Revisit Minimax Lower Bounds of Episodic Reinforcement Learning in Finite MDPs

Presenter: Hao Liang

The Chinese University of Hong Kong, Shenzhen

July 9, 2021

Mainly based on:
Domingues, Omar Darwiche, et al. "Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited." Algorithmic Learning Theory. PMLR, 2021.

Literature review

- In the average-reward setting, Jaksch et al. (2010) prove $\Omega(\sqrt{DSAT})$ lower bound
  - D: the diameter of the MDP
  - T: the total number of steps
- In the episodic setting, the total number of steps taken is HT and H is roughly the equivalent of the diameter D.
- Intuitively, the lower bound of Jaksch et al. (2010) should be translated to $\Omega\left(\sqrt{H^2SAT}\right)$ for episodic MDPs after T episodes.
- However, their construction only applies to stationary MDP.
- Jin et al. (2018) claim that the lower bound becomes $\Omega\left(\sqrt{H^3SAT}\right)$ by using the construction of Jaksch et al. (2010) and a mixing-time argument, but no complete proof.

Setting and Performance Measures

- Episodic Markov decision process (MDP) $\mathcal{M} \triangleq (\mathcal{S}, \mathcal{A}, H, \mu, p, r)$.
- $S := |\mathcal{S}|$ and $A := |\mathcal{A}|$
- $\Delta(\mathcal{A})$: the set of probability distributions over the action set
- $\mathcal{I}_h^t = \big((\mathcal{S} \times \mathcal{A})^{H-1} \times \mathcal{S}\big)^{t-1} \times (\mathcal{S} \times \mathcal{A})^{h-1} \times \mathcal{S}$ the set of possible histories up to step h of episode (not including rewards)
- $\big(s_1^1, a_1^1, s_2^1, a_2^1, \dots, s_H^1, \dots, s_1^t, a_1^t, s_2^t, a_2^t, \dots, s_h^t\big) \in \mathcal{I}_h^t$
- A Markov policy is a function $\pi : \mathcal{S} \times [H] \to \Delta(\mathcal{A})$
- A history-dependent policy is a sequence of functions $\boldsymbol{\pi} \triangleq (\pi_h^t)_{t \geq 1, h \in [H]}$ with $\pi_h^t : \mathcal{I}_h^t \to \Delta(\mathcal{A})$
- Π_{Markov} and Π_{Hist} the sets of Markov and history-dependent policies

Setting and Performance Measures

- A policy π interacting with an MDP induces a stochastic process $(S_h^t, A_h^t)_{t\geq 1, h\in[H]}$
- $I_h^t \triangleq \left(S_1^1, A_1^1, S_2^1, A_2^1, \ldots, S_H^1, \ldots, S_1^t, A_1^t, S_2^t, A_2^t, \ldots, S_h^t\right)$: the random history
- $\mathcal{F}_h^t$: the σ-algebra generated by I_h^t
- $\mathbb{P}_{\mathcal{M}}\left[I_H^T = i_H^T\right] = \prod_{t=1}^T \mu\left(s_1^t\right) \prod_{h=1}^{H-1} \pi_h^t\left(a_h^t \mid i_h^t\right) p_h\left(s_{h+1}^t \mid s_h^t, a_h^t\right)$
- Let $\mathbb{E}_{\mathcal{M}}$ be the corresponding expectation (implicitly dependent on π)
- In episode t, the value of a policy $\boldsymbol{\pi}$ in the MDP $\mathcal{M}$ is defined as

$$V^{\boldsymbol{\pi},t}\left(i_H^{t-1}, s\right) \triangleq \mathbb{E}_{\boldsymbol{\pi},\mathcal{M}}\left[\sum_{h=1}^H r_h\left(S_h^t, A_h^t\right) \mid I_H^{t-1} = i_H^{t-1}, S_1^t = s\right]$$

- For Markov policy, the value does not depend on i_H^{t-1}

$$V^{\pi}(s) \triangleq \mathbb{E}_{\pi,\mathcal{M}}\left[\sum_{h=1}^H r_h\left(S_h^1, A_h^1\right) \mid S_1^1 = s\right]$$

