigma^2\right) := \left\{ \left(\mathcal{N}\left(\mu_i, \sigma^2\right)\right)_i : \mu \in [0,1]^k \right\}$
- **Formal**: An environment class $\mathcal{E}$ is unstructured if $\mathcal{A}$ is finite and there exist sets of distributions $\mathcal{M}_a$ for each $a \in \mathcal{A}$ such that

$$\mathcal{E} = \{\nu = (P_a : a \in \mathcal{A}) : P_a \in \mathcal{M}_a \text{ for all } a \in \mathcal{A}\}$$

or, in short, $\mathcal{E} = \times_{a \in \mathcal{A}} \mathcal{M}_a$.

Structured bandit

Not our focus today. But for a quick tour.

- Let $\mathcal{A} = \{1, 2\}$ and $\mathcal{E} = \{(\mathcal{B}(\theta), \mathcal{B}(1-\theta)) : \theta \in [0,1]\}$.
- **Stochastic linear bandit.** Let $\mathcal{A} \subset \mathbb{R}^d$ and $\theta \in \mathbb{R}^d$ and

$$\nu_\theta = (\mathcal{N}(\langle a, \theta \rangle, 1) : a \in \mathcal{A}) \quad \text{and } \mathcal{E} = \left\{\nu_\theta : \theta \in \mathbb{R}^d\right\}$$

Notice that even if $\mathcal{A}$ is extremely large, the learner can deduce the true environment (recover θ) by playing just d actions that span $\mathbb{R}^d$.

Sturctured reward bandit (cont.)

Not our focus today. But for a quick tour.

- **Lipschitz** reward function for actions in continuous action set.
- Bounded **concave reward (convex loss)** function for actions in bounded convex and compact actions set: **Online Convex Optimization** framework [Hazan, 2016].
- **Reward structure gives us opportunities to obtain reward information of unchosen actions, which is suitable and useful in some practical problems.**

Contextual bandit

Not our focus today. But for a quick tour.

- Incorporate context information in the practical problem (e.g. news recommendation):
- Context $C_t \in \mathcal{C}$ at round t: profile and interests of news app user
- News to be recommended $i \in [k]$
- Reward structure: mean reward $r(c, i)$ is **linear** in the given feature $\psi(c, i) \in \mathbb{R}^d$ of contextual information of users and news, and unknown parameters θ_*:

$$r(c, i) = \langle \theta_*, \psi(c, i) \rangle \, , \text{ for all } (c, i) \in \mathcal{C} \times [k]$$

- **Contextual linear bandit**: in round t, learner is given the decision set $\mathcal{A}_t := \{\psi(C_t, i) : i \in [k]\} \subset \mathbb{R}^d$, from which it chooses an action A_t and receives rewards

$$X_t = \langle \theta_*, A_t \rangle + \eta_t, \quad \eta_t \sim \mathcal{N}(0, 1)$$

Applications already deployed in industry

- Recommendation (Toutiao News, Tik Tok, Taobao)
- Advertisement Placement (Google Ads, Microsoft Decision Services)
- Dynamic Pricing (DiDi, Uber, Salesforce)
- Online Network Routing (Maps app, DiDi)
- Algorithmic Component in AlphaGo (Upper Confidence Tree Search)
- Rate Adaptation in 802.11 wireless systems and other applications in cognitive radio networks, multi-channel communication systems (No idea whether deployed?)

Minimax regret

- Recall that $\mathcal{E}$ is a class of stochastic bandit environments (or instances)
- **Worst-case regret**: $R_n(\pi, \mathcal{E}) = \sup_{\nu \in \mathcal{E}} R_n(\pi, \nu)$
- Let Π be the set of all policies.
- **Minimax regret**:

$$R_n^*(\mathcal{E}) = \inf_{\pi \in \Pi} R_n(\pi, \mathcal{E}) = \inf_{\pi \in \Pi} \sup_{\nu \in \mathcal{E}} R_n(\pi, \nu)$$

- A core activity in bandit theory is to understand what makes $R_n^*(\mathcal{E})$ large or small.

Theorem 1.
Let $k > 1$ and $n \geq k - 1$. Then, for any policy π, there exists a k-armed bandit instance ν,

$$R_n\left(\pi, \nu\right) \geq c\sqrt{(k-1)n},$$

where c is a universal constant.

Minimax regret lower bound - Intuition (1)

- Generally, reduce our bandit problem to hypothesis testing:
- Fix any policy (or learner) π, then select two bandit problem instances ν and ν' s.t. the following hold simultaneously to make life difficult for learner (enlarge regrets)

$$sup_{\mathcal{E}} R_n\left(\pi, \mathcal{E}\right) \geq \max\left\{R_n\left(\pi, \nu\right), R_n\left(\pi, \nu'\right)\right\} \geq 1/2\left(R_n\left(\pi, \nu\right) + R_n\left(\pi, \nu'\right)\right),$$

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.

