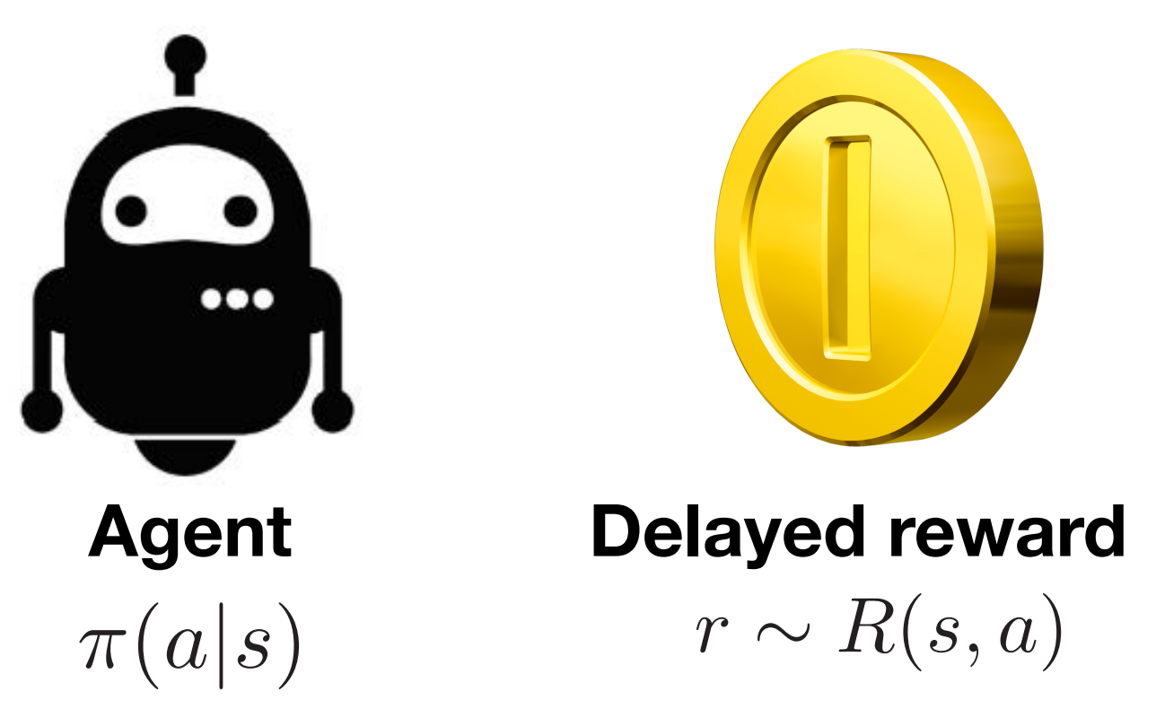
Download Paper

Error Bounds of Imitating Policies and Environments for Reinforcement Learning

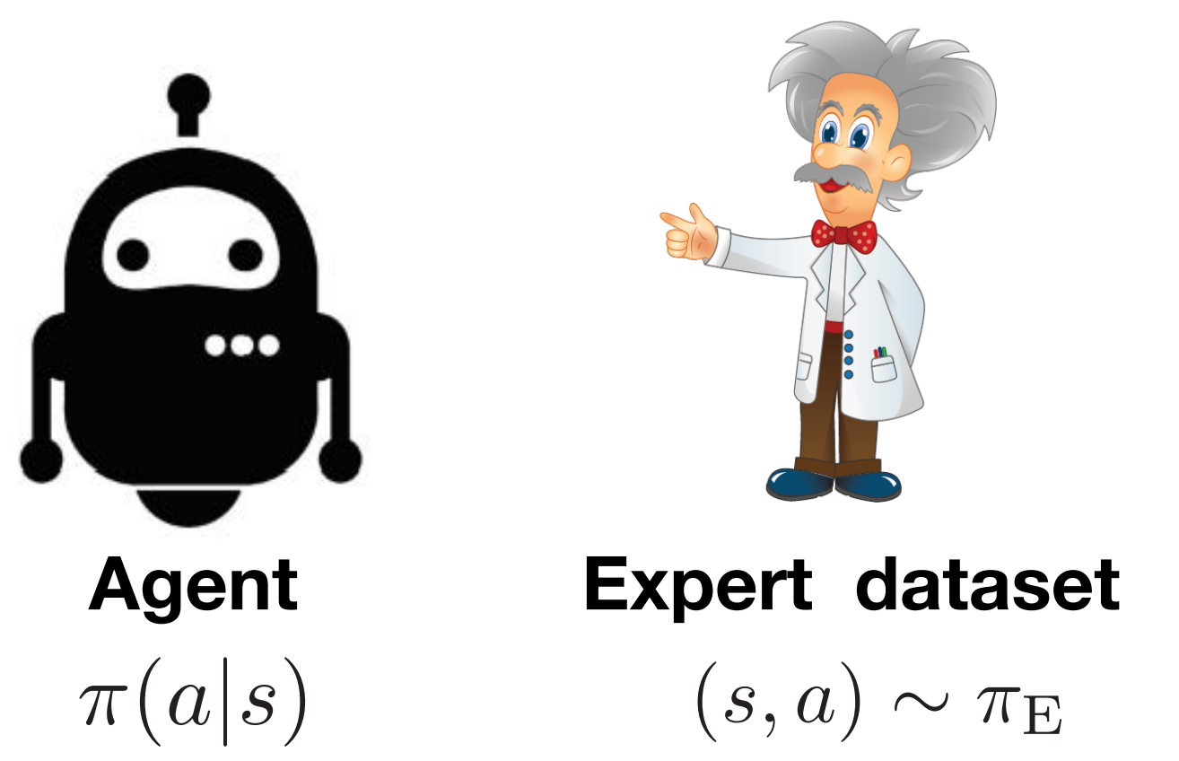
Tian Xu[1], Ziniu Li[2], and Yang Yu[1]

1: National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China
2: Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Shenzhen, Shenzhen 518172, China
xut@lamda.nju.edu.cn, ziniuli@link.cuhk.edu.cn, yuy@nju.edu.cn

Background and Setup



(a) RL from delayed rewards.



(b) IL from expert dataset.

- IL problem: minimize the **policy value gap** $V_{\pi_E} - V_\pi$ with the given expert demonstrations $\mathcal{D}$ collected by π_E:

$$\mathcal{D} = \{\mathrm{tr} = (s_1, a_1, s_2, a_2, \cdots, s_H, a_H); a_h \sim \pi_E(\cdot|s_h)\}.$$

- BC minimizes the discrepancy between action distributions:

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

- GAIL minimizes the JS divergence between ρ_{π_E} and ρ_π:

$$\min_{\pi \in \Pi} D_{JS}(\rho_\pi, \rho_{\pi_E}) =$$
$$\min_{\pi \in \Pi} \max_{D} \mathbb{E}_{\rho_\pi} [\log D(s,a)] + \mathbb{E}_{\rho_{\pi_E}} \left[\log(1 - D(s,a)) \right]$$

Motivation and Main Contribution

BC suffers compounding error issue due to one-step action distribution matching. GAIL achieves much empirical success while its theoretical understandings need more studies.

