The Fundamental Limits of Imitation Learning

Tian Xu

xut@lamda.nju.edu.cn

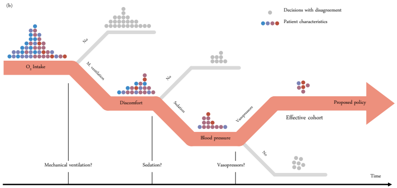
Nanjing University

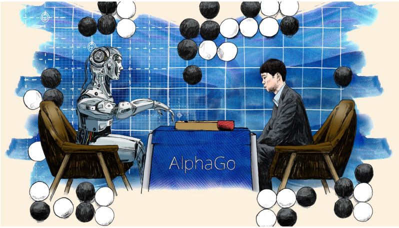
Mainly based on:
Toward the Fundamental Limits of Imitation Learning.
Provably Breaking the Quadratic Error Compounding Barrier in Imitation Learning, Optimally

June 25, 2021

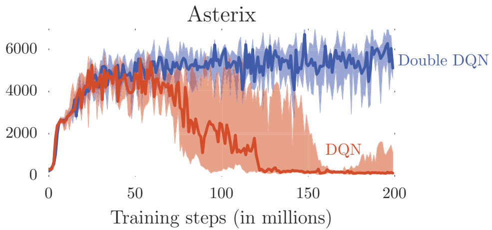
Outline

Reinforcement Learning (RL)





RL Challenges

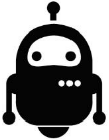


Double DQN requires million samples to solve Atari games [van Hasselt et al., 2016].

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 than RL.

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 finite 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.
- 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

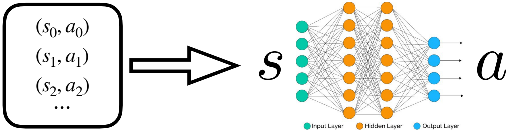
- In IL, our objective is to minimize the policy value gap:

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

- There are mainly two classes of methods: behavioral cloning (BC) [Pomerleau, 1991] and adversarial imitation learning (AIL) [Abbeel and Ng, 2004, Ho and Ermon, 2016].
  - BC: mimics expert actions with supervised learning.
  - AIL: firstly infers the reward function, then learns 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.

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

where $f_{\pi^*}^{t}(s)=\Pr_{\pi^*}(s_t=s)$.

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

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

where $f_{D}^{t}(s)=\frac{\sum_{i=1}^{m}\mathbb{I}(s_t^i=s)}{m}$.

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 $\hat{\pi} \in \Pi_{\text{mimic}}(D)$,

- *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.

Faster convergence of BC

- Connect policy value gap with the population risk [Ross et al., 2011]: $J(\pi^*) - J(\hat{\pi}) \leq H^2 \mathcal{L}_{\text{pop}}(\hat{\pi}, \pi^*)$.
- Upper bound the population risk with the missing mass: for each $\hat{\pi} \in \Pi_{\text{mimic}}(D)$,

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

$$= \frac{1}{H} \sum_{t=1}^{H} \underbrace{\sum_{s \in \mathcal{S}} f^t_{\pi^*}(s) \mathbb{I}(s_t \notin S_t(D))}_{\text{missing mass}}.$$

- For step $t \in [H]$, we consider the term $\sum_{s \in \mathcal{S}} f^t_{\pi^*}(s) \mathbb{I}(s_t \notin S_t(D))$, where $S_t(D) = \{(s^i_t, a^i_t)\}_{i=1}^m$ are i.i.d. drawn from $f^t_{\pi^*} \times \pi^*_t$.

Missing Mass

Definition 1 (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 <u>uncovered</u> in X^m.

Missing Mass

- Faster diminish rate of the expected missing mass:

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

- Faster concentration of missing mass [McAllester and Ortiz, 2003]: for any $\delta \in (0, \frac{1}{10}]$, w.p. $\ge 1 - \delta$,

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

- Faster diminish rate of policy value gap: $J\left(\pi^*\right) - J(\widehat{\pi}) \gtrsim \widetilde{\mathcal{O}}\left(\frac{H^2|\mathcal{S}|}{m}\right)$.

Outline

Compounding Error

- The planning horizon dependency of BC is $\mathcal{O}\left(H^2\right)$, causing a large policy value loss on long-horizon tasks.
- Under the non-interaction and active setting, the lower bound for any IL algorithms is of order $\Omega\left(H^2\right)$, implying that BC is already minimax optimal.
- Can we break this barrier if more environment information (i.e., the transition function) is provided to the learner?