Minimax regret lower bound - Construction (1)

- Fix any policy $\pi \in \Pi$. Consider a restricted class $\mathcal{E}_{\mathcal{N}}^k(1)$ of k-armed Gaussian bandits with unit variance and bounded mean vector $\mu \in [0,1]^k$
- Let $\Delta \in [0, 1/2]$ be some constant to be chosen later.
- Choose the first Gaussian bandit instance $\nu_\mu \in \mathcal{E}_{\mathcal{N}}^k(1)$ with mean vector:

$$\mu = (\underbrace{\Delta}_{\text{optimal arm in instance } \nu_\mu}, 0, 0, \ldots, 0).$$

- Recall that the interaction of instance ν_μ and policy π give rise to the distribution $\mathbb{P}_{\nu_\mu, \pi}$ on the sequence of outcomes $A_1, X_1, \cdots, A_n, X_n$.
- For brevity, let $\mathbb{P}_\mu := \mathbb{P}_{\nu_\mu, \pi}$, and $\mathbb{E}_\mu$ be expectations under $\mathbb{P}_\mu$

Minimax regret lower bound - Construction (2)

- Define $\mathbb{E}_\mu\left[T_j(n)\right] = \mathbb{E}_\mu\left[\sum_{t=1}^n \mathbb{I}(A_t = j)\right]$ as the expected #times of playing action j up to round n during the interaction between the environment ν_μ and learner π
- To choose the second instance, let $i = \operatorname{argmin}_{j>1} \mathbb{E}_\mu\left[T_j(n)\right]$
  - Since $\sum_{j=1}^k \mathbb{E}_\mu\left[T_j(n)\right] = n$, it holds that $\mathbb{E}_\mu\left[T_i(n)\right] \leq n/(k-1)$.
- The second bandit instance $\nu_{\mu'} \in \mathcal{E}_{\mathcal{N}}^k(1)$ with mean vector:

$$\mu' = (\Delta, 0, 0, \ldots, 0, \underbrace{2\Delta}_{\mu'_i \neq \mu_i,\ \text{optimal arm in instance } \nu_{\mu'}}, 0, \ldots, 0)$$

- Also we abbreviate $\mathbb{P}_{\mu'} := \mathbb{P}_{\nu_{\mu'},\pi}$.

Minimax regret lower bound - Intuition (2)

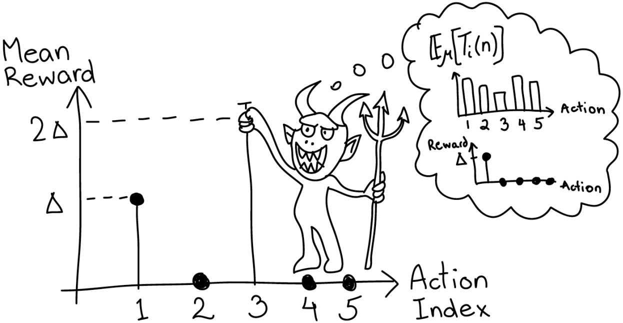


Figure: Given a policy and one environment, the evil antagonist picks another environment so that **the policy will suffer a large regret in at least one environment. Let event $A = \{T_1(n) \leq n/2\}$**

$$R_n(\pi, \nu_\mu) \geq \mathbb{P}_\mu(A)\frac{n\Delta}{2} \quad \text{and} \quad R_n(\pi, \nu_{\mu'}) > \mathbb{P}_{\mu'}(A^c)\frac{n\Delta}{2}.$$

Minimax regret lower bound

- Recall $R_n(\pi, \nu_\mu) = n\mu^*(\nu) - \mathbb{E}_\mu\left[\sum_{t=1}^n X_t\right]$ and event $A = \{T_1(n) \le n/2\}$
- Prove $R_n(\pi, \nu_\mu) \ge \mathbb{P}_\mu(A) \frac{n\Delta}{2}$:

$$\begin{aligned} R_n(\pi, \nu_\mu) &= \mathbb{E}_\mu\left[R_n(\pi, \nu_\mu) \mid A\right] \mathbb{P}_\mu(A) + \mathbb{E}_\mu\left[R_n(\pi, \nu_\mu) \mid A^c\right] \mathbb{P}_\mu(A^c) \\ &\ge \mathbb{E}_\mu\left[R_n(\pi, \nu_\mu) \mid A\right] \mathbb{P}_\mu(A) \ge \frac{\Delta n}{2} \mathbb{P}_\mu(A) \end{aligned}$$

- Prove $R_n(\pi, \nu_{\mu'}) > \mathbb{P}_{\mu'}(A^c) \frac{n\Delta}{2}$:

$$\begin{aligned} R_n(\pi, \nu_{\mu'}) &= \mathbb{E}_{\mu'}\left[R_n(\pi, \nu_{\mu'}) \mid A\right] \mathbb{P}_{\mu'}(A) + \mathbb{E}_{\mu'}\left[R_n(\pi, \nu_{\mu'}) \mid A^c\right] \mathbb{P}_{\mu'}(A^c) \\ &\ge \mathbb{E}_{\mu'}\left[R_n(\pi, \nu_{\mu'}) \mid A^c\right] \mathbb{P}_{\mu'}(A^c) > \frac{\Delta n}{2} \mathbb{P}_\mu(A^c) \end{aligned}$$

- $R_n(\pi, \nu_\mu) + R_n(\pi, \nu_{\mu'}) > \frac{n\Delta}{2}\left(\mathbb{P}_\mu(T_1(n) \le n/2) + \mathbb{P}_{\mu'}(T_1(n) > n/2)\right)$

Minimax regert lower bound

- $R_n(\pi, \nu_\mu) + R_n(\pi, \nu_{\mu'}) > \frac{n\Delta}{2} \left(\mathbb{P}_\mu (T_1(n) \leq n/2) + \mathbb{P}_{\mu'} (T_1(n) > n/2)\right)$
- Recall the Le Cam's method in reduction to binary testing, we will have similar argument:

Lemma 2 (Bretagnolle-Huber inequality).