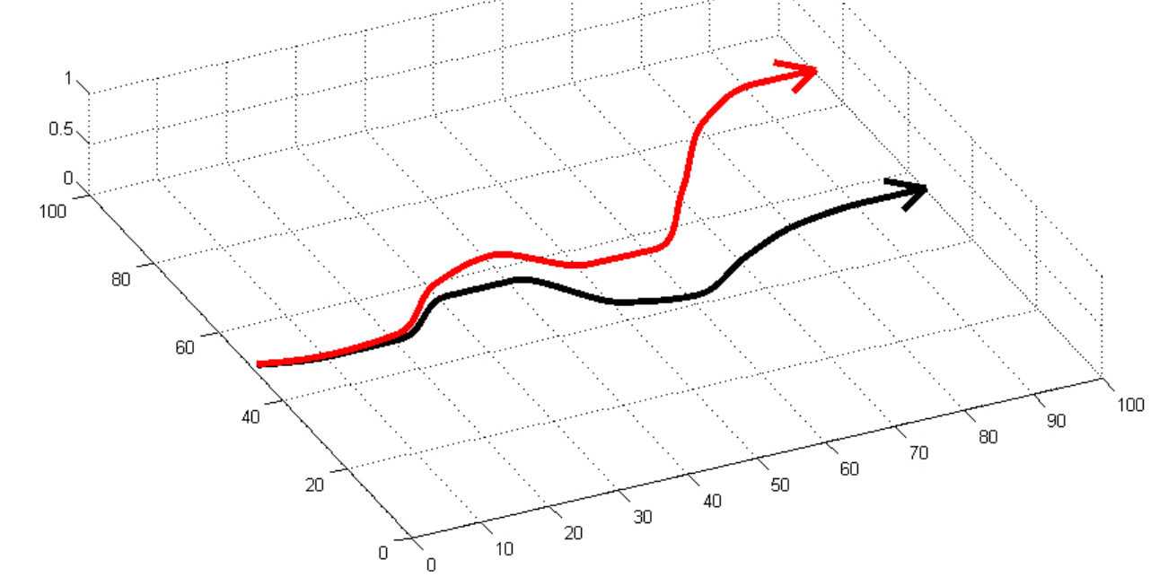


Figure 2. The compounding error issue

Theorem 1: Error Bound of BC

iven a stochastic expert policy π_E and an imitated policy π_{BC} with $\mathbb{E}_{s \sim d_{\pi_{\mathrm{E}}}} \left[D_{\mathrm{KL}} \left(\pi_{\mathrm{E}}(\cdot \mid s), \pi_{\mathrm{BC}}(\cdot \mid s) \right) \right] \leq \epsilon$, we have that

$$V_{\pi_{\mathrm{E}}} - V_{\pi_{\mathrm{BC}}} \leq \frac{\sqrt{2} R_{\max}}{(1-\gamma)^2} \sqrt{\epsilon}.$$

Theorem 2: Error Bound of GAIL

iven a stochastic expert policy π_E and an imitated policy π_{GA} with $d_{\mathcal{D}} \left(\widehat{\rho}_{\pi_E}, \widehat{\rho}_{\pi_{GA}} \right) \leq \min_{\pi \in \Pi} d_{\mathcal{D}} \left(\widehat{\rho}_{\pi_E}, \widehat{\rho}_\pi \right) + \varepsilon$, we have that

$$V_{\pi_{\mathrm{E}}} - V_{\pi_{\mathrm{BC}}} \leq \frac{\|r\|_{\mathcal{D}}}{1-\gamma} (\mathbf{Appr(II)} + \mathrm{Estm}(\mathcal{D}, m, \delta) + \hat{\epsilon}).$$

- GAIL enjoys a linear dependency on the effective planning horizon while BC suffers a quadratic dependency.

Proposition 1: Lower Bound without Interaction

iven expert dataset $D = \{(s_{\pi_E}^{(i)}, a_{\pi_E}^{(i)})\}_{i=1}^m$, for any algorithm Alg: $D \rightarrow \pi_I$, there exists an MDP $\mathcal{M}$ and a expert policy π_E,

$$V_{\pi_{\mathrm{E}}}^{\mathcal{M}} - V_{\pi_I}^{\mathcal{M}} \geq \min \left(\frac{1}{1-\gamma}, \frac{|\mathcal{S}|}{(1-\gamma)^2 m} \right).$$

- The quadratic dependency on the effective horizon can not be avoided for any algorithm when the interaction is not allowed.
- BC is already minimax optimal when the agent cannot interact with the environment.

