Off-Policy Imitation Learning from Observations

Objective of GAIL

$$\min_{\pi} \max_{x} J_{\mathrm{GAIL}}(\pi, x) := \mathbb{E}_{\mu^{E}(s,a)}[\log(x(s,a))] + \mathbb{E}_{\mu^{\pi}(s,a)}[\log(1 - x(s,a))].$$

- Requires expert action => Learning from observations
  - (BCO, GAILfO)
- On-policy (low sample efficiency)
  - off-policy algorithms rather than TRPO (SQIL, DAC)
  - offline reward estimation (PWIL)

Connection between LfD and LfO

Learning from Demonstration

$$\min J_{\mathrm{LfD}}(\pi) := \mathbb{D}_{\mathbf{KL}}[\mu^{\pi}(s,a)||\mu^{E}(s,a)]$$

Learning from Observation

$$\min J_{\mathrm{LfO}}(\pi) := \mathbb{D}_{\mathbf{KL}}[\mu^{\pi}(s,s')||\mu^{E}(s,s')].$$

	State Distribution	State-Action Distribution	Joint Distribution	Transition Distribution	Inverse-Action Distribution
Notation	$\mu^{\pi}(s)$	$\mu^{\pi}(s,a)$	$\mu^{\pi}(s,a,s')$	$\mu^{\pi}(s,s')$	$\mu^{\pi}(a\|s,s')$
Support	$\mathcal{S}$	$\mathcal{S}\times\mathcal{A}$	$\mathcal{S}\times\mathcal{A}\times\mathcal{S}$	$\mathcal{S}\times\mathcal{S}$	$\mathcal{A}\times\mathcal{S}\times\mathcal{S}$
Definition	$(1-\gamma)\sum_{t=1}^{\infty}\gamma^{t}\mu_{t}^{\pi}(s)$	$\mu^{\pi}(s)\pi(a\|s)$	$\mu^{\pi}(s,a)P(s'\|s,a)$	$\int_{\mathcal{A}}\mu^{\pi}(s,a,s')da$	$\frac{\mu^{\pi}(s,a)P(s'\|s,a)}{\mu^{\pi}(s,s')}$

Table 1: Summarization on different stationary distributions, with $\mu_t^{\pi}(s) = p(s_t = s|s_0 \sim p_0(\cdot), a_i \sim \pi(\cdot|s_i), s_{i+1} \sim P(\cdot|s_i, a_i)),\ \forall i < t)$.

Connection between LfD and LfO

$$\mathbb{D}_{\mathbf{KL}}[\mu^{\pi}(a|s,s')||\mu^{E}(a|s,s')] = \mathbb{D}_{\mathbf{KL}}[\mu^{\pi}(s,a)||\mu^{E}(s,a)] - \mathbb{D}_{\mathbf{KL}}[\mu^{\pi}(s,s')||\mu^{E}(s,s')]. \tag{3}$$

Remark 1. *In a non-injective MDP, the discrepancy of* $\mathbb{D}_{\mathbf{KL}}[\mu^{\pi}(a|s,s')||\mu^{E}(a|s,s')]$ *cannot be optimized without knowing expert actions. In a deterministic and injective MDP, it satisfies that* $\forall\, \pi : \mathcal{S} \to \mathcal{A},\ \mathbb{D}_{\mathbf{KL}}[\mu^{\pi}(a|s,s')||\mu^{E}(a|s,s')] = 0.$

$$\mathbb{D}_{\mathbf{KL}}\left[\mu^{\pi}(s,s')||\mu^{E}(s,s')\right] \leq \mathbb{E}_{\mu^{\pi}(s,s')}\left[\log\frac{\mu^{R}(s,s')}{\mu^{E}(s,s')}\right] + \mathbb{D}_{\mathbf{KL}}\left[\mu^{\pi}(s,a)||\mu^{R}(s,a)\right].$$

$$\mathbb{D}_{\mathbf{KL}}[P||Q] \leq \mathbb{D}_{f}[P||Q]$$

$$\mathbb{D}_{f}\left[\mu^{\pi}(s,a)||\mu^{R}(s,a)\right].$$

Similar “Surrogate Distribution”

surrogate expert buffer: $D_s = \{(s,a)_i\}$

$$W_1(\rho^*(s), \tilde{\rho}(s)) = \sup_{\|g_\phi\|_L \leq 1} \mathbb{E}_{s\sim\rho^*}[g_\phi(s)] - \mathbb{E}_{s\sim\tilde{\rho}}[g_\phi(s)]$$

perform AIRL on D_S

$$\max_{\omega} \mathbb{E}_{(s,a)\sim\mathcal{B}}\left[\log \frac{e^{f_\omega(s,a)}}{e^{f_\omega(s,a)} + \pi_\theta(a|s)}\right] + \mathbb{E}_{(s,a)\sim\pi_\theta}\left[\log \frac{\pi_\theta(a|s)}{e^{f_\omega(s,a)} + \pi_\theta(a|s)}\right]$$