Let P and Q be probability measures on the same measurable space $(\Omega, \mathcal{F})$, and let $A \in \mathcal{F}$ be an arbitrary event. Then,

$$P(A) + Q(A^c) \geq \frac{1}{2} \exp(-\mathrm{D}(P, Q)) \tag{2}$$

where $A^c = \Omega \backslash A$ is the complement of A.

- Then we have $R_n(\pi, \nu_\mu) + R_n(\pi, \nu_{\mu'}) > \frac{n\Delta}{4} \exp(-D(\mathbb{P}_\mu, \mathbb{P}_{\mu'}))$

Minimax regret lower bound - Divergence decomposition

Lemma 3 (Divergence decomposition).

Let $\nu = (P_1, \ldots, P_k)$ *be the reward distributions associated with one* k *-armed bandit, and let* $\nu' = (P_1', \ldots, P_k')$ *be the reward distributions associated with another* k*-armed bandit. Fix some policy* π *and let* $\mathbb{P}_\nu = \mathbb{P}_{\nu\pi}$ *and* $\mathbb{P}_{\nu'} = \mathbb{P}_{\nu'\pi}$ *be the probability measures on the sequence of outcomes induced by the n-round interconnection of* π *and* ν *(respectively,* π *and* ν' *). Then,*

$$\mathrm{D}\left(\mathbb{P}_\nu, \mathbb{P}_{\nu'}\right) = \sum_{i=1}^{k} \mathbb{E}_\nu\left[T_i(n)\right] \mathrm{D}\left(P_i, P_i'\right) \tag{3}$$

Minimax regret lower bound - Divergence decomposition (cont.)

- Density of $\mathbb{P}_\nu$ is $p_{\nu\pi}\left(a_1, x_1, \ldots, a_n, x_n\right) = \prod_{t=1}^{n} \pi\left(a_t \mid a_1, x_1, \ldots, a_{t-1}, x_{t-1}\right) p_{a_t}\left(x_t\right)$
- Density of $\mathbb{P}_{\nu'}$ is $p_{\nu\pi}\left(a_1, x_1, \ldots, a_n, x_n\right) = \prod_{t=1}^{n} \pi\left(a_t \mid a_1, x_1, \ldots, a_{t-1}, x_{t-1}\right) p'_{a_t}\left(x_t\right)$
-
$$\log \frac{d\mathbb{P}_\nu}{d\mathbb{P}_{\nu'}}\left(a_1, x_1, \ldots, a_n, x_n\right) = \sum_{t=1}^{n} \log \frac{p_{a_t}\left(x_t\right)}{p'_{a_t}\left(x_t\right)}$$
-
$$\mathrm{D}\left(\mathbb{P}_\nu, \mathbb{P}_{\nu'}\right) = \mathbb{E}_\nu\left[\log \frac{d\mathbb{P}_\nu}{d\mathbb{P}_{\nu'}}\left(A_1, X_1, \ldots, A_n, X_n\right)\right] = \sum_{t=1}^{n} \mathbb{E}_\nu\left[\log \frac{p_{A_t}\left(X_t\right)}{p'_{A_t}\left(X_t\right)}\right]$$
-
$$\mathbb{E}_\nu\left[\log \frac{p_{A_t}\left(X_t\right)}{p'_{A_t}\left(X_t\right)}\right] = \mathbb{E}_\nu\left[\mathbb{E}_\nu\left[\log \frac{p_{A_t}\left(X_t\right)}{p'_{A_t}\left(X_t\right)} \mid A_t\right]\right] = \mathbb{E}_\nu\left[\mathrm{D}\left(P_{A_t}, P'_{A_t}\right)\right]$$

Minimax regret lower bound - Divergence decomposition (cont.)

- Since $\sum_{i=1}^{k} \mathbb{I}\{A_t = i\} = 1 \quad a.s.$, by linearity of expectation

$$\sum_{t=1}^{n} \mathbb{E}_\nu \left[\mathrm{D}\left(P_{A_t}, P'_{A_t}\right)\right] = \sum_{i=1}^{k} \mathbb{E}_\nu \left[\sum_{t=1}^{n} \mathbb{I}\{A_t = i\} \,\mathrm{D}\left(P_{A_t}, P'_{A_t}\right)\right] = \sum_{i=1}^{k} \mathbb{E}_\nu \left[T_i(n)\right] \mathrm{D}\left(P_i, P'_i\right)$$

- Recall the construction of two instances, only difference is at index i:

$$\mu = (\underbrace{\Delta}_{\text{optimal arm in instance } \nu_\mu}, 0, 0, \ldots, 0)$$

$$\mu' = (\Delta, 0, 0, \ldots, 0, \underbrace{2\Delta}_{\mu'_i \neq \mu_i, \text{ optimal arm in instance } \nu_{\mu'}}, 0, \ldots, 0)$$

$$\mathrm{D}\left(\mathbb{P}_\mu, \mathbb{P}_{\mu'}\right) = \mathbb{E}_\mu \left[T_i(n)\right] \mathrm{D}(\mathcal{N}(0,1), \mathcal{N}(2\Delta, 1)) = \mathbb{E}_\mu \left[T_i(n)\right] \frac{(2\Delta)^2}{2} \leq \frac{2n\Delta^2}{k-1}$$

