The Fundamental Limits of Imitation Learning

Tian Xu

xut@lamda.nju.edu.cn

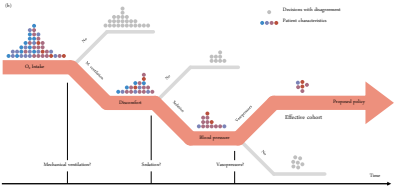
Nanjing University

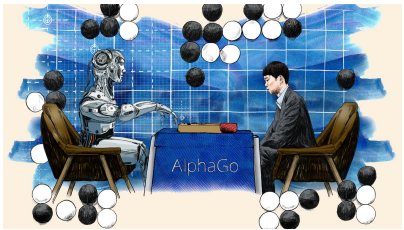
Mainly based on:
Toward the Fundamental Limits of Imitation Learning.

March 26, 2021

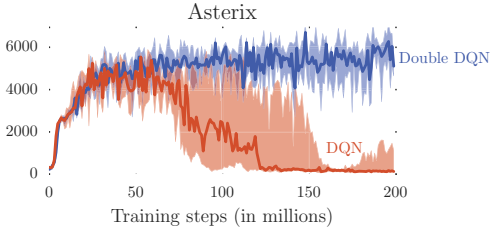
Outline

Reinforcement Learning (RL)





RL Challenges



Double DQN requires million samples to solve Atari games [van Hasselt et al., 2016].

Robot directly learns from human demonstrations.

- RL aims to learn the (near-) optimal decisions from interactions with environments
  - It often requires a large amount of samples.
  - It's hard to design proper reward function for each particular task.
- In some real-world scenarios, it is easy to obtain expert-level demonstrations.

Imitation Learning (IL)

Learner

$\pi(a|s)$

Expert

$(s, a) \sim \pi_{\mathrm{E}}$

- Given trajectories $D = \{(s_1^i, a_1^i, s_2^i, \cdots, s_H^i, a_H^i)\}_{i=1}^m$ collected by expert policy π_{E}, which is (near-) optimal.
- Agent directly learns a policy from D without explicit rewards.
- IL does not rely on trails-and-errors and could be more sample-efficient .

Markov Decision Process

- Consider a finite episodic Markov Decision Process $\left(\mathcal{S}, \mathcal{A}, H, \{P_h\}_{h\in[H]}, \{r_h\}_{h\in[H]}, \rho\right)$.
  - $\mathcal{S}$ and $\mathcal{A}$ are the state and action space, respectively.
  - $r_h(s,a) \in [0,1]$ is deterministic reward received after taking the action a in state s at step h.
  - $P_h(s'|s,a)$ specifies the transition probability of s' conditioned on s and a at step h.
  - H is the horizon length.
  - The initial state s_1 is sampled from the initial state distribution ρ.

Markov Decision Process

- A deterministic policy is a collection of functions $\pi_h : \mathcal{S} \to \mathcal{A}$ for all $h \in [H]$. We use Π_{det} to denote the set of all deterministic policies.
- We assume that the expert policy is deterministic and optimal.
- The policy value $J(\pi) = \mathbb{E}\left[\sum_{h=1}^{H} r_h(s_h, a_h)\right]$.

Settings

- There are mainly three settings in IL.
  - No-interaction: Provided with expert dataset, the learner is not allowed to interact with the MDP.
  - Known-transition: Besides expert dataset, the learner additionally knowns the MDP transition function.
  - Active: Without expert dataset in advance, the learner is allowed to interact with the MDP for m episodes and is provided access to an oracle which outputs the expert action $\pi^*(s)$ at the learner's current state s.
- Intuitively, the hardness of problems under different settings: No-interaction $\geq$ Known-transition, No-interaction $\geq$ ($\asymp$) Active.

Overview

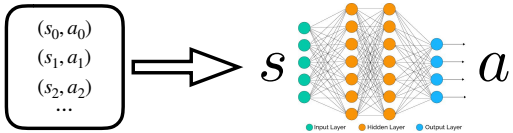
- In IL, our objective is to minimize the policy value gap:

$$\min_{\pi} J(\pi_E) - J(\pi) \quad \Longleftrightarrow \quad \max_{\pi} J(\pi)$$

- There are mainly two classes of methods: behavioral cloning (BC) [Pomerleau, 1991] and adversarial-based imitation [Abbeel and Ng, 2004, Ho and Ermon, 2016].
  - BC: mimic by action distribution matching with supervised learning.
  - Adversarial-based imitation: firstly infer the reward function, then learn a (sub-) optimal policy with the recovered reward.