MIMIC-MD

- Consider that the expert dataset D is equally divided into two parts $D = D_1 \cup D_2$.
- Recall the definition of $\Pi_{\text{mimic}}(D_1)$:

$$\Pi_{\text{mimic}}(D_1) \triangleq \left\{\pi \in \Pi : \forall t \in [H], s \in \mathcal{S}_t(D_1), \pi_t(\cdot \mid s) = \delta_{\pi_t^*(s)}\right\},$$

- Namely, $\Pi_{\text{mimic}}(D_1)$ is the set of BC policies on D_1.

MIMIC-MD

- Fixing $(s, a, t) \in \mathcal{S} \times \mathcal{A} \times [H]$, consider the set of trajectories $\mathcal{T}_t^{D_1}(s, a)$, each of which visits (s, a) at time t and at some time $\tau \leq t$ visits a state unvisited at time τ in D_1.
- Formally, $\mathcal{T}_t^{D_1}(s, a) \triangleq \left\{ \{(s_{t'}, a_{t'})\}_{t'=1}^{H} \mid s_t = s, a_t = a, \exists \tau \leq t : s_\tau \notin \mathcal{S}_\tau \left(D_1\right) \right\}$.
- Intuitively, $\mathcal{T}_t^{D_1}(s, a)$ is a set of trajectories that are not completely consistent with some trajectory in D_1.

MIMIC-MD

- The objective of MIMIC-MD:

$$\mathop{\arg\min}_{\pi \in \Pi_{\text{mimic}}(D_1)} \sum_{t=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \left| \Pr_{\pi}\left[\mathcal{T}_t^{D_1}(s,a)\right] - \frac{\sum_{\mathrm{tr} \in D_2} \mathbb{I}\left(\mathrm{tr} \in \mathcal{T}_t^{D_1}(s,a)\right)}{|D_2|} \right|$$

- Given D_1, $\frac{\sum_{\mathrm{tr} \in D_2} \mathbb{I}\left(\mathrm{tr} \in \mathcal{T}_t^{D_1}(s,a)\right)}{|D_2|}$ is an estimation of $\Pr_{\pi^*}\left[\mathcal{T}_t^{D_1}(s,a)\right]$ from the other half dataset D_2.
- For $\pi \in \Pi_{\text{mimic}}(D_1)$, π exactly takes the expert action on states covered in D_1.
- For a trajectory tr that is completely consistent with some trajectory in D_1 and $\pi \in \Pi_{\text{mimic}}(D_1)$, $\Pr_{\pi^*}(\mathrm{tr}) = \Pr_{\pi}(\mathrm{tr})$.
- This optimization problem cannot be exactly solved in polynomial time.

MIMIC-MD

Theorem 2

Consider $\widehat{\pi}$ is the solution of the above optimization problem, we have

$$J\left(\pi^{*}\right)-\mathbb{E}[J(\widehat{\pi}(D, P, \rho))] \lesssim \min \left\{H, \frac{|\mathcal{S}| H^{3/2}}{m}\right\}$$

- MIMIC-MD enjoys a horizon dependency of $\mathcal{O}\left(H^{3/2}\right)$, which is an improvement over the quadratic dependency of BC.
- MIMIC-MD keeps the faster rate of $\mathcal{O}\left(\frac{1}{m}\right)$ as in BC.

Analysis

Lemma 3

Fixing the expert dataset $D = D_1 \cup D_2$, *for any policy* $\widehat{\pi} \in \Pi_{mimic}(D_1)$, *we have*

$$J(\pi^*) - J(\widehat{\pi}) \leq \sum_{t=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \left| \Pr_{\widehat{\pi}} \left[\mathcal{T}_t^{D_1}(s,a) \right] - \Pr_{\pi^*} \left[\mathcal{T}_t^{D_1}(s,a) \right] \right|$$

- Since $\widehat{\pi}$ exactly takes the expert action on states covered in D_1, value loss only occurs on trajectories belong to $\mathcal{T}_t^{D_1}(s,a)$, a set of trajectories that are not completely agree with some trajectory in D_1.