$$\min_{\pi} J_{\text{opolo}}(\pi) \equiv \max_{\pi} \mathbb{E}_{\mu^{\pi}(s,s')}\left[-\log \frac{\mu^{R}(s,s')}{\mu^{E}(s,s')}\right] - \mathbb{D}_f[\mu^{\pi}(s,a)||\mu^{R}(s,a)]$$

$$-\mathbb{D}_f[\mu^{\pi}(s,a)||\mu^{R}(s,a)] = \inf_{x:S\times A\to R} \mathbb{E}_{(s,a)\sim\mu^{\pi}}[-x(s,a)] + \mathbb{E}_{(s,a)\sim\mu^{R}}[f_*(x(s,a))]$$

$$\equiv \max_{\pi} \min_{x:\mathcal{S}\times\mathcal{A}\to R} J_{\text{opolo}}(\pi, x) := \mathbb{E}_{\mu^{\pi}(s,a,s')}\left[\log \frac{\mu^{E}(s,s')}{\mu^{R}(s,s')} - x(s,a)\right] + \mathbb{E}_{\mu^{R}(s,a)}[f_*(x(s,a))].$$

pseudo-reward

$$Q(s,a) = \mathbb{E}_{s' \sim P(\cdot|s,a), a' \sim \pi(\cdot|s')} \left[-x(s,a) + \log \frac{\mu^E(s,s')}{\mu^R(s,s')} + \gamma Q(s',a') \right].$$

$$Q(s,a) = -x(s,a) + \mathbb{E}_{s' \sim P(\cdot|s,a), a' \sim \pi(\cdot|s')} \left[\log \frac{\mu^E(s,s')}{\mu^R(s,s')} + \gamma Q(s',a') \right] = -x(s,a) + \mathcal{B}^\pi Q(s,a).$$

$$\begin{aligned}
&\max_\pi \min_{x: S \times A \to R} J_{\text{opolo}}(\pi, x) \equiv \max_\pi \min_{Q: S \times A \to R} J_{\text{opolo}}(\pi, Q) \\
&:= \mathbb{E}_{(s,a,s') \sim \mu^\pi(s,a,s')} [\log \frac{\mu^E(s,s')}{\mu^R(s,s')} - (\mathcal{B}^\pi Q - Q)(s,a)] + \mathbb{E}_{(s,a) \sim \mu^R(s,a)} [f_*((\mathcal{B}^\pi Q - Q)(s,a))] \\
&= (1-\gamma) \mathbb{E}_{s_0 \sim p_0, a_0 \sim \pi(\cdot|s_0)} [Q(s_0,a_0)] + \mathbb{E}_{(s,a) \sim \mu^R(s,a)} [f_*((\mathcal{B}^\pi Q - Q)(s,a))]. \qquad (8)
\end{aligned}$$

no on-policy expectation left

Implementation Details

optimize the ratio using GAN

$$\max_{D:\mathcal{S}\times\mathcal{S}\to\mathbb{R}} \mathbb{E}_{(s,s')\sim\mu^E(s,s')}\Big[\log(D(s,s'))\Big] + \mathbb{E}_{(s,s')\sim\mu^R(s,s')}\Big[\log(1-D(s,s'))\Big];$$

additional regularization

$$\mathbb{D}_{\mathbf{KL}}[\pi_E(a|s)||\pi(a|s)] = \mathbb{D}_{\mathbf{KL}}[\mu^E(s'|s)||\mu^\pi(s'|s)] + \mathbb{D}_{\mathbf{KL}}[\mu^E(a|s,s')||\mu^\pi(a|s,s')],$$

$$\max_{P_I:\mathcal{S}\times\mathcal{S}\to\mathcal{A}} -\mathbb{D}_{\mathbf{KL}}[\mu^R(a|s,s')||P_I(a|s,s')] \equiv \max_{P_I:\mathcal{S}\times\mathcal{S}\to\mathcal{A}} \mathbb{E}_{(s,a,s')\sim\mu^R(s,a,s')}[\log P_I(a|s,s')].$$

Algorithm 1 *O*ff-*PO*licy *L*earning from *O*bservations (*OPOLO*)

Input: expert observations $\mathcal{R}_E$, off-policy-transitions $\mathcal{R}$, initial states $\mathcal{S}_0$, f- function, policy π_θ, critic Q_ϕ, discriminator D_w, inverse action model $P_{I\varphi}$, learning rate α.

for $n = 1, \ldots$ **do**

sample trajectory $\tau \sim \pi_\theta$, $\mathcal{R} \leftarrow \mathcal{R} \cup \tau$

update D_w: $w \leftarrow w + \alpha \hat{\mathbb{E}}_{(s,s')\sim\mathcal{R}_E}[\nabla_w \log(D_w(s,s')] + \hat{\mathbb{E}}_{(s,s')\sim\mathcal{R}}[\nabla_w \log(1 - D_w(s,s'))]$.

set $r(s,s') = -\log(1 - D_w(s,s'))$.

update $P_{I\varphi}$: $\varphi \leftarrow \varphi + \alpha \hat{\mathbb{E}}_{(s,a,s')\sim\mathcal{R}}[\nabla_\varphi \log(P_{I\varphi}(a|s,s'))]$.