Outline

Behavioral Cloning (BC)



- Given expert demonstrations: $D = \{(s_1^i, a_1^i, s_2^i, \cdots, s_H^i, a_H^i)\}_{i=1}^m$.
- BC reduces IL to supervised learning:
  - BC firstly splits trajectories into labeled data with states as inputs and actions as targets.
  - Then BC learns a mapping (e.g., neural networks) from state space to action space via any supervised learning methods.

Behavioral Cloning

- Mathematically, BC learns a policy to minimize the population $0-1$ risk.

$$\mathcal{L}_{\text{pop}}\left(\widehat{\pi}, \pi^*\right) = \frac{1}{H}\sum_{t=1}^{H}\mathbb{E}_{s_t\sim f^t_{\pi^*}}\left[\mathbb{E}_{a\sim\widehat{\pi}_t(\cdot|s_t)}\left[\mathbb{I}\left(a \neq \pi^*_t\left(s_t\right)\right)\right]\right],$$

  where $f^t_{\pi^*}(s) = \Pr_{\pi^*}(s_t = s)$.

- With expert dataset D, BC optimizes the following empirical risk.

$$\mathcal{L}_{\text{emp}}\left(\widehat{\pi}, \pi^*\right) = \frac{1}{H}\sum_{t=1}^{H}\mathbb{E}_{s_t\sim f^t_{D}}\left[\mathbb{E}_{a\sim\widehat{\pi}_t(\cdot|s_t)}\left[\mathbb{I}\left(a \neq \pi^*_t\left(s_t\right)\right)\right]\right],$$

  where $f^t_D(s) = \frac{\sum_{i=1}^{m}\mathbb{I}(s^i_t=s)}{m}$.

Behavioral Cloning

- BC does not need to interact with the MDP and optimizes the empirical risk in an offline manner.
- Given expert dataset D, we define $\Pi_{\mathrm{mimic}}(D)$ as the set of policies which are compatible with D.

$$\Pi_{\mathrm{mimic}}(D) \triangleq \left\{\pi \in \Pi : \forall t \in [H], s \in \mathcal{S}_t(D), \pi_t(\cdot \mid s) = \delta_{\pi_t^*(s)}\right\},$$

  where $\mathcal{S}_t(D) = \{s_t^i\}_{i=1}^m$ and δ_a is a distribution over $\mathcal{A}$ which puts all probability mass on a.
- It is easy to check that $\forall \hat{\pi} \in \Pi_{\mathrm{mimic}}(D)$, $\mathcal{L}_{\mathsf{emp}}(\pi, \pi^*) = 0$, meaning that the solution of BC lies in $\Pi_{\mathrm{mimic}}(D)$.

Upper Bound of BC

Theorem 1

Consider any policy $\hat{\pi} \in \Pi_{\mathrm{mimic}}(D)$,

- *The expected sub-optimality is bounded by,*

$$J(\pi^*) - \mathbb{E}[J(\widehat{\pi})] \lesssim \min\left\{H, \frac{|\mathcal{S}|H^2}{m}\right\}$$

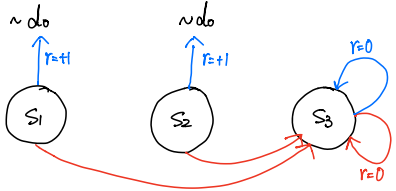
- *For any* $\delta \in (0, \min\{1, H/10\}]$, *w.p.* $\geq 1-\delta$, *the sub-optimality is bounded by,*

$$J(\pi^*) - J(\widehat{\pi}) \lesssim \frac{|\mathcal{S}|H^2}{m} + \frac{\sqrt{|\mathcal{S}|}H^2 \log(H/\delta)}{m}$$

Upper bound of BC

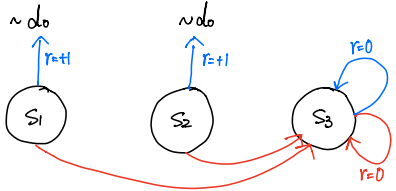
- BC enjoys a convergence rate of $\frac{1}{m}$, which is rare in decision-making tasks.
- The sub-optimality of BC grows quadratically w.r.t the horizon, which is referred to the phenomenon of compounding error.

