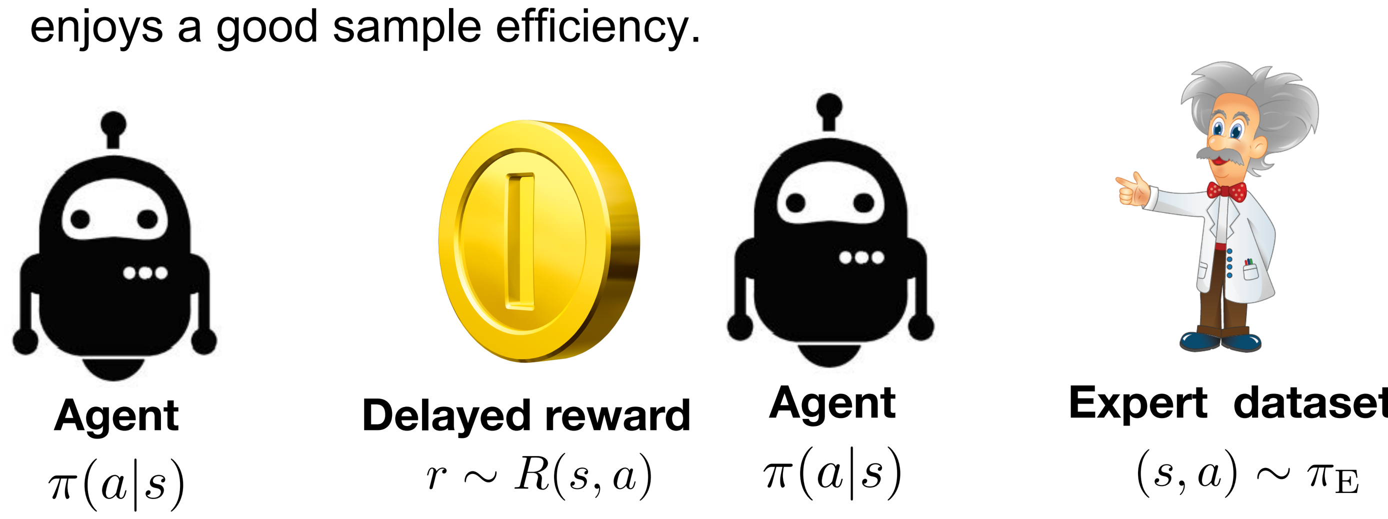
Error Bounds of Imitating Polices and Environments

Tian Xu, Nanjing University
Ziniu Li, The Chinese University of Hong Kong, Shenzhen & Polixir Technologies
Yang Yu, Nanjing University & Polixir Technologies

NEURAL INFORMATION PROCESSING SYSTEMS

Background

- Reinforcement learning (RL) learns from **delayed feedback** and may be not sample-efficient.
- Imitation learning (IL) learns from **expert demonstrations** and enjoys a good sample efficiency.



In IL, there are two famous methods: behaviorial cloning (BC) [1] and generative adversarial imitation learning (GAIL) [2].

- BC reduces IL to supervised learning and suffers from the **issue of compounding errors.**
- GAIL achieves better empirical performance than BC, but its theoretical understanding needs further studies.

Setup and IL algorithms:

- Infinite-horizon discounted MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, M^*, R, \gamma, d_0)$
  - S and A are finite state and finite action space
  - M^* is the transition function
  - R is the reward function bounded by R_{max}
  - γ is the discounted factor and d_0 is initial state distribution
- Policy $\pi: S \longrightarrow \Delta(A)$, policy value: $V_\pi = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | d_0, \pi, M^*]$
- **Effective planning horizon:** $\frac{1}{1-\gamma}$
- State distribution d_π and state-action distribution ρ_π
- The focus of IL: **policy value gap** $V_{\pi_E} - V_\pi$

BC: minimize the divergence between **policy distributions**

$$\min_{\pi \in \Pi} \mathbb{E}_{s \sim d_{\pi_E}} [D_{\mathrm{KL}}(\pi_E(\cdot|s), \pi(\cdot|s))]$$