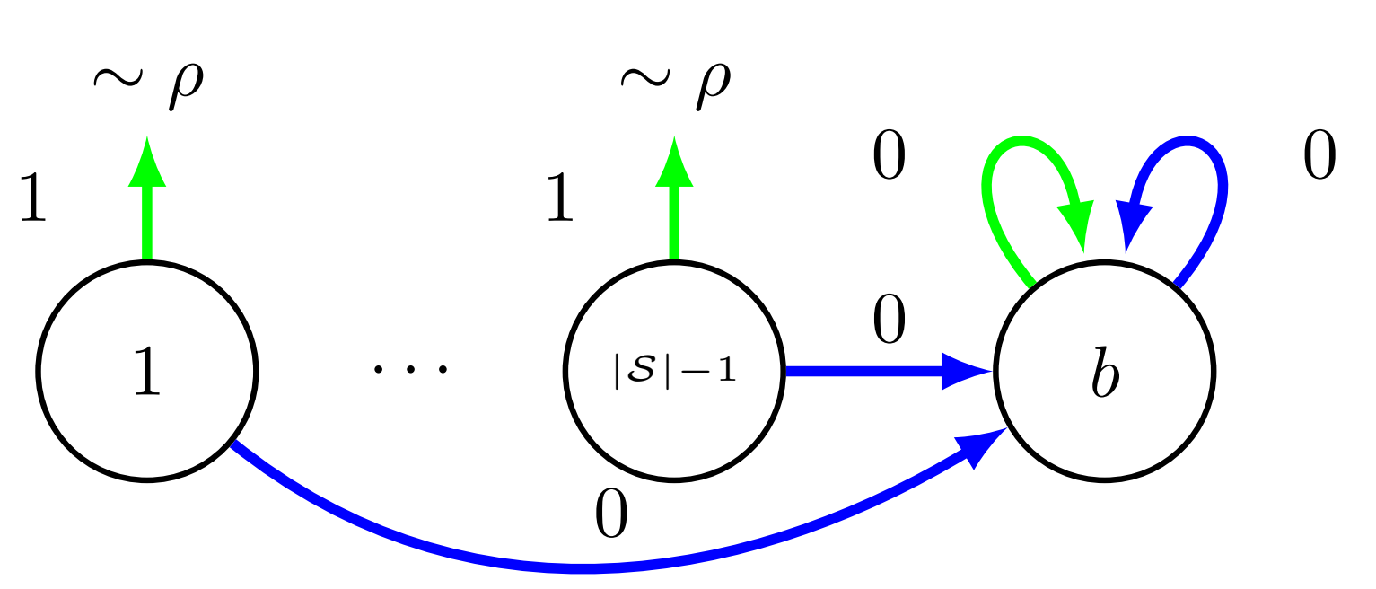
Proposition 2: Lower Bound with Interaction

hen the environment interaction is allowed, given expert dataset $D = \{(s_{\pi_E}^{(i)}, a_{\pi_E}^{(i)})\}_{i=1}^m$, for any algorithm Alg: $D \rightarrow \pi_I$, there exists an MDP $\mathcal{M}$ and a expert policy π_E,