An Illusrating Example



- Consider the three-state MDP. There are two actions $\mathcal{A} = \{B, R\}$. $d_0 = \left(\frac{1}{m+1}, 1 - \frac{1}{m+1}, 0\right)$.
- The expert policy $\pi_t^*(s) = B, \forall s \in \mathcal{S}, \forall t \in [H]$ and $J(\pi^*) = H$.
- The expert dataset $D = \{(s_t^i, a_t^i)_{t=1}^H\}_{i=1}^m$ where $s_t^i \overset{i.i.d.}{\sim} d_0, \ \forall t \in [H]$.

An Illusrating Example



- For each step $t \in [H]$, with a constant probability $((1 - \frac{1}{m+1})^m \geq e^{-1})$, s_1 is not covered in $S_t(D)$. The learner $\hat{\pi}$ does not know how to act when visiting s_1 at step t.
- For $\hat{\pi}$, at step t, if $\hat{\pi}$ does not make any mistakes before (or it has been transited into s_3), w.p. $\gtrsim \frac{1}{m+1}$, $\hat{\pi}$ encounters s_1, makes a mistake and suffers a sub-optimality of $H - t$.
- The total sub-optimality $\gtrsim \sum_{t=1}^{H} (1 - \frac{1}{m+1})^{t-1} \frac{1}{m+1} (H - t) \gtrsim \frac{H^2}{m}$.

Proof Idea

- We first bridge the connection between sub-optimality and the population $0-1$ risk.
- For each $\hat{\pi} \in \Pi_{\mathrm{mimic}}$, we upper bound the population risk $\mathcal{L}_{\mathrm{pop}}(\hat{\pi}, \pi^*)$ with a missing mass argument.

Analysis

Lemma 1 ([Ross et al., 2011])

For all policy $\hat{\pi}$, we have

$$J(\pi^*) - J(\hat{\pi}) \leq H^2 \mathcal{L}_{\text{pop}}(\hat{\pi}, \pi^*),$$

where $\mathcal{L}_{\text{pop}}(\hat{\pi}, \pi^*) = \frac{1}{H} \sum_{t=1}^{H} \mathbb{E}_{s_t \sim f_{\pi^*}^t} \left[\mathbb{E}_{a \sim \hat{\pi}_t(\cdot|s_t)} \left[\mathbb{I}(a \neq \pi_t^*(s_t)) \right] \right]$

Proof

- We first upper bound the sub-optimality with the total variation between two occupancy measures.

$$J(\pi^*) - J(\hat{\pi}) \le 2 \sum_{t=1}^{H} D_{\mathrm{TV}}(P_t^{\pi^*}, P_t^{\hat{\pi}}),$$

where $P_t^{\pi}(s,a) = \Pr_{\pi}(s_t = s, a_t = a)$.

- Then we derive a recursion formula of $D_{\mathrm{TV}}(P_t^{\pi^*}, P_t^{\hat{\pi}})$.

$$\underbrace{D_{\mathrm{TV}}(P_t^{\pi^*}, P_t^{\hat{\pi}})}_{\text{Accumulated error to step } t} \le \underbrace{\mathbb{E}_{s_t \sim f_{\pi^*}^t}\left[\mathbb{E}_{a \sim \widehat{\pi}_t(\cdot|s_t)}\left[\mathbb{I}\left(a \neq \pi_t^*(s_t)\right)\right]\right]}_{\text{Error at step } t} + \underbrace{D_{\mathrm{TV}}(P_{t-1}^{\pi^*}, P_{t-1}^{\hat{\pi}})}_{\text{Accumulated error to step } t-1}. \tag{1}$$

- Expanding the above formula yields the desired result.

Lemma 2

For each $\hat{\pi} \in \Pi_{\text{mimic}}(D)$, where $\Pi_{\text{mimic}}(D)$ is the set of policies which are compatible with D.

