Meta-Inverse Reinforcement Learning with Probabilistic Context Variables

Lantao Yu*, Tianhe Yu*, Chelsea Finn, Stefano Ermon
Department of Computer Science, Stanford University

Stanford ARTIFICIAL INTELLIGENCE

Highlights

- We aims at addressing two key limitations of existing inverse reinforcement learning (IRL) methods:
  - Learning reward functions from scratch and requiring large numbers of demonstrations to correctly infer the reward for each task.
  - Assuming demos are for one isolated task, while in practice it is more natural and scalable to obtain heterogeneous demos.
- We propose a new meta-inverse reinforcement learning framework based on latent probabilistic context variables termed PEMIRL.
- PEMIRL is capable of learning rewards from unstructured, multi-task demonstration data, and critically, use this experience to infer robust rewards for new, structurally-similar tasks from a single demonstration.
- We demonstrate the effectiveness of our approach compared to state-of-the-art imitation and inverse reinforcement learning methods on multiple continuous control tasks.

Backgrounds

Inverse RL Basic Principle: find a reward function r_ω that explains the expert behaviors. *(ill-defined problem)*
Maximum Entropy Inverse RL (MaxEnt IRL) (Ziebart et al., 2008) provides a general probabilistic framework that solves the reward ambiguity problem:

$$p_\omega(\tau) \propto \left[\eta(s_1)\prod_{t=1}^{T} P(s_{t+1}|s_t, a_t)\right] \exp\left(\sum_{t=1}^{T} r_\omega(s_t, a_t)\right), \quad \max_\omega \mathbb{E}_{\pi_E}[\log p_\omega(\tau)] = \mathbb{E}_{\tau\sim\pi_E}\left[\sum_{t=1}^{T} r_\omega(s_t, a_t)\right] - \log Z_\omega$$

where Z_ω is the *intractable* partition function, *i.e.*, an integral over all possible trajectories.
Adversarial Inverse RL (AIRL) (Fu et al., 2017) provides an efficient sampling-based approximation to MaxEnt IRL, with a special parameterization for discriminator that allows us to extract reward functions at optimality:

$$D_{\omega,\phi}(s, a, s') = \frac{\exp(f_{\omega,\phi}(s, a, s'))}{\exp(f_{\omega,\phi}(s, a, s')) + \pi(a|s)}, \quad f_{\omega,\phi}(s, a, s') = r_\omega(s, a) + \gamma h_\phi(s') - h_\phi(s)$$