Minimax regret lower bound

- Putting all together, for any fixed policy $\pi \in \Pi$:

$$\sup_{\nu \in \mathcal{E}_{\mathcal{N}}^{k}(1)} R_n(\pi, \nu) \geq (1/2)\left(R_n\left(\pi, \nu_\mu\right) + R_n\left(\pi, \nu_{\mu'}\right)\right) \geq \frac{n\Delta}{8} \exp\left(-\frac{2n\Delta^2}{k-1}\right)$$

- By choosing $\Delta = \sqrt{(k-1)/4n} \leq 1/2$ and lower bounding $\exp(-1/2)$,

$$\inf_{\pi \in \Pi} \sup_{\nu \in \mathcal{E}} R_n(\pi, \nu) \geq \inf_{\pi \in \Pi} \sup_{\nu \in \mathcal{E}_{\mathcal{N}}^{k}(1)} R_n(\pi, \nu) \geq \frac{1}{27}\sqrt{(k-1)n}$$

Minimax regret lower bound - Alternative construction

- Constant improvement.
- Fix policy $\pi \in \Pi$. Let Δ to be chosen later.
- Construct $k+1$ instances. For each instance $\nu_{\mu^{(i)}}$ indexed with $i \in \{0, 1, \ldots, k\}$, let the mean vector $\mu^{(i)} \in \mathbb{R}^k$ be $\mu_j^{(i)} = \mathbb{I}\{i = j\}\Delta$

$$
\begin{aligned}
\mu^{(0)} &= (0, \cdots, 0, \cdots, 0) \\
\mu^{(i)} &= (0, \cdots, \underbrace{\Delta}_{i\text{-arm}}, \ldots, 0), \quad \forall i \in \{1, 2, \cdots, k\}
\end{aligned}
$$

- Further abbreviate the notation $\mathbb{E}_i[\cdot] = \mathbb{E}_{\mu^{(i)}}[\cdot]$

Minimax regret lower bound - Alternative proof

- For any fixed policy $\pi \in \Pi$, let $R_i = R_n\left(\pi, \nu_{\mu^{(i)}}\right)$,

$$\sup_{\nu \in \mathcal{E}} R_n(\pi, \nu) \geq \max\{R_1, \cdots, R_k\} \geq (1/k) \sum_{i=1}^{k} R_i = \frac{\Delta}{k} \sum_{i=1}^{k} (n - \mathbb{E}_i\left[T_i(n)\right])$$

- Define a random variable J_n denoting the frequency of playing arm j up to round n under the interaction between policy π and instance $\nu_{\mu^{(i)}}$:

$$\mathbb{P}_i\left(J_n = j\right) = \mathbb{E}_i \frac{T_j(n)}{n}$$

- Then we rewrite,

$$(1/k) \sum_{i=1}^{k} R_i = n\Delta \left(1 - \frac{1}{k} \sum_{i=1}^{k} \mathbb{P}_i(J_n = i)\right)$$

Minimax regret lower bound - Alternative proof

- Let event $A = \{J_n = i\}$, then by Pinsker's inequality:

$$\mathbb{P}_i(A) - \mathbb{P}_0(A) \leq \sup_A \mathbb{P}_i(A) - \mathbb{P}_0(A) \leq \sqrt{\frac{1}{2}\mathrm{D}(\mathbb{P}_0, \mathbb{P}_i)}$$

- Since $\sum_{i=1}^k \mathbb{P}_0\left(J_n = i\right) = 1$, we have

$$\frac{1}{k}\sum_{i=1}^k \mathbb{P}_i\left(J_n = i\right) \leq \frac{1}{k} + \frac{1}{k}\sum_{i=1}^k \sqrt{\frac{1}{2}\mathrm{D}\left(\mathbb{P}_0, \mathbb{P}_i\right)}$$

- Then, we have the following important immediate result:

$$(1/k)\sum_{i=1}^k R_i \geq n\Delta\left(1 - \frac{1}{k} - \frac{1}{k}\sum_{i=1}^k \sqrt{\frac{1}{2}\mathrm{D}\left(\mathbb{P}_0, \mathbb{P}_i\right)}\right) \tag{4}$$

Minimax regret lower bound - Alternative proof

- By Divergence decomposition, $\mathrm{D}(\mathbb{P}_0, \mathbb{P}_i) = D(\mathcal{N}(0,1), \mathcal{N}(\Delta,1))\mathbb{E}_0[T_i(n)]$
- we have

$$\begin{aligned}(1/k)\sum_{i=1}^{k} R_i \geq n\Delta\left(1 - \frac{1}{k} - \frac{1}{k}\sum_{i=1}^{k}\sqrt{\frac{1}{2}\mathrm{D}\left(\mathbb{P}_0, \mathbb{P}_i\right)}\right) &= n\Delta\left(1 - \frac{1}{k} - \frac{1}{k}\sum_{i=1}^{k}\sqrt{\frac{\Delta^2}{4}\mathbb{E}_0[T_i(n)]}\right) \\ \geq n\Delta\left(1 - \frac{1}{k} - \frac{1}{k}\frac{\Delta}{2}\sqrt{k\sum_{i=1}^{k}\mathbb{E}_0[T_i(n)]}\right) &= n\Delta\left(1 - \frac{1}{k} - \frac{\Delta}{2k}\sqrt{kn}\right)\end{aligned}$$