- *The expected 0-1 population risk of $\hat{\pi}$ has an upper bound.*

$$\mathbb{E}\left[\mathcal{L}_{\text{pop}}(\hat{\pi}, \pi^*)\right] \leq \frac{4}{9}\frac{|\mathcal{S}|}{m}.$$

- $\forall \delta \in (0, \min\{1, H/10\})$, *w.p.* $1-\delta$, *we have*

$$\mathcal{L}_{\text{pop}}(\hat{\pi}, \pi^*) \leq \frac{4|\mathcal{S}|}{9m} + \frac{3\sqrt{|\mathcal{S}|}\log(H/\delta)}{m}.$$

Proof

- For each $\hat{\pi} \in \Pi_{\text{mimic}}(D)$, $\mathcal{L}_{\text{pop}}(\hat{\pi}, \pi^*) = \frac{1}{H}\sum_{t=1}^{H} \mathbb{E}_{s_t \sim f_{\pi^*}^t}\left[\mathbb{E}_{a\sim\hat{\pi}_t(\cdot|s_t)}\left[\mathbb{I}\left(a \neq \pi_t^*\left(s_t\right)\right)\right]\right] \leq \frac{1}{H}\sum_{t=1}^{H} \mathbb{E}_{s_t \sim f_{\pi^*}^t}\left[\mathbb{I}(s_t \notin S_t(D))\right] = \frac{1}{H}\sum_{t=1}^{H}\sum_{s\in\mathcal{S}} f_{\pi^*}^t(s)\mathbb{I}(s_t \notin S_t(D))$.
- For step t, we take expectation w.r.t. D and obtain that

$$\mathbb{E}\left[\sum_{s\in\mathcal{S}} f_{\pi^*}^t(s)\mathbb{I}(s_t \notin S_t(D))\right] = \sum_{s\in\mathcal{S}} f_{\pi^*}^t(s)\Pr(s_t \notin S_t(D)) = \sum_{s\in\mathcal{S}} f_{\pi^*}^t(s)(1-f_{\pi^*}^t(s))^m \overset{(1)}{\leq} \frac{4|\mathcal{S}|}{9m},$$

  where inequality (1) follows that $\max_{x\in[0,1]} x(1-x)^m \leq \frac{4}{9m}$.
- The term $\mathbb{E}\left[\sum_{s\in\mathcal{S}} f_{\pi^*}^t(s)\mathbb{I}(s_t \notin S_t(D))\right]$ is called "missing mass" which is the probability mass contributed by the items uncovered in dataset.

Missing mass

Definition 3 (Missing mass)

Let P be the probability distribution over $\mathcal{X}$. Suppose that X^m are i.i.d. drawn from P. Let $n_x(X^m) = \sum_{i=1}^{m} \mathbb{I}(X^i = x)$ denote the number of times that the symbol x is observed in X^m. Then the missing mass $m_0(p, X^m) = \sum_{x \in \mathcal{X}} p(x) \mathbb{I}\left(n_x(X^m) = 0\right)$ which is defined as the probability mass contributed by symbols are uncovered in X^m.

Missing mass

Theorem 4 ([McAllester and Ortiz, 2003])

Consider an arbitrary distribution P on X, and let $X^m \overset{i.i.d.}{\sim} P$ be a dataset of m samples drawn i.i.d. from P. Consider any $\delta \in (0, \frac{1}{10}]$. Then, w.p. $1-\delta$,

$$\mathfrak{m}_0\left(\nu, X^m\right) - \mathbb{E}\left[\mathfrak{m}_0\left(\nu, X^m\right)\right] \leq \frac{3\sqrt{|\mathcal{X}|}\log(1/\delta)}{m}.$$

Proof

- We have obtained the upper bound of the expected missing mass.

$$\mathbb{E}\left[\sum_{s\in\mathcal{S}} f^t_{\pi^*}(s)\mathbb{I}(s_t \notin S_t(D))\right] \leq \frac{4|\mathcal{S}|}{9m}.$$

- With the concentration argument for missing mass, we obtain the high probability bound. For any $\delta \in (0, \frac{1}{10}]$, w.p. $\geq 1-\delta$,

$$\sum_{s\in\mathcal{S}} f^t_{\pi^*}(s)\mathbb{I}(s_t \notin S_t(D) \leq \frac{4|\mathcal{S}|}{9m} + \frac{3\sqrt{|\mathcal{S}|}\log(H/\delta)}{m}.$$

Outline

Lower bound

Theorem 5

Under the no-interaction setting, for any learner $\hat{\pi}$, there exists an MDP $\mathcal{M}$ and a deterministic expert policy π^* such that the expected sub-optimality of the learner is lower bounded by,

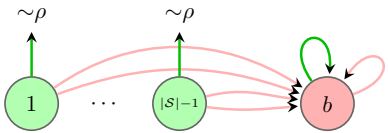
$$J_{\mathcal{M}}(\pi^*) - \mathbb{E}\left[J_{\mathcal{M}}(\widehat{\pi})\right] \gtrsim \min\left\{H, |\mathcal{S}|H^2/m\right\}$$

Furthermore this lower bound applies even when the learner operates in the active setting.