Proof

- Given D_1, for $t \in [H]$, define $\mathcal{E}_{D_1}^{\leq t} = \{\exists \tau < t : s_\tau \notin \mathcal{S}_\tau(D_1)\}$ as the event that the policy under consideration visits some state at time $\tau < t$ uncovered in D_1.
- $J(\pi^*) - J(\widehat{\pi}(D)) =$
$\sum_{t=1}^{H} \mathbb{E}_{\pi^*}\left[\left(\mathbb{I}\left(\left(\mathcal{E}_{D_1}^{\leq t}\right)^c\right) + \mathbb{I}\left(\mathcal{E}_{D_1}^{\leq t}\right)\right)\mathbf{r}_t(s_t, a_t)\right] - \mathbb{E}_{\widehat{\pi}}\left[\left(\mathbb{I}\left(\left(\mathcal{E}_{D_1}^{\leq t}\right)^c\right) + \mathbb{I}\left(\mathcal{E}_{D_1}^{\leq t}\right)\right)\mathbf{r}_t(s_t, a_t)\right]$
- As $\widehat{\pi} \in \Pi_{\text{mimic}}(D_1)$, $\sum_{t=1}^{H} \mathbb{E}_{\pi^*}\left[\mathbb{I}\left(\left(\mathcal{E}_{D_1}^{\leq t}\right)^c\right) \mathrm{r}_t(s_t, a_t)\right] = \sum_{t=1}^{H} \mathbb{E}_{\widehat{\pi}}\left[\mathbb{I}\left(\left(\mathcal{E}_{D_1}^{\leq t}\right)^c\right) \mathrm{r}_t(s_t, a_t)\right]$.
- $J(\pi^*) - J(\widehat{\pi}) =$
$\sum_{t=1}^{H} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mathbf{r}_t(s,a)\left(\mathrm{Pr}_{\pi^*}\left[\mathcal{E}_{D_1}^{\leq t}, s_t = s, a_t = a\right] - \mathrm{Pr}_{\widehat{\pi}}\left[\mathcal{E}_{D_1}^{\leq t}, s_t = s, a_t = a\right]\right)$

$$\leq \sum_{t=1}^{H} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left|\mathrm{Pr}_{\pi^*}\left[\mathcal{E}_{D_1}^{\leq t}, s_t = s, a_t = a\right] - \mathrm{Pr}_{\widehat{\pi}}\left[\mathcal{E}_{D_1}^{\leq t}, s_t = s, a_t = a\right]\right|$$

$$= \sum_{t=1}^{H} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left|\mathrm{Pr}_{\pi^*}\left[\mathcal{T}_t^{D_1}(s,a)\right] - \mathrm{Pr}_{\widehat{\pi}}\left[\mathcal{T}_t^{D_1}(s,a)\right]\right|$$

Analysis

Lemma 4

$$\sum_{t=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \mathbb{E} \left[\left| \mathrm{Pr}_{\pi^*} \left[\mathcal{T}_t^{D_1}(s,a) \right] - \frac{\sum_{\mathrm{tr} \in D_2} \mathbb{I}\left(\mathrm{tr} \in \mathcal{T}_t^{D_1}(s,a) \right)}{|D_2|} \right| \right] \leq \frac{8}{3} \frac{|\mathcal{S}| H^{\frac{3}{2}}}{N}$$

Proof

$$\sum_{t=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \mathbb{E}\left[\left| \operatorname{Pr}_{\pi^*}\left[\mathcal{T}_t^{D_1}(s,a)\right] - \frac{\sum_{\mathrm{tr} \in D_2} \mathbb{I}\left(\mathrm{tr} \in \mathcal{T}_t^{D_1}(s,a)\right)}{|D_2|} \right| \right]$$

$$\overset{(1)}{\leq} \sum_{t=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \left(\mathbb{E}\left[\left(\operatorname{Pr}_{\pi^*}\left[\mathcal{T}_t^{D_1}(s,a)\right] - \frac{\sum_{\mathrm{tr} \in D_2} \mathbb{I}\left(\mathrm{tr} \in \mathcal{T}_t^{D_1}(s,a)\right)}{|D_2|} \right)^2 \right] \right)^{1/2}$$

$$\leq \sum_{t=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \left(\frac{1}{|D_2|} \operatorname{Var}\left[\mathbb{I}\left(\mathrm{tr}_1 \in \mathcal{T}_t^{D_1}(s,a)\right]\right)^{1/2} \right)$$

$$\overset{(2)}{\leq} \sum_{t=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \left(\frac{1}{|D_2|} \operatorname{Pr}_{\pi^*}\left[\mathcal{T}_t^{D_1}(s,a)\right] \right)^{1/2}$$

Inequality (1) follows the Jensen Inequality, Inequality (2) follows that $\operatorname{Var}[X] = p(1-p) \leq p$ for a Bernoulli random variable X.