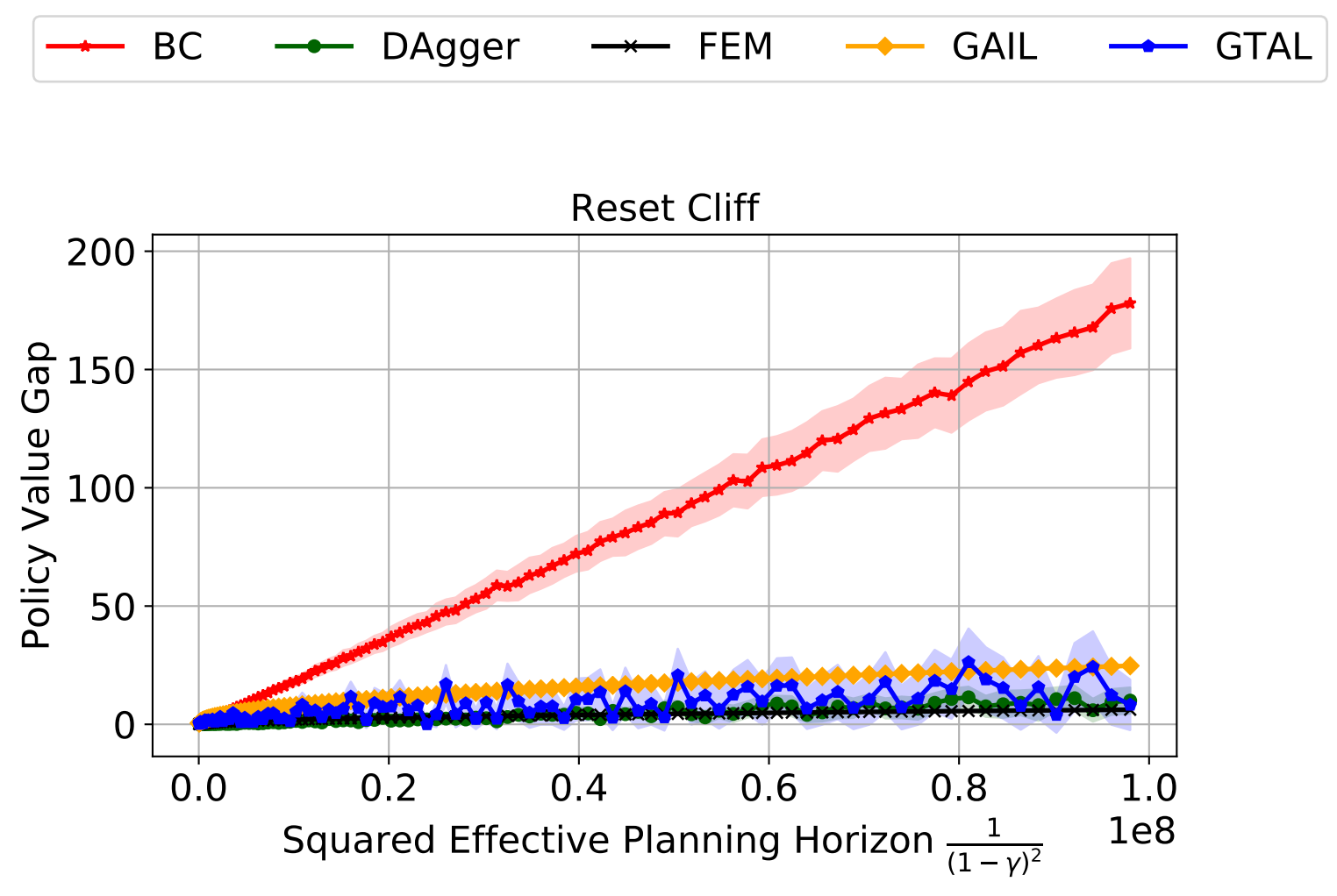
$$V_{\pi_{\mathrm{E}}}^{\mathcal{M}} - V_{\pi_I}^{\mathcal{M}} \geq \min \left(\frac{1}{1-\gamma}, \frac{|\mathcal{S}|}{(1-\gamma) m} \right).$$

- With the environment interaction, agent could have a knowledge of the transition function, which mitigates the compounding error issue.
- This helps explain why GAIL enjoys a linear dependency.