Lower bound

- The upper bound of BC meets the lower bound which implies that BC is already minimax optimal under the no-interaction setting in IL.
- This lower bound also holds in the active setting which suggests that the ability to actively query the expert does not reduce the hardness of problems.

The hard MDPs



- There are $|\mathcal{S}|$ states and $|\mathcal{A}|$ actions. At each state, there is an optimal action (the green arrow). The state b is a bad and absorbing state.
- The initial state distribution $\rho = \left(\frac{1}{m+1}, \cdots, \frac{1}{m+1}, 1 - \frac{|\mathcal{S}|-2}{m+1}, 0\right)$.
- At each state except b, when the agent takes the optimal action, it will be renewed according ρ and get $+1$ reward. Otherwise, it will be transited to b and can not get rewards anymore.

Proof

- For any learner $\hat{\pi}$, our target is to lower bound $\max_{\mathcal{M},\pi^*} J_{\mathcal{M}}(\pi^*) - \mathbb{E}\left[J_{\mathcal{M}}(\hat{\pi}(D))\right]$.
- It suffices to lower bound the Bayes expected sub-optimality $\mathbb{E}_{(\pi^*,\mathcal{M})\sim\mathcal{P}}\left[J_{\mathcal{M}}(\pi^*) - \mathbb{E}\left[J_{\mathcal{M}}(\hat{\pi}(D))\right]\right]$, where $\mathcal{P}$ is a joint distribution over MDPs and expert policies.
- The construction of $\mathcal{P}$: $\pi^* \sim \mathrm{Unif}\left(\Pi_{\mathrm{det}}\right)$ and $\mathcal{M} = \mathcal{M}[\pi^*]$ is determined by the MDP template constructed above.

Proof

- The correlation between $(\pi^*, \mathcal{M}[\pi^*])$ and D.
- Conditioned on $(\pi^*, \mathcal{M}[\pi^*])$, D is obtained by rolling out π^* on $\mathcal{M}[\pi^*]$.
- Conversely, conditioned on D, $(\pi^*, \mathcal{M}[\pi^*]) \sim \mathcal{P}(D)$ where $\pi^* \sim \mathrm{Unif}\left(\Pi_{\mathrm{mimic}}(D)\right)$ and $\mathcal{M} = \mathcal{M}[\pi^*]$.
- Then we have

$$\mathbb{E}_{(\pi^*, \mathcal{M}) \sim \mathcal{P}}\left[H - \mathbb{E}\left[J_{\mathcal{M}}(\widehat{\pi}(D))\right]\right] = \mathbb{E}\left[\mathbb{E}_{(\pi^*, \mathcal{M}) \sim \mathcal{P}(D)}\left[H - J_{\mathcal{M}}(\widehat{\pi}(D))\right]\right]$$

Auxiliary Lemma

Lemma 6

Define the stopping time τ as the first time t that the learner encounters a state $s_t \notin S_t(D)$ that has not been visited in D at time t. That is,

$$\tau = \begin{cases} \inf\{t : s_t \notin \mathcal{S}_t(D) \cup \{b\}\} & \exists t : s_t \notin \mathcal{S}_t(D) \cup \{b\} \\ H & \text{otherwise} \end{cases}$$

Then conditioned on the dataset D, we have

$$\mathbb{E}_{(\pi^*,\mathcal{M})\sim\mathcal{P}(D)}\left[J\left(\pi^*\right) - \mathbb{E}[J(\widehat{\pi}(D))]\right] \geq \left(1 - \frac{1}{|\mathcal{A}|}\right)\mathbb{E}_{(\pi^*,\mathcal{M})\sim\mathcal{P}(D)}\left[\mathbb{E}_{\widehat{\pi}(D)}[H - \tau]\right].$$

When $\hat{\pi}$ encounters an uncovered state at τ, with probability $\geq (1 - \frac{1}{|\mathcal{A}|})$, $\hat{\pi}$ takes an non-optimal action and suffers a sub-optimality of $H - \tau$.