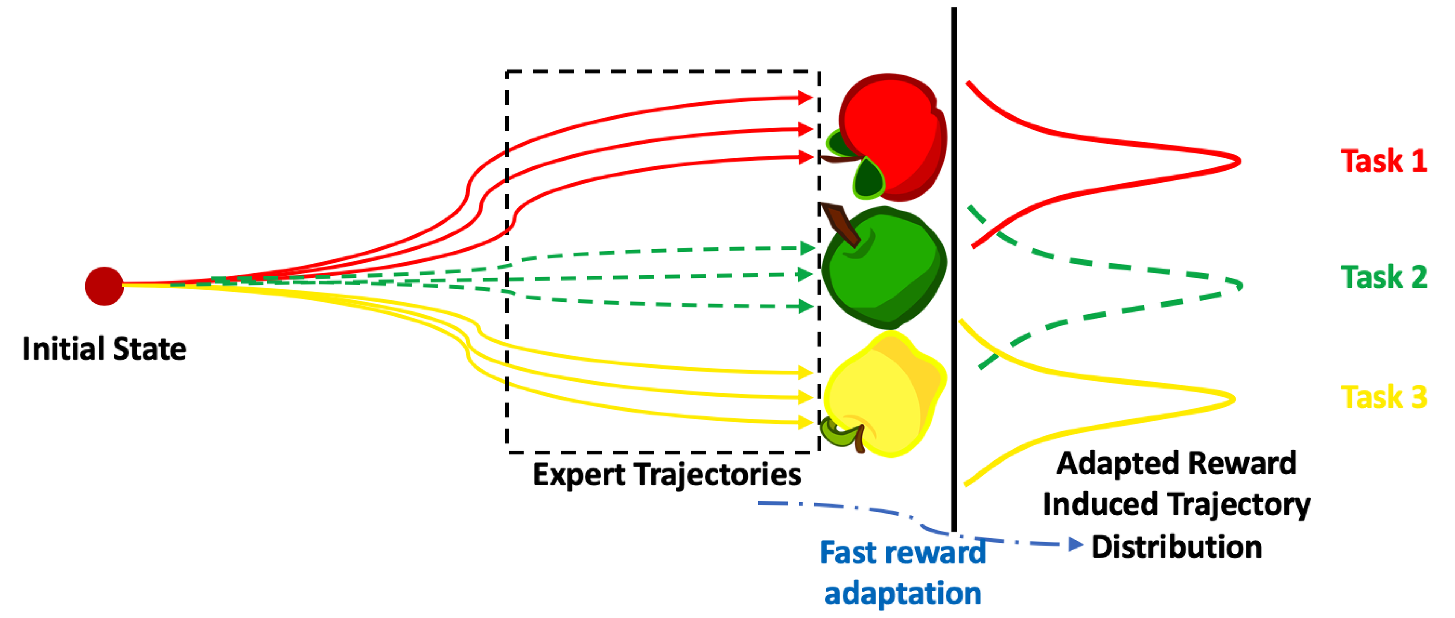
Under certain conditions, $r_\omega(s, a)$ is guaranteed to recover the ground-truth reward up to a constant.

Context-based Meta-Learning & Inverse Reinforcement Learning

Generalizing the notion of MDP with a probabilistic context variables $m \in \mathcal{M}$, where $\mathcal{M}$ is the (discrete or continuous) value space of m. We use $p(m)$ to denote the prior distribution over m.

- Context-dependent reward function $r : \mathcal{S} \times \mathcal{A} \times \mathcal{M} \to \mathbb{R}$; Context-dependent policy $\pi : \mathcal{S} \times \mathcal{M} \to \mathcal{P}(\mathcal{A})$.
- Expert policy: $\pi_E = \arg\max_\pi \mathbb{E}_{m\sim p(m),\ (s_{1:T}, a_{1:T})\sim p_\pi(\cdot|m)}\left[\sum_{t=1}^{T} r(s_t, a_t, m) - \log \pi(a_t|s_t, m)\right]$
- Marginal trajectory distribution of expert: $p_{\pi_E}(\tau) = \int_{\mathcal{M}} p(m)\prod_{t=1}^{T}\pi_E(a_t|s_t, m)P(s_{t+1}|s_t, a_t)dm$

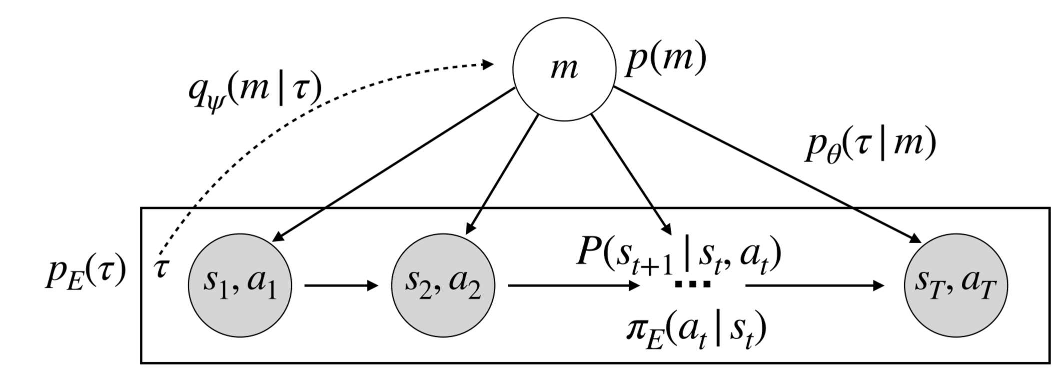
Meta-Inverse Reinforcement Learning (Meta-IRL): Given a set of **unstructured** demonstrations *i.i.d.* sampled from $p_{\pi_E}(\tau)$, **meta-learn** an inference model $q(m|\tau)$ and a reward function $f(s, a, m)$, such that given some new demonstration τ_E generated by sampling $m' \sim p(m), \tau_E \sim p_{\pi_E}(\tau|m')$, with $\hat{m}$ being inferred as $\hat{m} \sim q(m|\tau_E)$, the learned reward function $f(s, a, \hat{m})$ and the ground-truth reward $r(s, a, m')$ will induce the same set of optimal policies.