- Let $\Delta = c\sqrt{k/n}$ with some constant c, we get the final result

$$\inf_{\pi\in\Pi}\sup_{\nu\in\mathcal{E}} R_n(\pi,\nu) \geq \inf_{\pi\in\Pi}\max\{R_1, \cdots, R_k\} \geq \frac{1}{8}\sqrt{kn}$$

Outline

Bandit lower bounds

RL lower bounds

Discussions

RL hard MDP construction - Intuition

- Known deterministic reward $r(s=0)=0$ and $r(s=1)=1$
- All actions from the state 0 follow the same law $P(0,a)=(1-\delta_0,\delta_0)$.
- In state 1, $P(1,a)=(\delta_1,1-\delta_1)$ for $a\neq a^*$ and $P\left(1,a^*\right)=(\delta_1-\epsilon,1-\delta_1+\epsilon)$.
- For this simple class of MDP we will distinguish policies in terms of their action upon $s=1$, since this is the only action which can influence the evolution of the MDP.

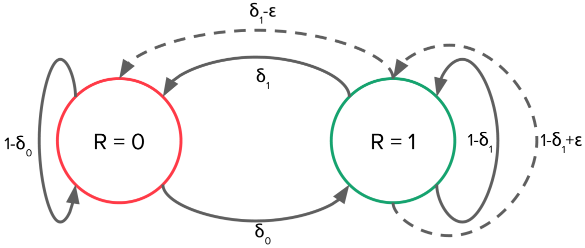


Figure: A hard-to-learn class of two state MDP. Dotted lines distinguish the unique optimal policy.

RL hard MDP construction - Intuition

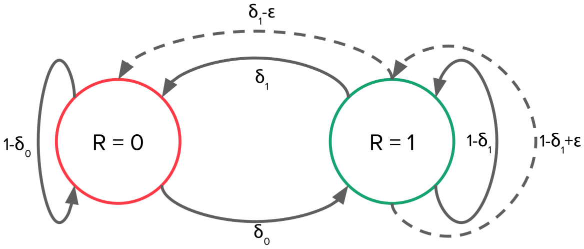


Figure: A hard-to-learn class of two state MDP. Dotted lines distinguish the unique optimal policy.

- We define $\theta_1 := \frac{\delta_0}{\delta_0+\delta_1}$ to be the average expected reward under the policy $a \neq a^*$.
- Let $\delta_1^* := \delta_1 - \epsilon$ for the distinguished optimal action,
- Correspondingly $\theta_1^* := \frac{\delta_0}{\delta_0+\delta_1^*}$ for the average expected reward under the optimal policy a^*

RL hard MDP construction - Intuition

- Assume $\delta_0 \geq \delta_1$, the agent should obtain expected regret $\Omega\left(\epsilon/\delta_0\right)$ every timestep it selects action $a_t \neq a^*$ whilst in state $s = 1$.

$$\begin{aligned}\theta_1^* - \theta_1 &= \frac{\delta_0}{\delta_0 + \delta_1 - \epsilon} - \frac{\delta_0}{\delta_0 + \delta_1} \\ &= \frac{\delta_0 \epsilon}{\left(\delta_0 + \delta_1\right)\left(\delta_0 + \delta_1 - \epsilon\right)} \\ &> \frac{\delta_0 \epsilon}{\left(\delta_0 + \delta_1\right)^2} > \frac{\delta_0 \epsilon}{\left(2\delta_0\right)^2} = \frac{\epsilon}{4\delta_0}\end{aligned}$$

- All other actions in any other state produce zero regret.
- The proportion of the time the agent spends in state $s = 1$ is lower bounded by θ_1.

RL regret lower bound

Construct A different hard instance by setting $a^* = i \in \{1, 2, \cdots, A\}$ and one additional $\mathrm{uniform}$ instance. Let the regret of policy π on ergodic RL problem under instance $\mathcal{M}_i$ be

$$R_T(\pi, \mathcal{M}_i) = \theta_1^* T - \mathbb{E}_i \left[\sum_{t=1}^{T} r\left(s_1, a_t\right) \right]$$

Lemma 4 (Informal).

In the environment of Figure 1*, when the optimal action on* s_1 *is* $a^* = i \in \mathcal{A}$*, for all* $\delta, \epsilon > 0$ *and all learning algorithms* π*,*

$$\frac{1}{A} \sum_{i \in \mathcal{A}} R_T(\pi, \mathcal{M}_i) \geq \theta_1 \frac{\epsilon}{4\delta_0} T \left(1 - \frac{1}{A} - \frac{1}{A} \sum_{i \in \mathcal{A}} \sqrt{\frac{1}{2} \mathrm{D}\left(P_{\mathrm{unif}}, P_i\right)} \right)$$