Proof

We apply the above useful lemma and obtain that

$$\begin{aligned}\mathbb{E}_{(\pi^*,\mathcal{M})\sim\mathcal{P}}\left[J\left(\pi^*\right)-\mathbb{E}[J(\widehat{\pi}(D))]\right] &\geq \left(1-\frac{1}{|\mathcal{A}|}\right)\mathbb{E}\left[\mathbb{E}_{(\pi^*,\mathcal{M})\sim\mathcal{P}(D)}\left[\mathbb{E}_{\widehat{\pi}(D)}[H-\tau]\right]\right], \\ &\overset{(1)}{\geq}\left(1-\frac{1}{|\mathcal{A}|}\right)\frac{H}{2}\mathbb{E}\left[\mathbb{E}_{(\pi^*,\mathcal{M})\sim\mathcal{P}(D)}\left[\Pr_{\widehat{\pi}(D)}[\tau\leq\lfloor H/2\rfloor]\right]\right] \\ &=\left(1-\frac{1}{|\mathcal{A}|}\right)\frac{H}{2}\mathbb{E}_{(\pi^*,\mathcal{M})\sim\mathcal{P}}\left[\mathbb{E}\left[\Pr_{\widehat{\pi}(D)}[\tau\leq\lfloor H/2\rfloor]\right]\right],\end{aligned}$$

where inequality (1) follows the Markov's inequality.

Auxiliary Lemma

Lemma 7

For any learner $\hat{\pi}$,

$$\mathbb{E}_{(\pi^*,\mathcal{M})\sim\mathcal{P}}\left[\mathbb{E}\left[\Pr_{\widehat{\pi}(D)}[\tau \le \lfloor H/2 \rfloor]\right]\right] \ge 1 - \left(1 - \frac{|\mathcal{S}|-2}{e(N+1)}\right)^{\lfloor H/2 \rfloor} \gtrsim \min\left\{1, \frac{|\mathcal{S}|H}{N}\right\}$$

Outline

Proof of Lemma 8

Lemma 8

Define the stopping time τ as the first time t that the learner encounters a state $s_t \notin S_t(D)$ that has not been visited in D at time t. That is,

$$\tau = \begin{cases} \inf\{t : s_t \notin \mathcal{S}_t(D) \cup \{b\}\} & \exists t : s_t \notin \mathcal{S}_t(D) \cup \{b\} \\ H & \text{otherwise} \end{cases}$$

Then conditioned on the dataset D, we have

$$\mathbb{E}_{(\pi^*,\mathcal{M})\sim\mathcal{P}(D)}\left[J\left(\pi^*\right) - \mathbb{E}[J(\widehat{\pi}(D))]\right] \geq \left(1 - \frac{1}{|\mathcal{A}|}\right)\mathbb{E}_{(\pi^*,\mathcal{M})\sim\mathcal{P}(D)}\left[\mathbb{E}_{\widehat{\pi}(D)}[H - \tau]\right].$$

Proof of Lemma 8

- We define an useful random variable τ_b to be the first time the learner first encounters the state b.

$$\tau_b = \begin{cases} \inf\{t : s_t = b\} & \exists t : s_t = b \\ H + 1 & \text{otherwise} \end{cases}$$

- We have $H - \mathbb{E}_{(\pi^*, \mathcal{M}) \sim \mathcal{P}(D)}[J(\widehat{\pi})] = \mathbb{E}_{(\pi^*, \mathcal{M}) \sim \mathcal{P}(D)}\left[\mathbb{E}_{\widehat{\pi}}\left[H - \tau_b + 1\right]\right]$.
- For each $t \in [H]$, we consider the probability $\Pr_{\hat{\pi}}(\tau_b = t + 1)$.

$$\begin{aligned} \Pr_{\hat{\pi}}(\tau_b = t+1) &\geq \Pr_{\hat{\pi}}(\tau_b = t+1, \tau = t) = \sum_{s \in \mathcal{S}} \Pr_{\hat{\pi}}(\tau_b = t+1, \tau = t, s_t = s) \\ &= \sum_{s \in \mathcal{S}} \Pr_{\hat{\pi}}(\tau_b = t+1 | \tau = t, s_t = s) \Pr_{\hat{\pi}}(\tau = t, s_t = s) \\ &= \sum_{s \in \mathcal{S}} \left(1 - \widehat{\pi}_t\left(\pi_t^*(s) \mid s\right)\right) \Pr_{\hat{\pi}}(\tau = t, s_t = s) \end{aligned}$$