Proof

$$
\begin{aligned}
\sum_{t=1}^{H} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mathbb{E}\left[\left(\frac{1}{|D_2|}\Pr_{\pi^*}\left[\mathcal{T}_t^{D_1}(s,a)\right]\right)^{1/2}\right] &\leq \sum_{t=1}^{H} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left(\frac{1}{|D_2|}\mathbb{E}\left[\Pr_{\pi^*}\left[\mathcal{T}_t^{D_1}(s,a)\right]\right]\right)^{1/2} \\
&\leq \sum_{t=1}^{H} \left(\frac{|\mathcal{S}|}{|D_2|}\right)^{1/2} \left(\sum_{s\in\mathcal{S}, a=\pi_t^*(s)} \mathbb{E}\left[\Pr_{\pi^*}\left[\mathcal{T}_t^{D_1}(s,a)\right]\right]\right)^{1/2} \\
&\leq \sum_{t=1}^{H} \left(\frac{|\mathcal{S}|}{|D_2|}\right)^{1/2} \left(\mathbb{E}\left[\Pr_{\pi^*}\left[\varepsilon_{D_1}^{\leq t}\right]\right]\right)^{1/2}
\end{aligned}
$$

- $\Pr_{\pi^*}\left[\varepsilon_{D_1}^{\leq t}\right]$ is the probability that π^* visits a state at some time $\tau \leq t$ uncovered in D_1. This term is closely related to the missing mass.

Proof

- Connect $\Pr_{\pi^*}\left[\mathcal{E}_{D_1}^{\leq t}\right]$ with missing mass.

$$\Pr_{\pi^*}\left[\mathcal{E}_{D_1}^{\leq t}\right] = \Pr_{\pi^*}\left[\exists \tau \leq t : s_\tau \notin \mathcal{S}_\tau\left(D_1\right)\right] = \sum_{\tau=1}^{t} \Pr_{\pi^*}\left[\forall \tau' < \tau, s_{\tau'} \in \mathcal{S}_{\tau'}\left(D_1\right), s_\tau \notin \mathcal{S}_\tau\left(D_1\right)\right]$$

$$\leq \sum_{\tau=1}^{t} \Pr_{\pi^*}\left[s_\tau \notin \mathcal{S}_\tau\left(D_1\right)\right] = \sum_{\tau=1}^{t} \sum_{s \in \mathcal{S}} \Pr_{\pi^*}\left[s_\tau = s\right] \mathbb{I}\left(s \notin \mathcal{S}_\tau\left(D_1\right)\right)$$

$$\leq \sum_{\tau=1}^{H} \underbrace{\sum_{s \in \mathcal{S}} \Pr_{\pi^*}\left[s_\tau = s\right] \mathbb{I}\left(s \notin \mathcal{S}_\tau\left(D_1\right)\right)}_{\text{missing mass at time } \tau}$$

- We have shown that $\mathbb{E}\left[\sum_{s \in \mathcal{S}} \Pr_{\pi^*}\left[s_\tau = s\right] \mathbb{I}\left(s \notin \mathcal{S}_\tau\left(D_1\right)\right)\right] \leq \frac{4|S|}{9|D_1|}$.
- $J(\pi^*) - \mathbb{E}\left[J(\widehat{\pi})\right] \leq \sum_{t=1}^{H} \left(\frac{|\mathcal{S}|}{|D_2|}\right)^{1/2} \left(\mathbb{E}\left[\Pr_{\pi^*}\left[\varepsilon_{D_1}^{\leq t}\right]\right]\right)^{1/2} \leq \frac{4}{3} \frac{|\mathcal{S}| H^{3/2}}{m} \lesssim \frac{|\mathcal{S}| H^{3/2}}{m}$.

Outline

Lower Bound

Theorem 5

Suppose $H \geq 2$ and $N \geq 7$. There exists a three-state MDP $\mathcal{M}$ and an expert policy π^* such that, for every learner $\widehat{\pi}$,

$$\Pr\left(J(\pi^*) - J(\widehat{\pi}) \gtrsim \frac{H^{3/2}}{m} \right) \geq c',$$

for some constants $c, c' > 0$. The probability is taken over the randomness of the expert dataset D.