$$
\begin{aligned}
\frac{1}{A}\sum_{i\in\mathcal{A}} R_T(\pi,\mathcal{M}_i) &\geq \theta_1 \frac{\epsilon}{4\delta_0} T \left(1 - \frac{1}{A} - \frac{1}{A}\sum_{i\in\mathcal{A}} \sqrt{\frac{1}{2}\mathrm{D}\left(P_{\mathrm{unif}}, P_i\right)}\right) \\
&\geq \theta_1 \frac{\epsilon}{4\delta_0} T \left(1 - \frac{1}{A} - \sqrt{\frac{1}{2}\frac{\epsilon^2}{\delta_1}\frac{\theta_1 T}{A}}\right) \quad \text{for all } \epsilon \\
&\geq \frac{1}{4}\frac{\epsilon\theta_1 T}{\delta_0}\left(1 - \frac{1}{A} - \sqrt{\frac{\epsilon^2\theta_1 T}{2\delta_1 A}}\right) \\
&\geq \frac{1}{4}\cdot\sqrt{\frac{\delta_1 A}{8\theta_1 T}\cdot\frac{\theta_1 T}{\delta_0}}\left(1 - \frac{1}{A} - \frac{1}{4}\right) \quad \text{setting } \epsilon = \sqrt{\frac{\delta_1 A}{8\theta_1 T}} \\
&\geq \frac{1}{32\sqrt{2}}\sqrt{\frac{\delta_1\theta_1}{\delta_0^2}AT}
\end{aligned}
$$

Diameter of MDP

- $T_{\mu}^{M}(s, s')$ for the expected number of time steps to get from state s to s' in MDP M under policy μ.
- The one-way diameter of an MDP is defined

$$D_{\mathrm{ow}}(M) := \max_{s} \min_{\mu} T_{\mu}^{M}(s, \bar{s}), \text{ where } \bar{s} \text{ is any state with optimal value bias.}$$

- From construction of the simple two-state MDP, it is clear that $D_{\mathrm{ow}} = \frac{1}{\delta_0}$, since the only state with optimal value bias is $s = 1$ and the expected time from $s = 0$ to $s = 1$ is $\frac{1}{\delta_0}$. We now examine behavior of the remaining free parameters using the definition $\theta_1 = \delta_0 / (\delta_0 + \delta_1)$
- (Notice!) For finite horizon MDP with horizon H, $D_{\mathrm{ow}} = \Theta(H)$

- For any choice of $\delta_1 > 0$,

$$\sqrt{\frac{\delta_1 \theta_1}{\delta_0^2}} = D_{\text{ow}} \sqrt{\delta_1 \theta_1} = D_{\text{ow}} \sqrt{\frac{\delta_1 / D_{\text{ow}}}{\delta_1 + 1/D_{\text{ow}}}} = \sqrt{\frac{D_{\text{ow}}}{1 + \frac{1}{\delta_1 D_{\text{ow}}}}} = O(\sqrt{D_{\text{ow}}})$$

- Establish a lower bound $\Omega(\sqrt{D_{\text{ow}} SAT})$ for ergodic RL problem and imply lower bound $\Omega(\sqrt{HSAT})$ for finite horizon problem.
- For finite horizon problem, UCBVI achieves the lower bound under large T and large finite state-action space.
- Jaksch et al. [2010] establish $\Omega(\sqrt{DSAT})$, and design the UCRL2 algorithm achieving $\tilde{O}(DS\sqrt{AT})$ upper bound, where $D(M) := \max_{s,s'} \min_\mu T_\mu^M(s, s') \geq D_{\text{ow}}$ is the diameter of the MDP.
- Tossou et al. [2019] design algorithm UCRL-V and close the gap for ergodic RL problem $\tilde{O}(\sqrt{DSAT})$

Instance (Gap)-dependent lower bound

- Regret for bandit
- PAC for bandit
- Regret for episodic RL [Simchowitz and Jamieson, 2019]

$$\lim_{K\to\infty} \frac{\mathbb{E}^{\mathcal{M}}\left[\mathrm{Regret}_K\right]}{\log T} \gtrsim (1-\alpha) \sum_{x,a:\mathrm{gap}_1(x,a)>0} \frac{H^2}{\mathrm{gap}_1(x,a)}$$

Outline

Bandit lower bounds

RL lower bounds

Discussions

Performance measures for RL

- Expected Regret: There exists a function $F_{\mathrm{ER}}(S, A, H, T)$ such that

$$\mathbb{E}[R(T)] \leq F_{\mathrm{ER}}(S, A, H, T)$$

- High Probability Regret: There exists a function $F_{\mathrm{HPR}}(S, A, H, T, \log(1/\delta))$ such that

$$\mathbb{P}\left(R(T) > F_{\mathrm{HPR}}(S, A, H, T, \log(1/\delta))\right) \leq \delta$$

- Probably approximately correct (PAC): $(\varepsilon, \delta) - PAC$: There exists a polynomial function $F_{\mathrm{PAC}}(S, A, H, 1/\varepsilon, \log(1/\delta))$ such that

$$\mathbb{P}\left(N_\varepsilon > F_{\mathrm{PAC}}(S, A, H, 1/\varepsilon, \log(1/\delta))\right) \leq \delta$$