Proof of Lemma 8

- Taking expectation yields that

$$\begin{aligned}
&\mathbb{E}_{(\pi^*,\mathcal{M})\sim\mathcal{P}(D)}\left[\left(1-\widehat{\pi}_t\left(\pi_t^*(s)\mid s\right)\right)\Pr_{\hat{\pi}}\left(\tau=t,s_t=s\right)\right]\\
\overset{(1)}{=}&\ \mathbb{E}_{(\pi^*,\mathcal{M})\sim\mathcal{P}(D)}\left[\left(1-\widehat{\pi}_t\left(\pi_t^*(s)\mid s\right)\right)\right]\mathbb{E}_{(\pi^*,\mathcal{M})\sim\mathcal{P}(D)}\left[\Pr_{\hat{\pi}}\left(\tau=t,s_t=s\right)\right]\\
\overset{(2)}{=}&\ \left(1-\frac{1}{|\mathcal{A}|}\right)\mathbb{E}_{(\pi^*,\mathcal{M})\sim\mathcal{P}(D)}\left[\Pr_{\hat{\pi}}\left(\tau=t,s_t=s\right)\right],
\end{aligned}$$

Equation (1) holds since $\pi_1^*,\cdots,\pi_{t-1}^*$ and π_t^* are independent. Equation (2) follows that at states uncovered in D, the expert action is uniformly drawn from the action space.

Proof of Lemma 8

- Taking summation over s yields
  $\mathbb{E}_{(\pi^*,\mathcal{M})\sim\mathcal{P}(D)}\left[\Pr_{\hat{\pi}}\left(\tau_b = t+1\right)\right] \geq (1-\frac{1}{|\mathcal{A}|})\mathbb{E}_{(\pi^*,\mathcal{M})\sim\mathcal{P}(D)}\left[\Pr_{\hat{\pi}}\left(\tau = t\right)\right]$.
- For the sub-optimality, we have

$$\begin{aligned} H - \mathbb{E}_{(\pi^*,\mathcal{M})\sim\mathcal{P}(D)}\left[J(\widehat{\pi})\right] &= \mathbb{E}_{(\pi^*,\mathcal{M})\sim\mathcal{P}(D)}\left[\mathbb{E}_{\widehat{\pi}}\left[H-\tau_b+1\right]\right] \\ &\geq (1-\frac{1}{|\mathcal{A}|})\mathbb{E}_{(\pi^*,\mathcal{M})\sim\mathcal{P}(D)}\left[\mathbb{E}_{\widehat{\pi}}\left[H-\tau\right]\right] \end{aligned}$$

Proof of Lemma 9

Lemma 9

For any learner $\hat{\pi}$,

$$\mathbb{E}_{(\pi^*,\mathcal{M})\sim\mathcal{P}}\left[\mathbb{E}\left[\Pr_{\widehat{\pi}(D)}[\tau \le \lfloor H/2 \rfloor]\right]\right] \ge 1 - \left(1 - \frac{|\mathcal{S}|-2}{e(N+1)}\right)^{\lfloor H/2 \rfloor} \gtrsim \min\left\{1, \frac{|\mathcal{S}|H}{N}\right\}$$

Proof of Lemma 9

- For each $t \in [H]$, we consider $\Pr_{\widehat{\pi}(D)}[\tau = t]$.

$$\begin{aligned}
\Pr_{\widehat{\pi}(D)}[\tau = t] &= \Pr_{\widehat{\pi}(D)}\left[s_t \notin \mathcal{S}_t(D) \cup \{b\}, \forall t' < t, s_{t'} \in \mathcal{S}_{t'}(D) \cup \{b\}\right] \\
&= \Pr_{\widehat{\pi}(D)}\left[s_t \notin \mathcal{S}_t(D) \cup \{b\}, \forall t' < t, s_{t'} \in \mathcal{S}_{t'}(D) \backslash \{b\}\right] \\
&= \Pr_{\pi^*}\left[s_t \notin \mathcal{S}_t(D) \cup \{b\}, \forall t' < t, s_{t'} \in \mathcal{S}_{t'}(D) \backslash \{b\}\right] \\
&= \left(1 - \rho\left(\mathcal{S}_t(D) \backslash \{b\}\right)\right) \prod_{t'=1}^{t-1} \rho\left(\mathcal{S}_{t'}(D) \backslash \{b\}\right)
\end{aligned}$$

- $\Pr_{\widehat{\pi}(D)}[\tau \leq \lfloor H/2 \rfloor] = 1 - \prod_{t=1}^{\lfloor H/2 \rfloor} \rho\left(\mathcal{S}_t(D) \backslash \{b\}\right)$.