Setting and Performance Measures

- The optimal value function $V^*(s) \triangleq \max_{\pi \in \Pi} V^\pi(s)$ achieved by (Markov) π^*
- Markov policies suffices

$$V^*(s) \geq V^{\boldsymbol{\pi},t}\left(i_H^{t-1}, s\right)$$

- Define the average value functions over the initial state as

$$\rho^{\pi,t}\left(i_H^{t-1}\right) \triangleq \mathbb{E}_{s\sim\mu}\left[V^{\boldsymbol{\pi},t}\left(i_H^{t-1}, s\right)\right], \quad \rho^{\pi} \triangleq \mathbb{E}_{s\sim\mu}\left[V^{\pi}(s)\right], \quad \rho^* \triangleq \rho^{\pi^*}$$

- The expected regret of an algorithm π in an MDP $\mathcal{M}$ after T episodes is defined as

$$\mathcal{R}_T(\boldsymbol{\pi}, \mathcal{M}) \triangleq \mathbb{E}_{\boldsymbol{\pi},\mathcal{M}}\left[\sum_{t=1}^{T}\left(\rho^* - \rho^{\boldsymbol{\pi},t}\left(I_H^{t-1}\right)\right)\right]$$

Lower Bound Recipe

- Consider a class $\mathcal{C}$ of hard MDPs instances
  - the optimal policy is difficult to identify
  - close to each other, but the behavior of an algorithm is expected to be different
- Use a change of distribution between two well-chosen MDPs to obtain inequalities on the expected number of visits of certain state-action pairs in one of them

Intuition of Hard MDPs

- From a high-level perspective, the family of MDPs behave like MABs with $\Theta(HSA)$ arms.
- To gain some intuition, assume that $S = 4$ and consider the MDP in Figure 1

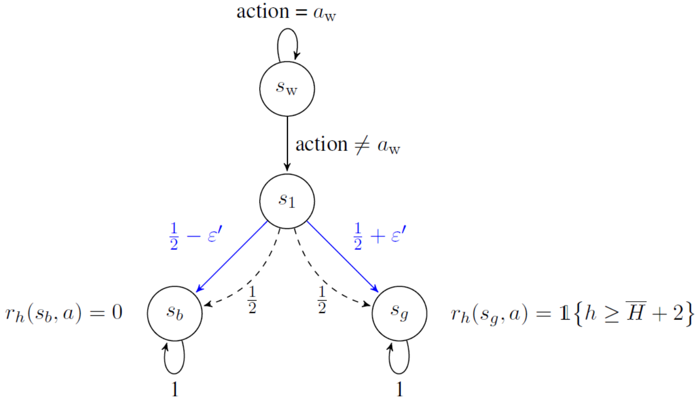

Figure: Illustration of the class of hard MDPs for $S = 4$

Intuition of Hard MDPs

- Can stay in s_{w} up to a stage $\bar{H} < H$
- a^* in state s_1 increases by ϵ the probability of reaching s_g, must be taken at stage h^*
- Optimal policy: choose the right moment $h \in [\bar{H}]$ to leave s_w, then choose a^* in s_1
- Total of $\bar{H}A$ possible choices/"arms", maximal rewards is $\Theta(\bar{H})$
- By analogy with the existing minimax regret bound for MAB, choosing $\bar{H} = \Theta(H)$ yields

$$\Omega(H\sqrt{HAT})$$

Generalization to $S > 4$

Assumption 1.

$S \geq 6, A \geq 2$; $\exists d \in \mathbb{N}$ *s.t.* $S = 3 + \left(A^d - 1\right)/(A-1)$ *(implying* $d = \Theta\left(\log_A S\right)$*)*; $H \geq 3d$.