- The lower bound of $\Omega\left(\frac{H^{3/2}}{m}\right)$ implies that MIMIC-MD is minimax optimal when the transition function is known.

Analysis

Lemma 6

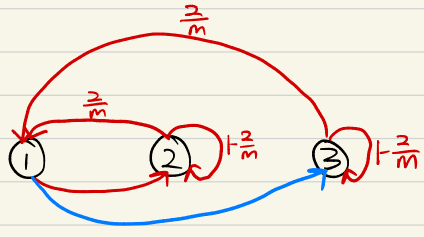
Suppose there exist a three-state MDP $\mathcal{M}$ and expert policy π^* such that for every learner $\widehat{\pi}$, $\Pr\left(\left|J_{\mathcal{M}}(\pi^*) - J_{\mathcal{M}}(\widehat{\pi})\right| \gtrsim \frac{H^{3/2}}{m}\right) \geq c'$ for some constant $0 < c' \leq 1$. Then there exist a three-state MDP $\mathcal{M}$ and expert policy π^* such that for every learner $\widehat{\pi}$, $\Pr\left(J_{\mathcal{M}}(\pi^*) - J_{\mathcal{M}}(\widehat{\pi}) \gtrsim \frac{H^{3/2}}{m}\right) \geq \frac{c'}{2}$.

- Given expert policy π^*, the learner cannot distinguish between $\mathcal{M} = (\rho, P, r)$ and $\mathcal{M}' = (\rho, P, 1 - r)$ from expert dataset only with state-action pairs.
- For an arbitrary policy π, we have that $J_{\mathcal{M}}(\pi) + J_{\mathcal{M}'}(\pi) = H$. Therefore, the learner needs to upper bound the two-sided error.
- This assumption on the problem class seems strange since the expert policy π^* cannot perform good on both $\mathcal{M}$ and $\mathcal{M}'$.

Analysis

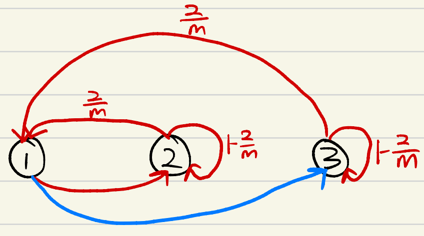
- We aim to prove that there exist $\mathcal{M}$ and π^*, for every learner $\widehat{\pi}$, $\Pr\left(\left| J_{\mathcal{M}}(\pi^*) - J_{\mathcal{M}}(\widehat{\pi}) \right| \gtrsim \frac{H^{3/2}}{m} \right) \geq c'$ for some constant $0 < c' \leq 1$.
- We consider the following three-state MDP.

Three-state MDP



- There are three states $\mathcal{S} = \{1, 2, 3\}$ and two actions $\mathcal{A} = \{R, B\}$.
- On state 1, if the agent takes action R, it deterministically goes to state 2. Otherwise, it deterministically goes to state 3.
- On states 2 and 3, no matter which action is taken, the agent goes to state 1 with a probability of $\frac{2}{m}$ and stays absorbing with a probability of $1 - \frac{2}{m}$.

Three-state MDP



- The reward equals 1 on state 2 and 0 on the other state-action pairs.
- Only actions on state 1 are meaningful and the optimal policy is $\pi_t^*(\cdot|1) = (\pi_t^*(R|1), \pi_t^*(B|1)) = (1, 0)$ for $t \in [H]$.

Analysis

- Given the three-state MDP $\mathcal{M}$, there exists π^*, for every learner $\widehat{\pi}$, $\Pr\left(|J_{\mathcal{M}}(\pi^*) - J_{\mathcal{M}}(\widehat{\pi})| \gtrsim \frac{H^{3/2}}{m}\right) \geq c'$.
- It suffices to find a prior distribution $\mathcal{D}$ over π^* such that $\mathbb{E}_{\pi^* \sim \mathcal{D}}\left[\Pr\left(|J_{\mathcal{M}}(\pi^*) - J_{\mathcal{M}}(\widehat{\pi})| \gtrsim \frac{H^{3/2}}{m}\right)\right] \geq c'$.
- For $t \in [H]$, $\pi_t^* (r \mid 1) \sim \mathrm{Unif}(\{0, 1\})$.