Proof of Lemma 9

- We take expectation w.r.t the expert dataset.

$$\begin{aligned}
\mathbb{E}\left[\prod_{t=1}^{\lfloor H/2\rfloor}\rho\left(\mathcal{S}_t(D)\backslash\{b\}\right)\right] &= \prod_{t=1}^{\lfloor H/2\rfloor}\mathbb{E}\left[\rho\left(\mathcal{S}_t(D)\backslash\{b\}\right)\right] \\
&= \prod_{t=1}^{\lfloor H/2\rfloor}\sum_{s}\rho(s)\Pr(s\in\mathcal{S}_t(D)\backslash\{b\}) \\
&= \prod_{t=1}^{\lfloor H/2\rfloor}\sum_{s}\rho(s)(1-(1-\rho(s))^m) \\
&= (1-\gamma)^{\lfloor H/2\rfloor},
\end{aligned}$$

  where $\gamma = \sum_s \rho(s)(1-\rho(s))^m$ is the missing mass.

- $\mathbb{E}\left[\Pr_{\widehat{\pi}(D)}[\tau \le \lfloor H/2\rfloor]\right] = 1-(1-\gamma)^{\lfloor H/2\rfloor}$.

Proof of Lemma 9

- Then we lower bound the missing mass.

$$\gamma = \sum_{s\in\mathcal{S}} \rho(s)(1-\rho(s))^m \geq \frac{|\mathcal{S}|-2}{m+1}\left(1-\frac{1}{m+1}\right)^m \overset{(1)}{\geq} \frac{|\mathcal{S}|-2}{e(m+1)}.$$

  where inequality (1) follows that $(1+\frac{1}{m})^m \leq e$.

- $\mathbb{E}_{(\pi^*,\mathcal{M})\sim\mathcal{P}}\left[\mathbb{E}\left[\Pr_{\widehat{\pi}}[\tau \leq \lfloor H/2 \rfloor]\right]\right] = 1-(1-\gamma)^{\lfloor H/2 \rfloor} \geq 1-\left(1-\frac{|\mathcal{S}|-2}{e(N+1)}\right)^{\lfloor H/2 \rfloor} \overset{(2)}{\gtrsim} \min\left\{1, \frac{|\mathcal{S}|H}{N}\right\}$, where inequality (2) follows that $(1+\frac{x}{N})^N \leq \exp(x) \leq 1+\frac{x}{2}$ when $x \in (-1,0)$.

Bibliography

[Abbeel and Ng, 2004] Abbeel, P. and Ng, A. Y. (2004).

Apprenticeship learning via inverse reinforcement learning.

In Machine Learning, Proceedings of the Twenty-first International Conference (ICML 2004), Banff, Alberta, Canada, July 4-8, 2004, volume 69.

[Ho and Ermon, 2016] Ho, J. and Ermon, S. (2016).

Generative adversarial imitation learning.

In Advances in Neural Information Processing Systems 29 (NeurIPS'16), pages 4565–4573.

Bibliography (cont.)

[McAllester and Ortiz, 2003] McAllester, D. A. and Ortiz, L. E. (2003).

Concentration inequalities for the missing mass and for histogram rule error.

J. Mach. Learn. Res., 4:895–911.

[Pomerleau, 1991] Pomerleau, D. (1991).

Efficient training of artificial neural networks for autonomous navigation.

Neural Computation, 3(1):88–97.

[Ross et al., 2011] Ross, S., Gordon, G. J., and Bagnell, D. (2011).

A reduction of imitation learning and structured prediction to no-regret online learning.

In Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS'11), pages 627–635.

Bibliography (cont.)

[van Hasselt et al., 2016] van Hasselt, H., Guez, A., and Silver, D. (2016).

Deep reinforcement learning with double q-learning.

In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA, pages 2094–2100. AAAI Press.

Thank you!

Feel free to contact me for more discussions!

xut@lamda.nju.edu.cn