Regret bounds

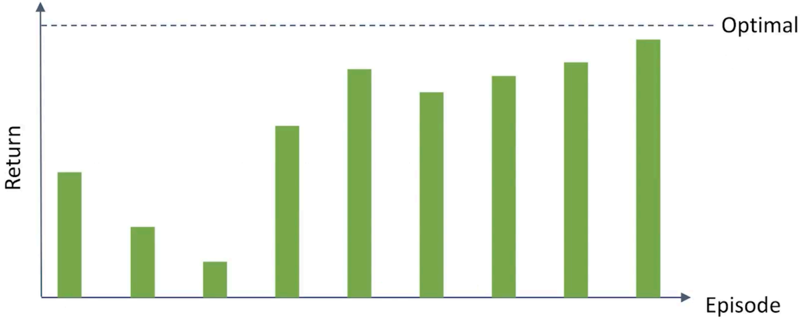


Figure: Return (accumulated reward) at each episode

Regret bounds

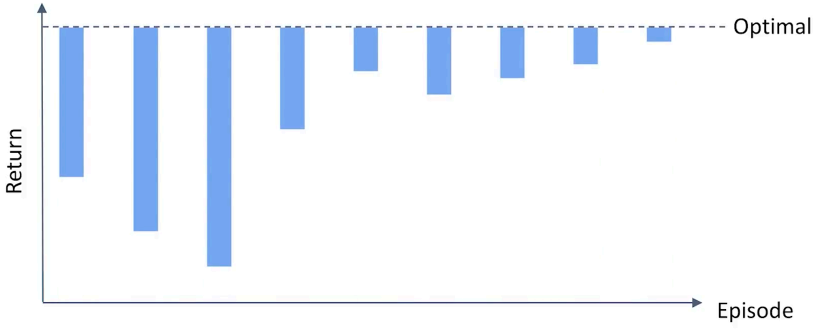


Figure: Optimality gap of the return at each episode

Regret bounds

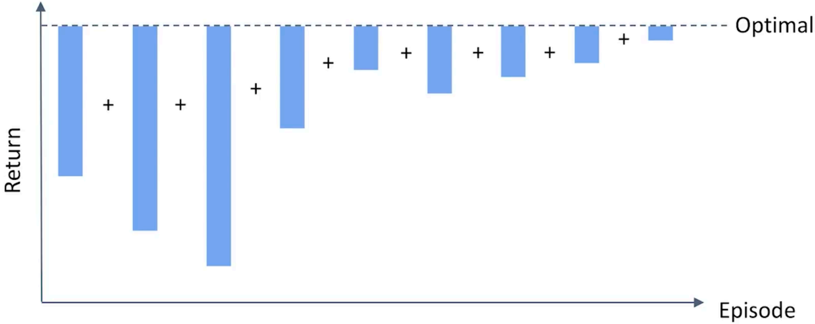


Figure: Bound on sum of differences between optimal and achieved performance (with high probability)

PAC bounds

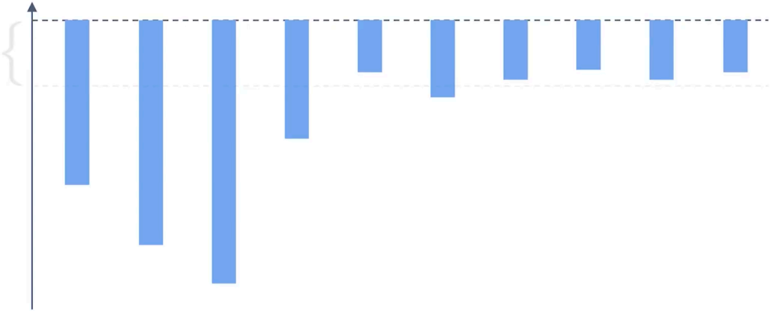


Figure: Optimality gap of the return at each episode

PAC bounds

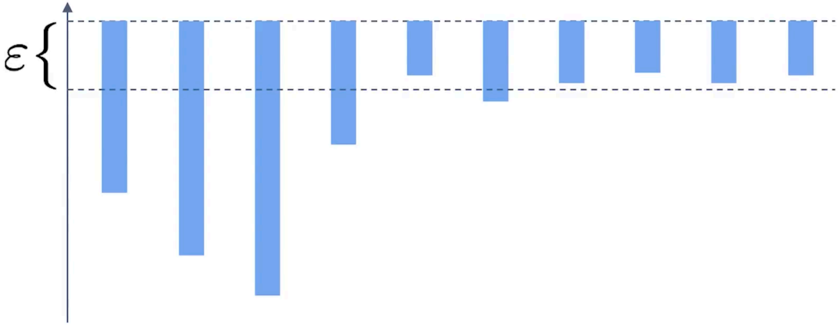


Figure: Compare optimality gap with fixed threshold ϵ. All episodes that do not achieve ϵ-optimal are considered as 'mistakes'.

PAC bounds

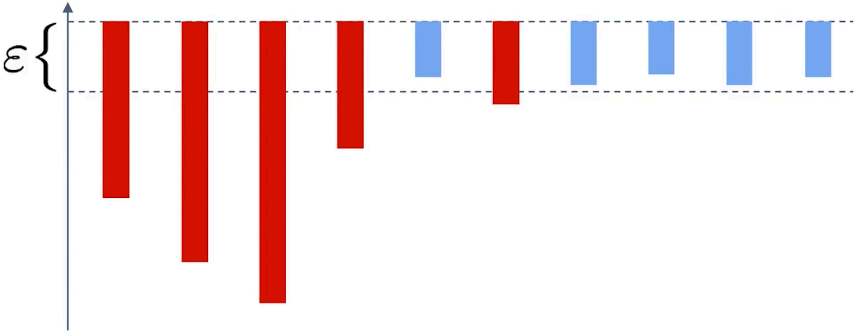


Figure: Bound on #episodes where performance is not ϵ-optimal (with high probability)

Limitations of the performance measures and beyond

- Limitations of PAC bounds

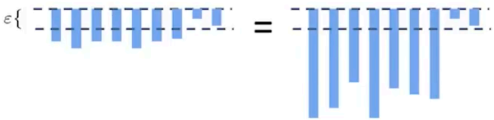


Figure: No guarantee of how bad 'mistakes' are

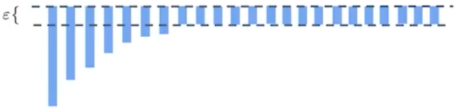


Figure: Allow not converging to optimal

- Limitation of Regret bounds

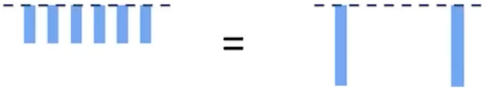


Figure: Only bound the total sum of optimality gap (errors)

- PAC bounds and Regret bounds are not directly comparable!