GAIL: minimize the divergence between **state-action distributions**

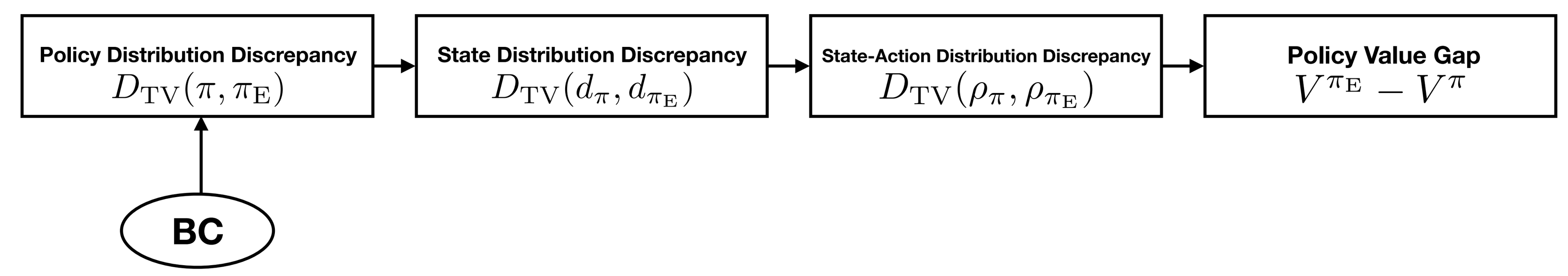
$$\min_{\pi \in \Pi} D_{\mathrm{JS}}(\rho_{\pi_E}, \rho_\pi)$$

Error Bounds of Imitating Polices

Behavioral Cloning:

Theorem 1: Given an expert policy π_E and an imitated policy π_{BC} with $\mathbb{E}_{s \sim d_{\pi_E}}[D_{\mathrm{KL}}(\pi_E(\cdot|s), \pi_{BC}(\cdot|s))] \leq \epsilon$ (which can be achieved BC), we have that $V_{\pi_E} - V_{\pi_{BC}} \leq \frac{2\sqrt{2}R_{\max}}{(1-\gamma)^2}\sqrt{\epsilon}$

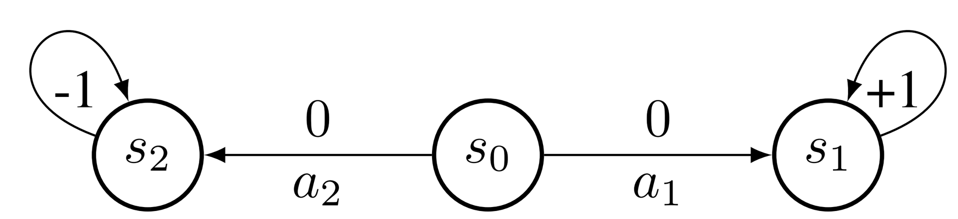
- The error bound of BC has a **quadratic** dependency on the effective horizon, verifying the issue of compounding errors from theoretical view.
- The proof is based on the following coherent error-propagation analysis:



Corollary 1: Suppose that π_E and π_{BC} are deterministic and the provided function class Π satisfies realizability. $\forall\, \delta \in (0,1)$, w.p. $\geq 1-\delta$, we have that

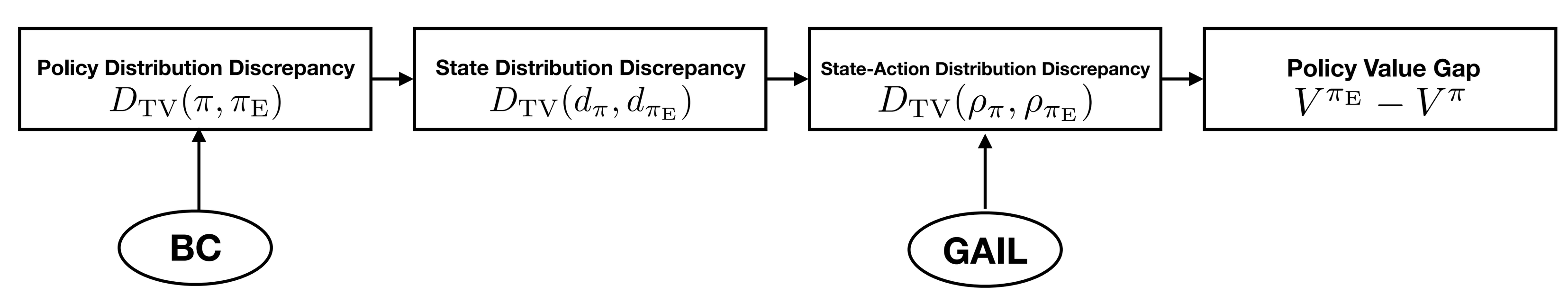
$$V_{\pi_E} - V_{\pi_{BC}} \leq \frac{2\sqrt{2}R_{\max}}{(1-\gamma)^2}\left(\frac{1}{m}\log(|\Pi|) + \frac{1}{m}\log(\frac{1}{\delta})\right)$$

The following example shows that the quadratic dependency of BC is unavoidable in the worst case.