Meta-IRL with Probabilistic Context Variables

Under the framework of MaxEnt IRL, we first parametrize two components:

- Context variable inference model $q_\psi(m|\tau)$.
- Context-dependent reward function $f_\theta(s, a, m)$.



We would like to maximize the the mutual information between two random variables m and τ under joint distribution $p_\theta(m, \tau) = p(m)p_\theta(\tau|m)$: $I_{p_\theta}(m; \tau) = \mathbb{E}_{m\sim p(m), \tau\sim p_\theta(\tau|m)}[\log p_\theta(m|\tau) - \log p(m)]$, subject to:

- Desideratum 1. Matching conditional distributions: $\mathbb{E}_{p(m)}\left[D_{\mathrm{KL}}(p_{\pi_E}(\tau|m)||p_\theta(\tau|m))\right] = 0$
- Desideratum 2. Matching posterior distributions: $\mathbb{E}_{p_\theta(\tau)}[D_{\mathrm{KL}}(p_\theta(m|\tau)||q_\psi(m|\tau))] = 0$

With Lagrangian duality and Lagrangian multipliers taking specific values, we have the relaxed problem:

$$\min_{\theta,\psi} \mathbb{E}_{p(m)}\left[D_{\mathrm{KL}}(p_{\pi_E}(\tau|m)||p_\theta(\tau|m))\right] + \mathbb{E}_{p_\theta(m,\tau)}\left[\log\frac{p(m)}{p_\theta(m|\tau)} + \log\frac{p_\theta(m|\tau)}{q_\psi(m|\tau)}\right]$$

$$\equiv \max_{\theta,\psi} -\mathbb{E}_{p(m)}\left[D_{\mathrm{KL}}(p_{\pi_E}(\tau|m)||p_\theta(\tau|m))\right] + \mathbb{E}_{m\sim p(m), \tau\sim p_\theta(\tau|m)}[\log q_\psi(m|\tau)]$$

We can leverage adversarial reward learning (AIRL) to optimize this objective.

Experiments

Empirical evaluations in various continuous control tasks demonstrate the effectiveness of our framework:

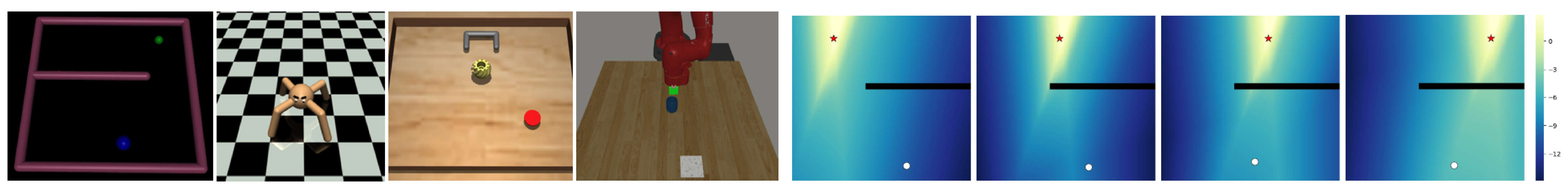


Figure: **Experimental domains** (left) and **visualizations of adapted rewards on Point-Maze** (right).