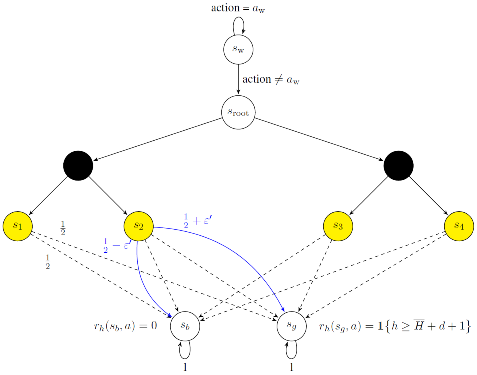

Figure: Illustration of the class of hard MDPs

Generalization to $S > 4$

- $\bar{H} \leq H - d$: a parameter of the class of MDPs.
- $\mathcal{L} = \{s_1, s_2, \ldots, s_L\}$: the set of L leaves of the tree.
- Define an $\mathrm{MDP}\, \mathcal{M}_{(h^*, \ell^*, a^*)}$ for each $(h^*, \ell^*, a^*) \in \{1+d, \ldots, \bar{H}+d\} \times \mathcal{L} \times \mathcal{A}$
- Deterministic transition for each state in the tree.
- The transitions from s_w are given by

$$p_h\left(s_{\mathrm{w}} \mid s_{\mathrm{w}}, a\right) \triangleq \mathbb{I}\left\{a = a_{\mathrm{w}}, h \leq \bar{H}\right\} \quad \text{and} \quad p_h\left(s_{\text{root}} \mid s_{\mathrm{w}}, a\right) \triangleq 1 - p_h\left(s_{\mathrm{w}} \mid s_{\mathrm{w}}, a\right)$$

- After stage $\bar{H}$, the agent has to traverse the tree down to the leaves.

Generalization to $S > 4$

- The transitions from any leaf $s_i \in \mathcal{L}$ are given by

$$p_h(s_g \mid s_i, a) \triangleq \frac{1}{2} + \Delta_{(h^*,\ell^*,a^*)}(h, s_i, a) \quad \text{and} \quad p_h(s_b \mid s_i, a) \triangleq \frac{1}{2} - \Delta_{(h^*,\ell^*,a^*)}(h, s_i, a)$$

- $\Delta_{(h^*,\ell^*,a^*)}(h, s_i, a) \triangleq \mathbb{I}\{(h, s_i, a) = (h^*, s_{\ell^*}, a^*)\} \cdot \varepsilon'$, for some $\varepsilon' \in [0, 1/2]$ that is the second parameter
- The reward function depends only on the state

$$\forall a \in \mathcal{A}, \quad r_h(s, a) \triangleq \mathbb{I}\left\{s = s_g, h \geq \bar{H} + d + 1\right\}$$

- Does not miss any reward if it chooses to stay at s_{w} until stage $\bar{H}$.
- Optimal policy: choose an action a^* at stage h^* and leaf ℓ^*
- Define a reference MDP $\mathcal{M}_0$ where $\Delta_0(h, s_i, a) \triangleq 0$ for all (h, s_i, a)
- Define $\mathcal{C}_{\bar{H},\varepsilon'} \triangleq \{\mathcal{M}_0\} \bigcup \left\{\mathcal{M}_{(h^*,\ell^*,a^*)}\right\}_{(h^*,\ell^*,a^*) \in \{1+d,\ldots,\bar{H}+d\} \times \mathcal{L} \times \mathcal{A}}$.

Change of Distribution Tools

Lemma 1.

Let $\mathcal{M}$ and $\mathcal{M}'$ be two MDPs that are identical except for their transition probabilities. For any stopping time τ with respect to $(\mathcal{F}_H^t)_{t\geq 1}$ that satisfies $\mathbb{P}_{\mathcal{M}}[\tau < \infty] = 1$

$$\mathrm{KL}\left(\mathbb{P}_{\mathcal{M}}^{I_H^\tau}, \mathbb{P}_{\mathcal{M}'}^{I_H^\tau}\right) = \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{h\in[H-1]} \mathbb{E}_{\mathcal{M}}\left[N_{h,s,a}^\tau\right] \mathrm{KL}\left(p_h(\cdot \mid s,a), p_h'(\cdot \mid s,a)\right),$$

where $N_{h,s,a}^\tau \triangleq \sum_{t=1}^\tau \mathbb{I}\left\{(S_h^t, A_h^t) = (s,a)\right\}$.

Lemma 2 (Lemma 1, Garivier et al., 2019).