Analysis

Lemma 7

$$\mathbb{E}_{\pi^* \sim \mathcal{D}}\left[\Pr_D\left(|J_{\mathcal{M}}(\pi^*) - J_{\mathcal{M}}(\widehat{\pi})| \lesssim \frac{H^{3/2}}{m}\right)\right] \leq \frac{1}{2} + \mathbb{E}_D\left[\Pr_{\pi_1^*, \pi_2^*}\left(|J_{\mathcal{M}}(\pi_1^*) - J_{\mathcal{M}}(\pi_2^*)| \lesssim \frac{H^{3/2}}{m} \middle| D\right)\right],$$

where π_1^* and π_2^* are two independent samples drawn from the posterior distribution conditioned on the expert dataset D.

Proof

$$2\mathbb{E}_{\pi^* \sim \mathcal{D}}\left[\mathbb{E}_D\left[\mathbb{I}\left(|J_{\mathcal{M}}(\pi^*) - J_{\mathcal{M}}(\widehat{\pi})| \lesssim \frac{H^{3/2}}{m}\right)\right]\right] = 2\mathbb{E}_D\left[\mathbb{E}_{\pi^*}\left[\mathbb{I}\left(|J_{\mathcal{M}}(\pi^*) - J_{\mathcal{M}}(\widehat{\pi})| \lesssim \frac{H^{3/2}}{m}\right)\middle| D\right]\right]$$

$$= \mathbb{E}_D\left[\mathbb{E}_{\pi_1^*}\left[\mathbb{I}\left(|J_{\mathcal{M}}(\pi_1^*) - J_{\mathcal{M}}(\widehat{\pi})| \lesssim \frac{H^{3/2}}{m}\right)\middle| D\right]\right] + \mathbb{E}_D\left[\mathbb{E}_{\pi_2^*}\left[\mathbb{I}\left(|J_{\mathcal{M}}(\pi_2^*) - J_{\mathcal{M}}(\widehat{\pi})| \lesssim \frac{H^{3/2}}{m}\right)\middle| D\right]\right]$$

$$\overset{(1)}{\leq} 1 + \mathbb{E}_D\left[\mathbb{E}_{\pi_1^*, \pi_2^*}\left[\mathbb{I}\left(|J_{\mathcal{M}}(\pi_1^*) - J_{\mathcal{M}}(\widehat{\pi})| + |J_{\mathcal{M}}(\pi_2^*) - J_{\mathcal{M}}(\widehat{\pi})| \lesssim \frac{H^{3/2}}{m}\right)\middle| D\right]\right]$$

$$\overset{(2)}{\leq} 1 + \mathbb{E}_D\left[\mathbb{E}_{\pi_1^*, \pi_2^*}\left[\mathbb{I}\left(|J_{\mathcal{M}}(\pi_1^*) - J_{\mathcal{M}}(\pi_2^*)| \lesssim \frac{H^{3/2}}{m}\right)\middle| D\right]\right]$$

- Inequality (1) follows that $\mathbb{I}(x \leq a) + \mathbb{I}(y \leq b) \leq 1 + \mathbb{I}(x + y \leq a + b)$.
- Inequality (2) follows that $|J_{\mathcal{M}}(\pi_1^*) - J_{\mathcal{M}}(\widehat{\pi})| + |J_{\mathcal{M}}(\pi_2^*) - J_{\mathcal{M}}(\widehat{\pi})| \lesssim \frac{H^{3/2}}{m} \rightarrow |J_{\mathcal{M}}(\pi_1^*) - J_{\mathcal{M}}(\pi_2^*)| \lesssim \frac{H^{3/2}}{m}$.

Analysis

Lemma 8

Conditioned on the expert dataset D, the expert policy $\pi^* \sim \mathrm{Unif}\left(\Pi_{\mathrm{mimic}}(D)\right)$. In other words, at time $t \in [H]$ such that state 1 is unvisited in any trajectory in the expert dataset, $\pi^*_t(R \mid 1) \sim \mathrm{Unif}(\{0, 1\})$.

Proof

- Note that for $t \in [H]$, $\Pr_{\pi}(s_t = 1)$ is the same for all policies and we denote it as $\Pr(s_t = 1)$.
- For a fixed time $t \in [H]$, we consider the random variables $\pi_t^*(R|1)$ and $D_t = \{(s_t^i, a_t^i)\}_{i=1}^m$.
- We list the joint probabilities as follows. WR means that D_t contains state 1 and the corresponding action is R and WO means that D_t does not cover state 1.