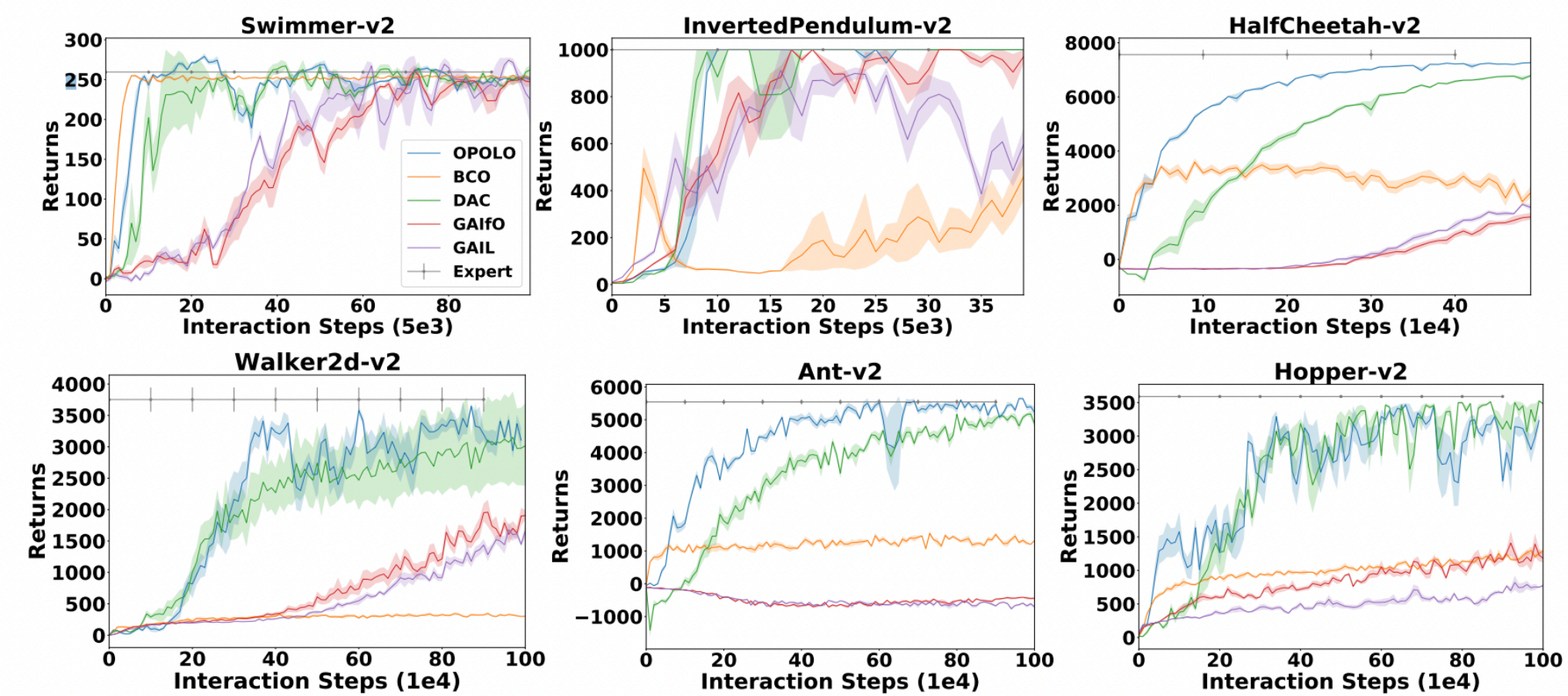
update π_θ and Q_ϕ :