Consider a measurable space $(\Omega, \mathcal{F})$ equipped with two distributions $\mathbb{P}_1$ and $\mathbb{P}_2$. For any $\mathcal{F}$ -measurable function $Z : \Omega \to [0,1]$, we have

$$\mathrm{KL}\left(\mathbb{P}_1, \mathbb{P}_2\right) \geq \mathrm{kl}\left(\mathbb{E}_1[Z], \mathbb{E}_2[Z]\right),$$

where $\mathbb{E}_1$ and $\mathbb{E}_2$ are the expectations under $\mathbb{P}_1$ and $\mathbb{P}_2$ respectively.

Regret Lower Bound

Using change of distributions between MDPs in the class $\mathcal{C}_{\bar{H},\varepsilon}$ can prove the following result.

Theorem 3.

Under Assumption 1, for any algorithm $\boldsymbol{\pi}$, there exists an MDP $\mathcal{M}_{\boldsymbol{\pi}}$ whose transitions depend on the stage h, such that, for $T \geq HSA$

$$\mathcal{R}_T\left(\boldsymbol{\pi}, \mathcal{M}_{\boldsymbol{\pi}}\right) \geq \frac{1}{48\sqrt{6}}\sqrt{H^3SAT}.$$

Regret of π in $\mathcal{M}_{(h^*,\ell^*,a^*)}$

- The mean reward gathered by π in $\mathcal{M}_{(h^*,\ell^*,a^*)}$ is given by

$$\mathbb{E}_{(h^*,\ell^*,a^*)}\left[\sum_{t=1}^{T}\sum_{h=1}^{H} r_h\left(S_h^t, A_h^t\right)\right] = \sum_{t=1}^{T}\mathbb{E}_{(h^*,\ell^*,a^*)}\left[\sum_{h=\bar{H}+d+1}^{H}\mathbb{I}\left\{S_h^t = s_g\right\}\right]$$
$$= (H-\bar{H}-d)\sum_{t=1}^{T}\mathbb{P}_{(h^*,\ell^*,a^*)}\left[S_{\bar{H}+d+1}^t = s_g\right].$$

- For any $h \in \{1+d,\dots,\bar{H}+d\}$
$\mathbb{P}_{(h^*,\ell^*,a^*)}\left[S_{h+1}^t = s_g\right] =$
$\mathbb{P}_{(h^*,\ell^*,a^*)}\left[S_h^t = s_g\right] + \frac{1}{2}\mathbb{P}_{(h^*,\ell^*,a^*)}\left[S_h^t \in \mathcal{L}\right] + \mathbb{I}\left\{h = h^*\right\}\mathbb{P}_{(h^*,\ell^*,a^*)}\left[S_h^t = s_{\ell^*}, A_h^t = a^*\right]\varepsilon.$
- Indeed, if $S_{h+1}^t = s_g$, we have either $S_h^t = s_g$ or $S_{h+1}^t \in \mathcal{L}$.

Regret of π in $\mathcal{M}_{(h^*,\ell^*,a^*)}$

- The mean reward gathered by π in $\mathcal{M}_{(h^*,\ell^*,a^*)}$ is given by

$$\mathbb{E}_{(h^*,\ell^*,a^*)}\left[\sum_{t=1}^{T}\sum_{h=1}^{H} r_h\left(S_h^t, A_h^t\right)\right] = \sum_{t=1}^{T}\mathbb{E}_{(h^*,\ell^*,a^*)}\left[\sum_{h=\bar{H}+d+1}^{H}\mathbb{I}\left\{S_h^t = s_g\right\}\right]$$
$$= (H-\bar{H}-d)\sum_{t=1}^{T}\mathbb{P}_{(h^*,\ell^*,a^*)}\left[S_{\bar{H}+d+1}^t = s_g\right].$$

- For any $h \in \{1+d,\ldots,\bar{H}+d\}$
$\mathbb{P}_{(h^*,\ell^*,a^*)}\left[S_{h+1}^t = s_g\right] =$
$\mathbb{P}_{(h^*,\ell^*,a^*)}\left[S_h^t = s_g\right] + \frac{1}{2}\mathbb{P}_{(h^*,\ell^*,a^*)}\left[S_h^t \in \mathcal{L}\right] + \mathbb{I}\left\{h = h^*\right\}\mathbb{P}_{(h^*,\ell^*,a^*)}\left[S_h^t = s_{\ell^*}, A_h^t = a^*\right]\varepsilon.$
- Indeed, if $S_{h+1}^t = s_g$, we have either $S_h^t = s_g$ or $S_{h+1}^t \in \mathcal{L}$.