Empirical Study



(a) The Reset Cliff MDP



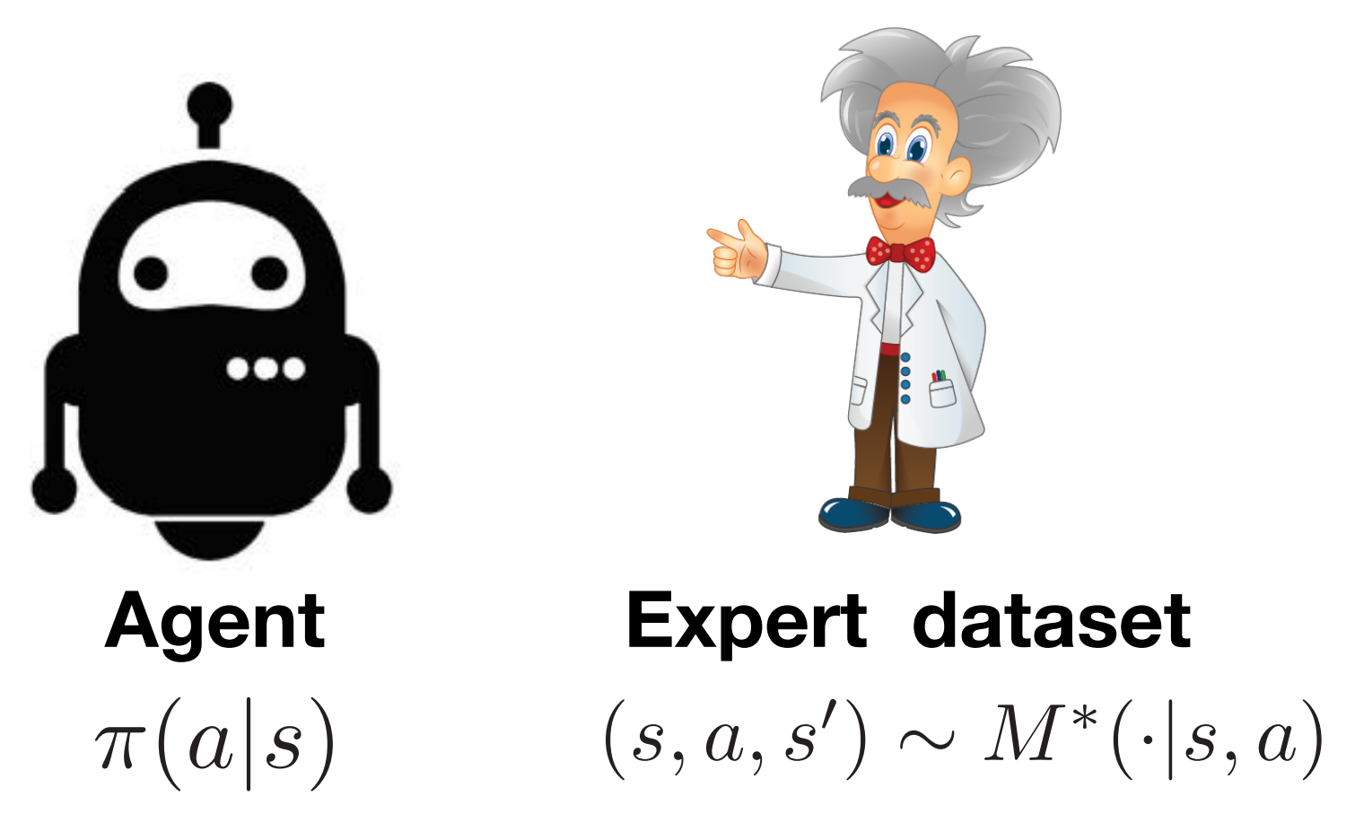
(b) On effective planning horizon.

Remark

- Reset Cliff highlights the compounding error issue. On states uncovered in expert demonstrations, BC randomly selects an action. It is very likely that the agent picks up a non-expert action, goes to the bad state and gets 0 reward forever.
- GAIL can avoid this issue via imitating expert state-action distribution. To see this, consider a good state (e.g., state 1) at time step h appears in the dataset, the agent must follow the expert actions at all time steps before h; otherwise, the agent will go to the bad state and cannot visit the observed good state at time step h, which violates the state-action distribution matching principle.

Error Bounds of Imitating Environments

By treating environment transition model as dual agent, learning the environment transition function can also be treated by imitation learning. We use the policy evaluation error $\left| V_\pi^{M^*} - V_\pi^{M_\theta} \right|$ to measure the performance.



Imitate Environment via BC:

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

Lemma 1

iven a learned transition model M_θ by BC with $\mathbb{E}_{(s,a) \sim \rho_{\pi_D^{M^*}}} \left[D_{\mathrm{KL}} \left(M^*(\cdot \mid s,a), M_\theta(\cdot \mid s,a) \right) \right] \leq \varepsilon_m$, for an arbitrary policy with bounded divergence ε_π, we have

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

Imitate Environment via GAIL:

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

where μ^M is the state-action-next-state distribution under M.

Lemma 2

iven a learned transition model M_θ by GAIL with $D_{JS}(\mu^{M_\theta}, \mu^{M^*}) \leq \varepsilon_m$, we have

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

Conclusion

This paper presents the error bounds of BC and GAIL for imitating policies and environments, mainly showing that GAIL can achieve a linear dependency on the effective planning horizon while BC has a quadratic dependency. The lower bound indicates that the quadratic dependency on the effective planning horizon is minimax optimal when the agent can not interact with the environment. Our analysis also suggests that a GAIL-style transition learner can replace the BC-like learner to improve the performance, which shed light on model-based reinforcement learning.