Inverse Constrained Reinforcement Learning

Usman Anwar[1], Shehryar Malik[1], Alireza Aghasi[2], Ali Ahmed[1]

[1]Information Technology University, Lahore.
[2]Georgia State University, USA.

July 17, 2021

Motivation

- AI safety.
- Important for agent to know what *never* to do.
- Manual constraint specification not possible.
- We study the problem of 'constraint inference' in perspective of embodied agents trained through reinforcement learning.

Motivation

- AI safety.
- Important for agent to know what *never* to do.
- Manual constraint specification not possible.
- We study the problem of 'constraint inference' in perspective of embodied agents trained through reinforcement learning.

Contributions

- Learning constraints in high dimensional continuous settings.
- Transfer to new agents across morphology and dynamics.

Preliminaries & Notation

- $\mathcal{M}$ represents a nominal MDP.
- Augmenting $\mathcal{M}$ with some constraint set $\mathcal{C}$ results in a constrained MDP $\mathcal{M}^{\mathcal{C}}$.
- $\mathcal{M}$ and $\mathcal{M}^{\mathcal{C}}$ have the same reward function but may differ in their optimal policies $\pi_{\mathcal{M}}$ and $\pi_{\mathcal{M}^{\mathcal{C}}}$.
- We represent true constraint set with $\mathcal{C}^*$, which is known to the demonstrating agent, but unknown to the RL agent.

Inverse Constrained Reinforcement Learning

Find the constraint set which best explains the demonstrations $\mathcal{D}$ and nominal MDP $\mathcal{M}$.

$$\mathcal{C}^* \leftarrow \operatorname*{argmax}_{\mathcal{C}} p_{\mathcal{M}}(\mathcal{D}|\mathcal{C}). \tag{1}$$

Inverse Constrained Reinforcement Learning

Maximum Entropy Model

We assume that all trajectories τ in the dataset $\mathcal{D}$ are distributed according to the maximum entropy distribution.

$$\pi_{\mathcal{M}^{\mathcal{C}}}(\tau) = \frac{\exp(\beta r(\tau))}{Z_{\mathcal{M}^{\mathcal{C}}}} \mathbb{1}^{\mathcal{M}^{\mathcal{C}}}(\tau). \tag{2}$$

where

- $\mathbb{1}^{\mathcal{M}^{\mathcal{C}}}$ is an indicator function which is 0 if τ belongs to constraint set $\mathcal{C}$
- Indicator function distributes over individual state action pairs, i.e.,

$$\mathbb{1}^{\mathcal{M}^{\mathcal{C}}}(\tau) = \prod_{i=1}^{T} \mathbb{1}^{\mathcal{M}^{\mathcal{C}}}(s_t, a_t).$$

Inverse Constrained Reinforcement Learning

Maximum Entropy Model

We assume that all trajectories τ in the dataset $\mathcal{D}$ are distributed according to the maximum entropy distribution.

$$\pi_{\mathcal{M}^{\mathcal{C}}}(\tau) = \frac{\exp(\beta r(\tau))}{Z_{\mathcal{M}^{\mathcal{C}}}} \mathbb{1}^{\mathcal{M}^{\mathcal{C}}}(\tau). \tag{2}$$

where

- $\mathbb{1}^{\mathcal{M}^{\mathcal{C}}}$ is an indicator function which is 0 if τ belongs to constraint set $\mathcal{C}$
- Indicator function distributes over individual state action pairs, i.e.,

$$\mathbb{1}^{\mathcal{M}^{\mathcal{C}}}(\tau) = \prod_{i=1}^{T} \mathbb{1}^{\mathcal{M}^{\mathcal{C}}}(s_t, a_t).$$

Observation

Learning $\mathbb{1}^{\mathcal{M}^{\mathcal{C}}}$ is equivalent to learning the constraint set $\mathcal{C}$.

Inverse Constrained Reinforcement Learning

Final Objective

Use a classifier ζ_θ parametrized by θ to approximate the indicator function $\mathbb{1}^{\mathcal{M}^{\mathcal{C}}}(\tau)$:

$$\nabla_\theta \mathcal{L}(\theta) = \underbrace{\mathbb{E}_{\tau\sim\pi^{\mathcal{C}^*}}\left[\nabla_\theta \log \zeta_\theta(\tau)\right]}_{\text{expert}} - \overbrace{\mathbb{E}_{\hat{\tau}\sim\pi^{\zeta_\theta}}\left[\nabla_\theta \log \zeta_\theta(\hat{\tau})\right]}^{\text{RL agent}}, \tag{3}$$

Inverse Constrained Reinforcement Learning - Training Tricks

Regularizer

$$R(\theta) = \delta \sum_{\tau \sim \{\mathcal{D}, \mathcal{S}\}} [\zeta_\theta(\tau) - 1] \tag{4}$$

Importance Sampling

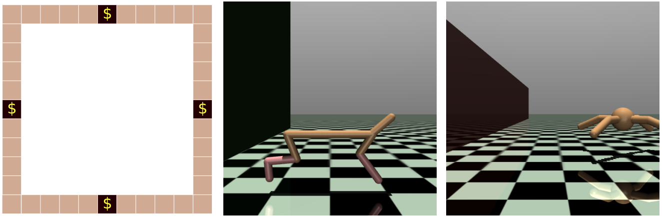
$$\omega(s_t, a_t) = \frac{\zeta_\theta(s_t, a_t)}{\zeta_{\bar{\theta}}(s_t, a_t)}. \tag{5}$$

KL Based Early Stopping

$$\begin{aligned} D_{\mathrm{KL}}(\pi_{\bar{\theta}} || \pi_\theta) &\leq 2 \log \bar{\omega} \\ D_{KL}(\pi_\theta || \pi_{\bar{\theta}}) &\leq \frac{\mathbb{E}_{\tau \sim \pi_{\bar{\theta}}} \left[(\omega(\tau) - \bar{\omega}) \log \omega(\tau) \right]}{\bar{\omega}}. \end{aligned} \tag{6}$$

Experiments

Results: Learning Constraints



(a) LapGridWorld (b) HalfCheetah (c) Ant

Figure: The environments used in the experiments for learning constraints.