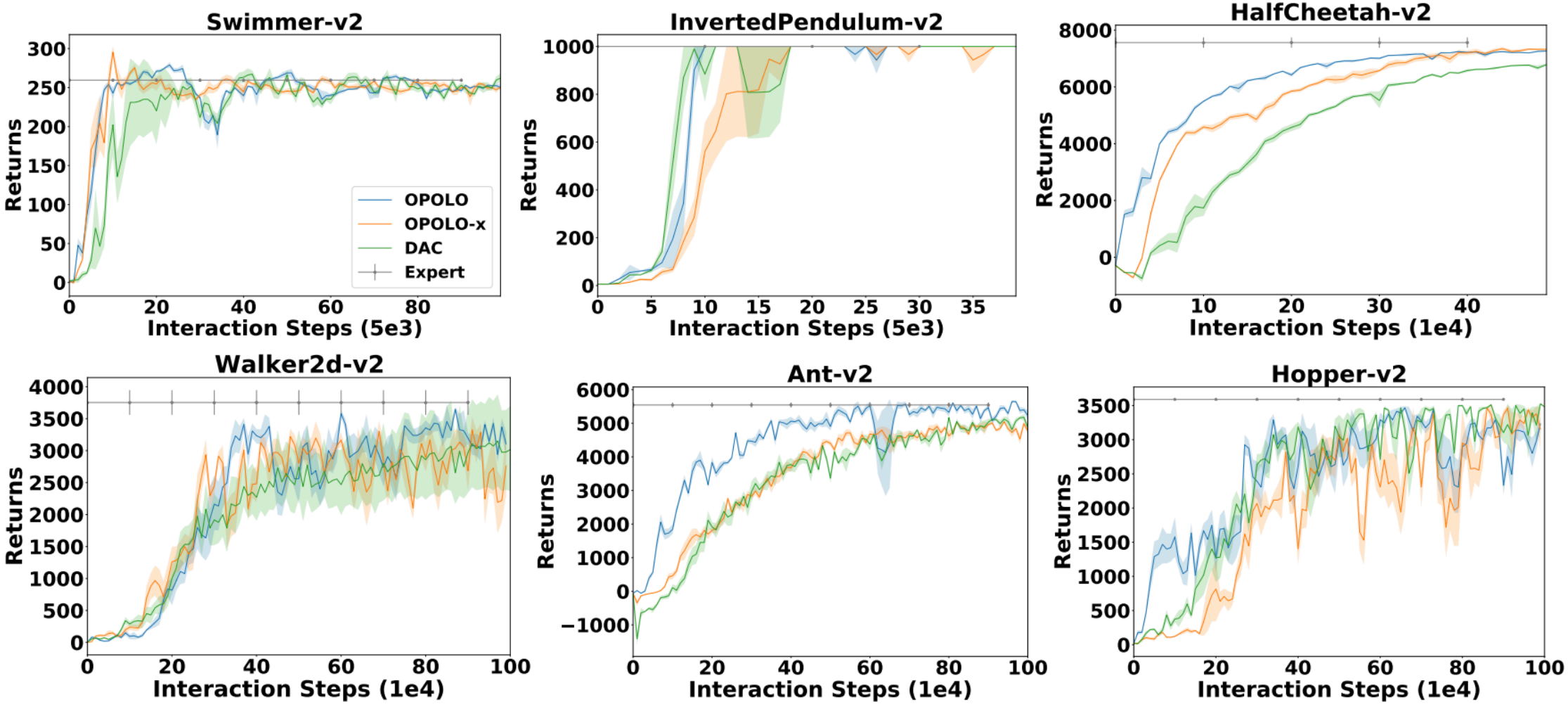
$$J(\pi_\theta, Q_\phi) = (1-\gamma)\hat{\mathbb{E}}_{s\sim\mathcal{S}_0}[Q_\phi(s, \pi_\theta(s))] + \hat{\mathbb{E}}_{(s,a,s')\sim\mathcal{R}}\left[f^*\left(r(s,s') + \gamma Q_\phi(s', \pi_\theta(s')) - Q_\phi(s,a)\right)\right].$$

$$J_{\mathrm{Reg}}(\pi_\theta) = \mathbb{E}_{(s,s')\sim\mathcal{R}_E, a\sim P_{I\varphi}(\cdot|s,s')}[\log \pi_\theta(a|s)].$$

$$\phi \leftarrow \phi - \alpha J_{\nabla\phi}(\pi_\theta, Q_\phi); \quad \theta \leftarrow \theta + \alpha\left(J_{\nabla\theta}(\pi_\theta, Q_\phi) + J_{\nabla\theta} J_{\mathrm{Reg}}(\pi_\theta)\right).$$

end for


Swimmer-v2
InvertedPendulum-v2
HalfCheetah-v2
Walker2d-v2
Ant-v2
Hopper-v2
Returns
Interaction Steps (5e3)
Interaction Steps (1e4)
OPOLO
BCO
DAC
GAIfO
GAIL
Expert


Swimmer-v2
InvertedPendulum-v2
HalfCheetah-v2
Walker2d-v2
Ant-v2
Hopper-v2
Returns
Interaction Steps (5e3)
Interaction Steps (1e4)
OPOLO
OPOLO-x
DAC
Expert

http://lamda.nju.edu.cn

End