Limitations of the performance measures and beyond

- Motivation beyond PAC and regret performance measures:
  - (Safety issues) High-stake applications like robotics and healthcare, etc
  - Compare algorithms beyond experiments
- Uniform PAC [Dann et al., 2017]: For all $\varepsilon > 0$ jointly: bound #episodes where performance is not ε -optimal (with high probability)
- Individual Policy Certificates (IPOC) [Dann et al., 2019]
  - IPOC imply both PAC and Regret
  - First algorithm achieve both PAC lower bound and Regret lower bound

Current construction is not suitable for Long horizon problems

- **Reward Uniformity** The regularity assumption of Dann and Brunskill (2015) is the standard $r_h \in [0,1]$ (and hence $\sum_{h=1}^{H} r_h \in [0,H]\Big)$. To remove the dependence on H due to reward scaling, we should normalize their cumulative reward to [0,1] by dividing reward by H. Now compare their assumption (after normalization) to ours:
- Standard assumption (e.g., Dann and Brunskill, 2015): $r_h \in \left[0, \frac{1}{H}\right]$, and hence $\sum_{h=1}^{H} r_h \in [0,1]$
- More general assumption (e.g., Krishnamurthy et al., 2016): $r_h \geq 0$, and $\sum_{h=1}^{H} r_h \in [0,1]$
- It is clear that our assumption is strictly weaker, despite that it might seem more restrictive at the first glance. (A key subtlety here is on the interpretation of ϵ : only after normalization does represent the relative suboptimality gap (Kakade, 2003, Chapter 2.2.3).)
- In fact, requiring $r_h \in \left[0, \frac{1}{H}\right]$ effectively imposes a uniformity requirement on rewards, and cannot model environments with sparse rewards-for which we believe long horizons are most challenging-in a tight manner.

Long horizon problems

- **Asymptotics** The other assumption they have is $\epsilon \in \left[0, \frac{1}{H}\right]$ (after normalization). For some of our motivating scenarios, such an asymptotic situation is uninteresting: for example, the horizon of a control task can be, say, $H \sim 10^6$, when we control motors that respond in millisecond intervals ("flat RL"), but the horizon may reduce significantly if pre-defined macro actions are available ("hierarchical RL"). In this case, learning a policy 10^{-6} close to optimal is unnecessary, and to show the advantage of hierarchical RL we are interested in the regime of $\epsilon \gg 1/H$

Long horizon problems

- The lower bound construction given by Dann and Brunskill [2015] is as follows: the agent chooses an action in the first step, transitions to either a good state or a bad state with action-dependent probabilities, and then loops in the good / bad state for the remaining time steps receiving either $+1$ or 0 reward per time step. Once we normalize total reward, the construction is exactly a multi-armed bandit with Bernoulli distributed rewards, which obviously will not yield any H dependence.
- Another type of lower bound constructions in literature utilize lazy Markov chains [Jaksch et al., 2010] typically there are a good state and a bad state, and under all actions the agent will stay in its current state and only transition to the other state with small probabilities. The small probabilities of switching states are set as $O\left(\frac{1}{H}\right)$; As H increases, the MDP simply becomes lazier, and can be emulated by sampling episodes from an MDP with smaller H and adding uninformative "elapsing" time steps.

References I

C. Dann and E. Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, Advances in Neural Information Processing Systems 28, pages 2818–2826. Curran Associates, Inc., 2015. URL http://papers.nips.cc/paper/5827-sample-complexity-of-episodic-fixed-horizon-reinforcement-learning.pdf.

C. Dann, T. Lattimore, and E. Brunskill. Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning. In Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, pages 5717–5727, USA, 2017. Curran Associates Inc. ISBN 978-1-5108-6096-4.

References II

C. Dann, L. Li, W. Wei, and E. Brunskill. Policy certificates: Towards accountable reinforcement learning. In International Conference on Machine Learning, pages 1507–1516, 2019.

E. Hazan. Introduction to online convex optimization. Foundations and Trends in Optimization, 2(3-4):157–325, 2016.

T. Jaksch, R. Ortner, and P. Auer. Near-optimal regret bounds for reinforcement learning. Journal of Machine Learning Research, 11(Apr):1563–1600, 2010.

H. Robbins. Some aspects of the sequential design of experiments. Bulletin of the American Mathematical Society, 58(5):527–535, 1952.

M. Simchowitz and K. G. Jamieson. Non-asymptotic gap-dependent regret bounds for tabular mdps. In Advances in Neural Information Processing Systems, pages 1153–1162, 2019.

References III

A. Tossou, D. Basu, and C. Dimitrakakis. Near-optimal optimistic reinforcement learning using empirical bernstein inequalities. arXiv preprint arXiv:1905.12425, 2019.