A ``hard'' deterministic MDP for BC. Digits on arrows are corresponding rewards. Initial state is s_0 while s_1 and s_2 are two absorbing states.

Generative Adversarial Imitation Learning:

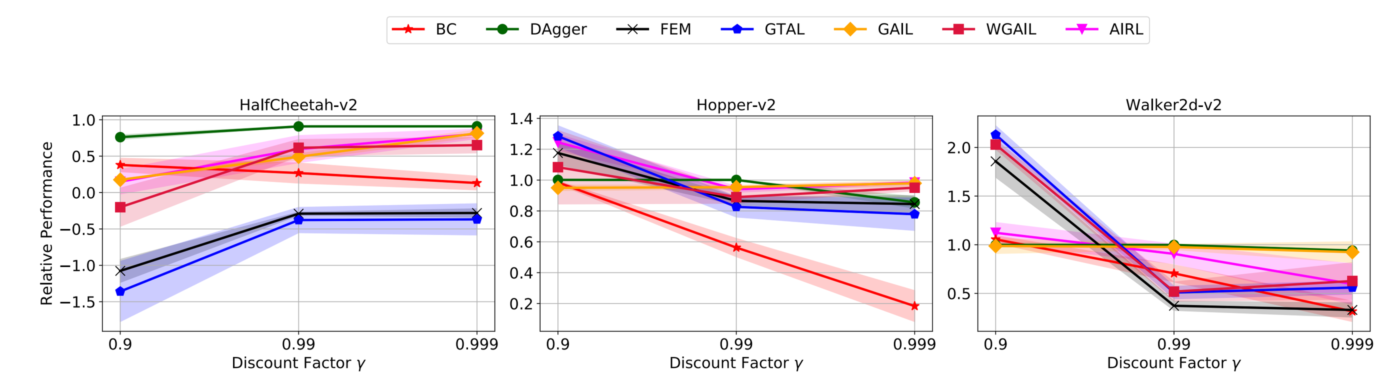


Theorem 2: Given an expert policy π_E and an imitated policy π_{GA} with $d_{\mathcal{D}}(\hat{\rho}_{\pi_E}, \hat{\rho}_{\pi_{GA}}) - \inf_{\pi \in \Pi} d_{\mathcal{D}}(\hat{\rho}_{\pi_E}, \hat{\rho}_\pi) \leq \hat{\epsilon}$ (which can be achieved GAIL), w.p. $\geq 1-\delta$, we have that

$$V_{\pi_{\mathrm{E}}} - V_{\pi_{\mathrm{GA}}} \leq \frac{\|r\|_{\mathcal{D}}}{1-\gamma}\Big(\underbrace{\inf_{\pi \in \Pi} d_{\mathcal{D}}(\hat{\rho}_{\pi_E}, \hat{\rho}_\pi)}_{\mathrm{Appr}(\Pi)} + \underbrace{2\hat{\mathcal{R}}^{(m)}_{\rho_{\pi_E}}(\mathcal{D}) + 2\hat{\mathcal{R}}^{(m)}_{\rho_{\pi_{GA}}}(\mathcal{D}) + 12\Delta\sqrt{\frac{\log(2/\delta)}{m}}}_{\mathrm{Estm}(\mathcal{D}, m, \delta)} + \hat{\epsilon}\Big),$$

- Compared to BC, GAIL enjoys a **linear** dependency on the effective horizon.
- Moreover, theorem 2 suggests seeking **a trade-off on the complexity of discriminator class $\mathcal{D}$**

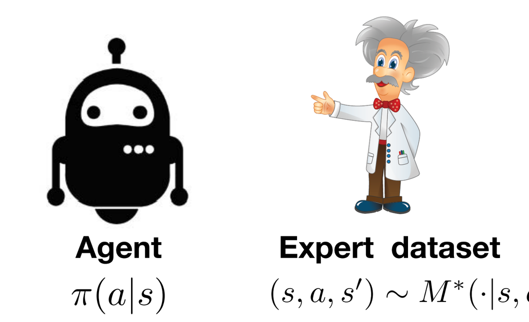
Experiments:



As $\gamma \to 1$, the effective planning horizon increases, BC is worse than GAIL, and other adversarial-based methods.