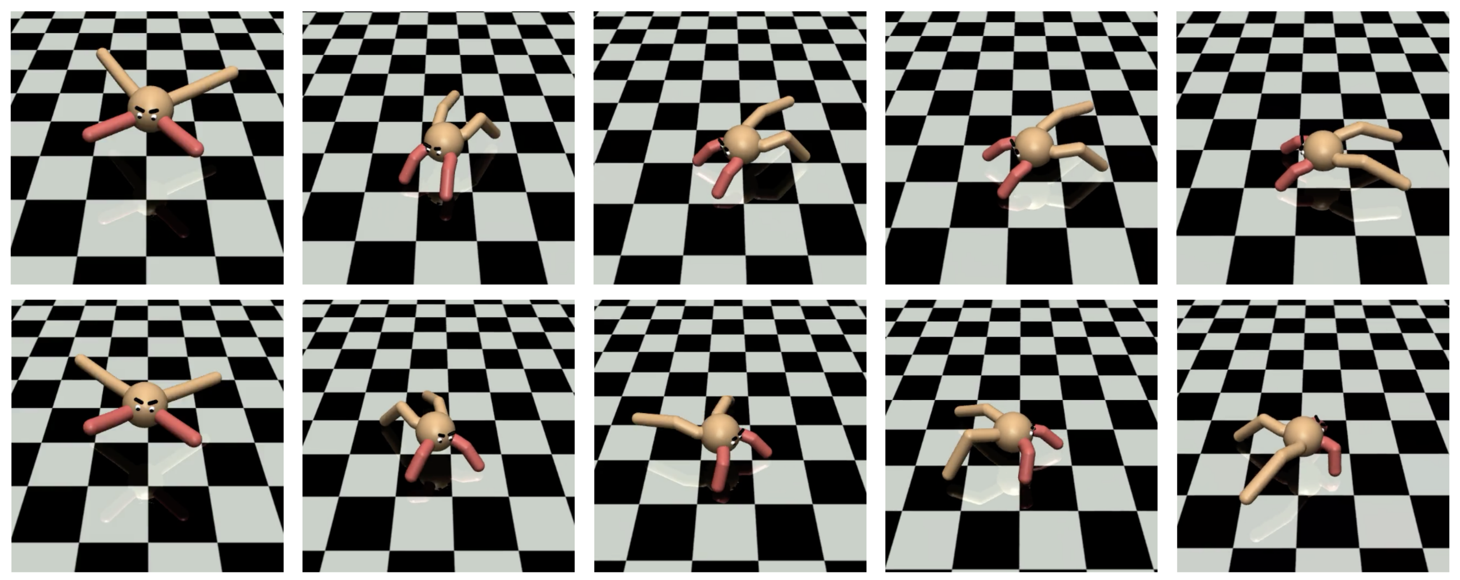


Figure: Results of the disabled ant running forward and backward respectively with adapted rewards.

	Method	Point-Maze-Shift	Disabled-Ant
Policy Generalization	Meta-IL	-28.61 ± 3.71	-27.86 ± 10.31
	Meta-InfoGAIL	-29.40 ± 3.05	-51.08 ± 4.81
	PEMIRL	-28.93 ± 3.59	-46.77 ± 5.54
Reward Adaptation	AIRL	-29.07 ± 4.12	-76.21 ± 10.35
	Meta-InfoGAIL	-29.72 ± 3.11	-38.73 ± 6.41
	PEMIRL (ours)	$\mathbf{-9.04} \pm 1.09$	$\mathbf{152.62} \pm 11.75$
	Expert	-5.37 ± 0.86	331.17 ± 17.82

Figure: Results on direct policy generalization and reward adaptation to challenging situations.