Regret of $\boldsymbol{\pi}$ in $\mathcal{M}_{(h^*,\ell^*,a^*)}$

- Using the facts that $\mathbb{P}_{(h^*,\ell^*,a^*)}\left[S_{1+d}^t = s_g\right] = 0$ and $\sum_{h=1+d}^{\bar{H}+d} \mathbb{P}_{(h^*,\ell^*,a^*)}\left[S_h^t \in \mathcal{L}\right] = 1$

$$\mathbb{P}_{(h^*,\ell^*,a^*)}\left[S_{\bar{H}+d+1}^t = s_g\right] = \sum_{h=1+d}^{\bar{H}+d} \frac{1}{2}\mathbb{P}_{(h^*,\ell^*,a^*)}\left[S_h^t \in \mathcal{L}\right] + \mathbb{I}\left\{h = h^*\right\}\mathbb{P}_{(h^*,\ell^*,a^*)}\left[S_h^t = s_{\ell^*}, A_h^t =\right.$$

$$= \frac{1}{2} + \varepsilon\mathbb{P}_{(h^*,\ell^*,a^*)}\left[S_{h^*}^t = s_{\ell^*}, A_{h^*}^t = a^*\right].$$

- π^*: traverse the tree at step $h^* - d$ then go to the leaf s_{ℓ^*} and performs action a^*
- The optimal value in any of the MDPs is $\rho^* = (H - \bar{H} - d)(1/2 + \varepsilon)$
- The regret of $\boldsymbol{\pi}$ in $\mathcal{M}_{(h^*,\ell^*,a^*)}$ is then

$$\mathcal{R}_T\left(\boldsymbol{\pi}, \mathcal{M}_{(h^*,\ell^*,a^*)}\right) = T(H - \bar{H} - d)\varepsilon\left(1 - \frac{1}{T}\mathbb{E}_{(h^*,\ell^*,a^*)}\left[N_{(h^*,\ell^*,a^*)}\right]\right),$$

where $N_{(h^*,\ell^*,a^*)}^T = \sum_{t=1}^T \mathbb{I}\left\{S_{h^*}^t = s_{\ell^*}, A_{h^*}^t = a^*\right\}$.

Maximum regret of π over all possible $\mathcal{M}_{(h^*,\ell^*,a^*)}$

- We first lower bound the maximum of the regret by the mean over all instances

$$\max_{(h^*,\ell^*,a^*)} \mathcal{R}_T\left(\boldsymbol{\pi}, \mathcal{M}_{(h^*,\ell^*,a^*)}\right) \geq \frac{1}{\bar{H}LA} \sum_{(h^*,\ell^*,a^*)} \mathcal{R}_T\left(\boldsymbol{\pi}, \mathcal{M}_{(h^*,\ell^*,a^*)}\right)$$

$$\geq T(H-\bar{H}-d)\varepsilon\left(1-\frac{1}{\bar{H}LAT}\sum_{(h^*,\ell^*,a^*)} \mathbb{E}_{(h^*,\ell^*,a^*)}\left[N^T_{(h^*,\ell^*,a^*)}\right]\right)$$

- Need an upper bound on $\sum_{(h^*,\ell^*,a^*)} \mathbb{E}_{(h^*,\ell^*,a^*)}\left[N^T_{(h^*,\ell^*,a^*)}\right]$
- Relate each expectation to the expectation of the same quantity under $\mathcal{M}_0$

Upper bound on $\sum \mathbb{E}_{(h^*,\ell^*,a^*)}\left[N_{(h^*,\ell^*,a^*)}\right]$