	WR	WB	WO
1	$\frac{1}{2}(1-(1-\Pr(s_t=1))^m)$	0	$\frac{1}{2}(1-\Pr(s_t=1))^m$
0	0	$\frac{1}{2}(1-(1-\Pr(s_t=1))^m)$	$\frac{1}{2}(1-\Pr(s_t=1))^m$

- $\Pr\left(\pi_t^*(R|1) = 1 \middle| D_t = WR\right) = 1, \Pr\left(\pi_t^*(R|1) = 0 \middle| D_t = WB\right) = 1,$
  $\Pr\left(\pi_t^*(R|1) = 0 \middle| D_t = WO\right) = \Pr\left(\pi_t^*(R|1) = 0 \middle| D_t = WO\right) = \frac{1}{2}.$

Analysis

- We want to prove that $\mathbb{E}_D\left[\mathrm{Pr}_{\pi_1^*,\pi_2^*}\left(|J_{\mathcal{M}}(\pi_1^*) - J_{\mathcal{M}}(\pi_2^*)| \gtrsim \frac{H^{3/2}}{m}\middle| D\right)\right] \geq c$ for some constant $0 < c \leq 1$.
- It is easy to calculate that $J_{\mathcal{M}}\left(\pi^*\right) = \sum_{t=1}^{H-1}\left(\sum_{t'=t+1}^{H}\left(1-\frac{2}{m}\right)^{H-t'}\right)\mathrm{Pr}\left(s_t = 1\right)\pi_t^*\left(R \mid 1\right) + \sum_{t=1}^{H}\left(1-\frac{2}{m}\right)^{t-1}$.

Analysis

- Conditioned on the expert dataset D,

$$J_{\mathcal{M}}(\pi_1^*) - J_{\mathcal{M}}(\pi_2^*) = \sum_{t=1}^{H-1} \left(\sum_{t'=t+1}^{H} \left(1 - \frac{2}{m}\right)^{H-t'} \right) \Pr\left(s_t = 1\right) X_t \mathbb{I}\left(1 \notin \mathcal{S}_t(D)\right),$$

  where X_t are i.i.d. random variables distributed as

$$X_t = \begin{cases} -1, & \text{w.p. } \frac{1}{4} \\ 0, & \text{w.p. } \frac{1}{2} \\ +1, & \text{w.p. } \frac{1}{4} \end{cases}$$

  .

- Let $Z_D = J_{\mathcal{M}}(\pi_1^*) - J_{\mathcal{M}}(\pi_2^*) = \sum_{t=1}^{H-1} \kappa_t X_t$, $\mathbb{E}\left[Z_D | D\right] = 0$ and $\mathrm{Var}\left[Z_D | D\right] = \mathbb{E}\left[Z_D^2 | D\right]$.

Key Inequality

Lemma 9 (Paley-Zygmund Argument)

For a random variable X, we have that

$$\Pr\left(X \geq \theta \mathbb{E}\left[X\right]\right) \geq (1-\theta)^2 \frac{\left(\mathbb{E}\left[X\right]\right)^2}{\mathbb{E}\left[X^2\right]}.$$

- A common strategy is to prove a lower bound of $\mathbb{E}[X]$. Set θ as a constant and lower bound $\frac{(\mathbb{E}[X])^2}{\mathbb{E}[X^2]}$ by a constant.

Analysis

- Applying Paley-Zygmund Argument on random variable Z_D^2 yields

$$\Pr\left(Z_D^2 \geq \theta \mathbb{E}\left[Z_D^2 | D\right] | D\right) \geq (1-\theta)^2 \frac{\left(\mathbb{E}\left[Z_D^2 | D\right]\right)^2}{\mathbb{E}\left[Z_D^4 | D\right]}.$$

- It is easy to derive that $\frac{\left(\mathbb{E}[Z_D^2|D]\right)^2}{\mathbb{E}[Z_D^4|D]} \geq \frac{1}{3}$. Choosing $\theta = \frac{1}{10}$ yields

$$\Pr\left(Z_D^2 \geq \frac{1}{10}\mathbb{E}\left[Z_D^2 | D\right] | D\right) \geq \frac{27}{100}.$$

- It suffices to prove that $\Pr_D\left(\mathbb{E}\left[Z_D^2 | D\right] \gtrsim \frac{H^3}{m^2}\right) \geq c$ for $c > 0$.