Error Bounds of Imitating Environments

By treating environment transition model as dual agent, learning the transition function can also be treated by imitation learning.



Imitate Environments via BC:

$$\min_\theta \mathbb{E}_{(s,a) \sim \rho^{M^*}_{\pi_D}}\left[D_{\mathrm{KL}}\big(M^*(\cdot|s,a), M_\theta(\cdot|s,a)\big)\right]$$

Lemma 3: Given a learned transition model M_θ by BC with $\mathbb{E}_{(s,a)\sim\rho^{M^*}_{\pi_D}}[D_{\mathrm{KL}}(M^*(\cdot|s,a), M_\theta(\cdot|s,a))] \leq \epsilon_m$, for an arbitrary bounded divergence policy π with $\max_s D_{\mathrm{KL}}\big(\pi(\cdot|s), \pi_D(\cdot|s)\big) \leq \epsilon_\pi$, we have

$$|V^{M^*}_\pi - V^{M_\theta}_\pi| \leq \frac{\sqrt{2}R_{\max}\gamma}{(1-\gamma)^2}\sqrt{\epsilon_m} + \frac{2\sqrt{2}R_{\max}}{(1-\gamma)^2}\sqrt{\epsilon_\pi}$$

Imitate Environments via GAIL:

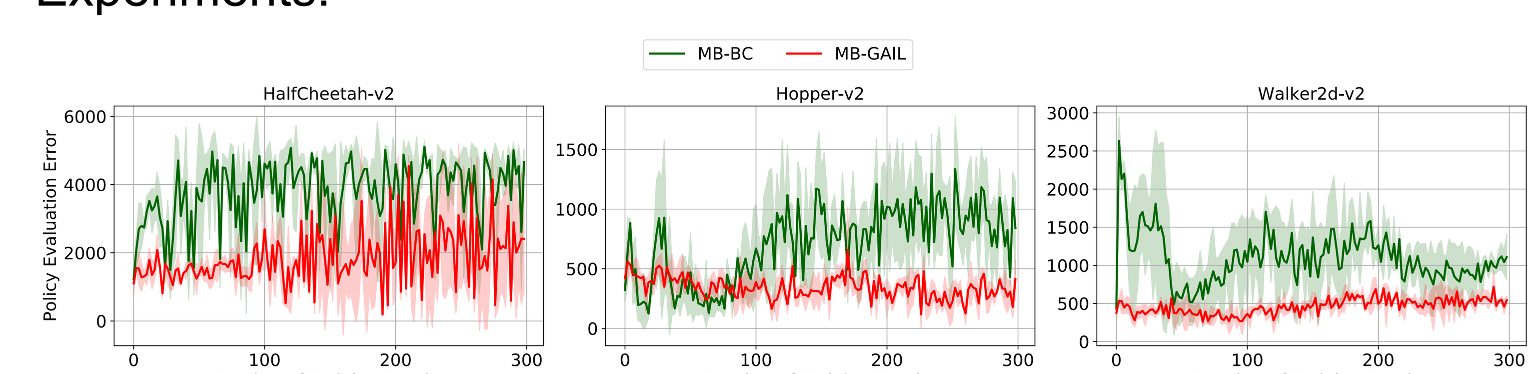
$$\min_\theta D_{\mathrm{JS}}(\mu^{M_\theta}, \mu^{M^*})$$

Lemma 4: Given a learned transition model M_θ by GAIL with $D_{\mathrm{JS}}(\mu^{M_\theta}, \mu^{M^*}) \leq \epsilon_m$, under the same assumption of lemma 3, we have

$$|V^{M_\theta}_\pi - V^{M^*}_\pi| \leq \frac{2\sqrt{2}R_{\max}}{1-\gamma}\sqrt{\epsilon_m} + \frac{2\sqrt{2}R_{\max}}{(1-\gamma)^2}\sqrt{\epsilon_\pi}$$

Learning the environment transition with GAIL-style learner can mitigate the model-bias when evaluating policies.

Experiments:



References

[1] Dean Pomerleau. Efficient training of artificial neural networks for autonomous navigation. Neural Computation, 1991.
[2] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In NeurIPS'16, 2016.