- Since $N^T_{(h^*,\ell^*,a^*)}/T \in [0,1]$, Lemma gives us

$$\mathrm{kl}\left(\frac{1}{T}\mathbb{E}_0\left[N^T_{(h^*,\ell^*,a^*)}, \frac{1}{T}\mathbb{E}_{(h^*,\ell^*,a^*)}\left[N^T_{(h^*,\ell^*,a^*)}\right]\right]\right) \leq \mathrm{KL}\left(\mathbb{P}_0^{I_H^T}, \mathbb{P}_{(h^*,\ell^*,a^*)}^{I_H^T}\right)$$

- By Pinsker's inequality, $(p-q)^2 \leq (1/2)\mathrm{kl}(p,q)$, it implies

$$\frac{1}{T}\mathbb{E}_{(h^*,\ell^*,a^*)}\left[N_{(h^*,\ell^*,a^*)}\right] \leq \frac{1}{T}\mathbb{E}_0\left[N^T_{(h^*,\ell^*,a^*)}\right] + \sqrt{\frac{1}{2}\mathrm{KL}\left(\mathbb{P}_0^{I_H^T}, \mathbb{P}_{(h^*,\ell^*,a^*)}^{I_I^T}\right)}$$

- Since $\mathcal{M}_0$ and $\mathcal{M}_{(h^*,\ell^*,a^*)}$ only differ at stage h^* when $(s,a) = (s_{\ell^*}, a^*)$, Lemma gives

$$\mathrm{KL}\left(\mathbb{P}_0^{I_H^T}, \mathbb{P}_{(h^*,\ell^*,a^*)}^{I_H^T}\right) = \mathbb{E}_0\left[N^T_{(h^*,\ell^*,a^*)}\right]\mathrm{kl}(1/2, 1/2+\varepsilon)$$

Upper bound on $\sum \mathbb{E}_{(h^*,\ell^*,a^*)}\left[N_{(h^*,\ell^*,a^*)}\right]$

- For $\varepsilon \leq 1/4$, we have $\mathrm{kl}(1/2, 1/2+\varepsilon) \leq 4\varepsilon^2$ (to be checked)

$$\frac{1}{T}\mathbb{E}_{(h^*,\ell^*,a^*)}\left[N_{(h^*,\ell^*,a^*)}\right] \leq \frac{1}{T}\mathbb{E}_0\left[N^T_{(h^*,\ell^*,a^*)}\right] + \sqrt{2}\varepsilon\sqrt{\mathbb{E}_0\left[N^T_{(h^*,\ell^*,a^*)}\right]}$$

- The sum of $N^T_{(h^*,\ell^*,a^*)}$ over all instances $(h^*,\ell^*,a^*) \in \{1+d,\ldots,\bar{H}+d\} \times \mathcal{L} \times \mathcal{A}$ is

$$\sum_{(h^*,\ell^*,a^*)} N^T_{(h^*,\ell^*,a^*)} = \sum_{t=1}^{T}\sum_{h^*=1+d}^{\bar{H}+d} \mathbb{I}\left\{S^t_{h^*} \in \mathcal{L}\right\} = T$$

- Summing over all instances and using the Cauchy-Schwartz inequality

$$\begin{aligned}\frac{1}{T}\sum_{(h^*,\ell^*,a^*)}\mathbb{E}_{(h^*,\ell^*,a^*)}\left[N^T_{(h^*,\ell^*,a^*)}\right] &\leq 1 + \sqrt{2}\varepsilon\sum_{(h^*,\ell^*,a^*)}\sqrt{\mathbb{E}_0\left[N_{(h^*,\ell^*,a^*)}\right]}\\ &\leq 1+\sqrt{2}\varepsilon\sqrt{\bar{H}LAT}.\end{aligned}$$

Discussion

- The proof uses Assumption 1 stating that
  - there exists an integer d such that $S = 3 + \left(A^d - 1\right)/(A - 1)$
  - $H \geq 3d$,
- They can be relaxed to the case we cannot build a full tree.
- The proof can be easily adapted to stationary case and recover $\Omega\left(\sqrt{H^2SAT}\right)$.
- The author also proves a sample complexity lower bound for best policy identification in a non-stationary MDP.
- The proof relies on the same construction of hard MDPs.