Analysis

- We first lower bound the prior variance: $\mathbb{E}\left[Z_D^2\right] = \mathbb{E}\left[\mathbb{E}\left[Z_D^2|D\right]\right] \gtrsim \frac{H^3}{m^2}$.

$$\mathbb{E}\left[Z_D^2 \mid D\right] = \frac{1}{2}\sum_{t=1}^{H}\kappa_t^2 = \frac{1}{2}\sum_{t=1}^{H}\left(\sum_{t'=t+1}^{H}\left(1-\frac{2}{N}\right)^{H-t'}\right)^2 \left(\Pr\left(s_t=1\right)\right)^2 \mathbb{I}(1 \in \mathcal{S}_t(D))$$

$$\mathbb{E}\left[Z_D^2\right] = \frac{1}{2}\sum_{t=1}^{H}\underbrace{\left(\sum_{t'=t+1}^{H}\left(1-\frac{2}{N}\right)^{H-t'}\right)^2}_{\Omega(H^2)} \underbrace{\left(\Pr\left(s_t=1\right)\right)^2}_{\Omega(1/m^2)} \Pr\left(1 \in \mathcal{S}_t(D)\right) \gtrsim \frac{H^3}{m^2}$$

- We again utilize the Paley-Zygmund Argument on random variable $\mathbb{E}\left[Z_D^2|D\right]$:

$$\Pr_D\left(\mathbb{E}\left[Z_D^2|D\right] \geq \frac{1}{10}\mathbb{E}\left[Z_D^2\right]\right) \geq \frac{81}{100}\frac{\left(\mathbb{E}\left[Z_D^2\right]\right)^2}{\mathbb{E}\left[\mathbb{E}\left[Z_D^2|D\right]^2\right]} \geq \frac{9}{25}.$$

Analysis

- Now we have (i) $\Pr\left(Z_D^2 \geq \frac{1}{10}\mathbb{E}\left[Z_D^2|D\right]|D\right) \geq \frac{27}{100}$, (ii) $\Pr_D\left(\mathbb{E}\left[Z_D^2|D\right] \gtrsim \frac{H^3}{m^2}\right) \geq \frac{9}{25}$.
- We want to prove $\mathbb{E}_D\left[\Pr\left(Z_D^2 \gtrsim \frac{H^3}{m^2}\middle| D\right)\right] \geq c$ for $c > 0$. Let $\mathcal{E}$ be the event that $\mathbb{E}\left[Z_D^2|D\right] \gtrsim \frac{H^3}{m^2}$.

$$
\begin{aligned}
\mathbb{E}_D\left[\Pr\left(Z_D^2 \gtrsim \frac{H^3}{m^2}\middle| D\right)\right] &\geq \Pr(\mathcal{E})\mathbb{E}_D\left[\Pr\left(Z_D^2 \gtrsim \frac{H^3}{m^2}\middle| D\right)\middle|\mathcal{E}\right] \\
&\geq \Pr(\mathcal{E})\mathbb{E}_D\left[\Pr\left(Z_D^2 \geq \frac{1}{10}\mathbb{E}\left[Z_D^2|D\right]\middle| D\right)\middle|\mathcal{E}\right] \\
&\geq \frac{9}{25}\frac{27}{100}.
\end{aligned}
$$

Outline

Summary

Setting		Value gap
No-interaction / Active	BC	$\widetilde{\mathcal{O}}\left(\frac{H^2 \lvert\mathcal{S}\rvert}{m}\right)$
	Lower bound	$\widetilde{\Omega}\left(\frac{H^2 \lvert\mathcal{S}\rvert}{m}\right)$
Known transition	MIMIC-MD	$\widetilde{\mathcal{O}}\left(\frac{H^{3/2} \lvert\mathcal{S}\rvert}{m}\right)$
	Lower bound	$\widetilde{\Omega}\left(\frac{H^{3/2} \lvert\mathcal{S}\rvert}{m}\right)$

Future Direction

- The known transition setting is not practical and a more common setting is that the agent does not know the exact transition function but can interact with the environment.
- The exploration issue in IL: how many environment interactions are required to achieve a desired policy value gap ?
  - Upper bound: BC does not need exploration but suffers from the compounding error issue. AIL optimizes policy in each iteration and requires exploration.
  - Lower bound: the characteristics of IL, the learner cannot observe true rewards but have access to expert demonstrations.

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.

[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