Results: Learning Constraints

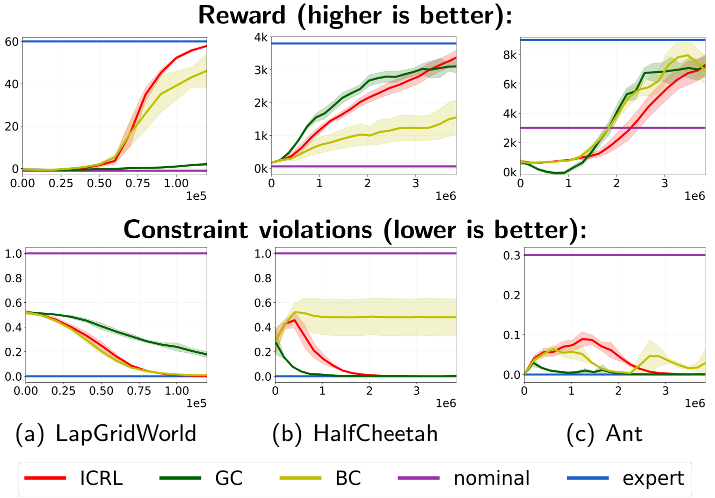
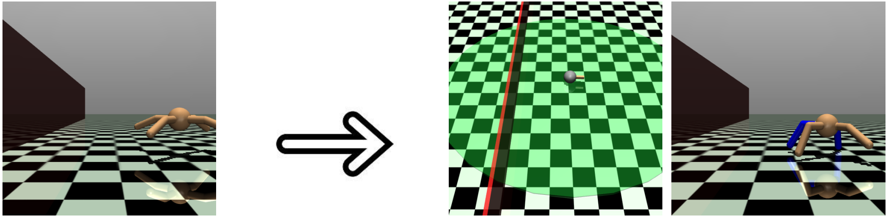


Figure: Performance of agents during training over several seeds (5 in LapGridWorld, 10 in others). The x-axis is the number of timesteps taken in the environment. The shaded regions correspond to the standard error.

Results: Transferring Constraints



(a) Ant (b) Point (c) AntBroken

Figure: Constraints learned in ant environment were transferred to point and ant broken environments.

Results: Transferring Constraints

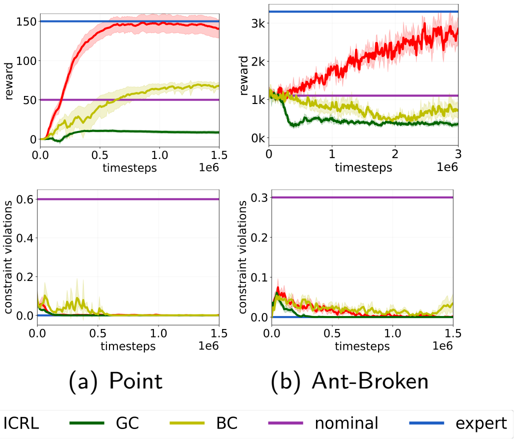


Figure: Transferring constraints. The x-axis is the number of timesteps taken in the environment. All plots were smoothed and averaged over 5 seeds. The shaded regions correspond to the standard error.

Ablation Studies

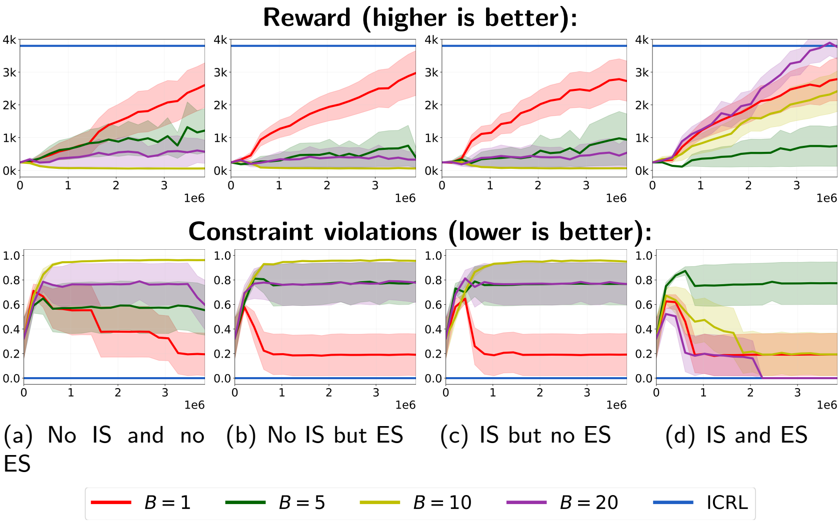


Figure: Ablation studies on the HalfCheetah environment. All plots were averaged over 5 seeds. IS refers to importance sampling and ES to early stopping. The x-axis corresponds to the number of timesteps the agent takes in the environment. Shaded regions correspond to the standard error.

Limitations & Future Work

- Maximum Causal Entropy & stochastic MDPs.
- Soft Constraints.
- Off-policy constraint learning.
- Robust imitation learning.

Thank You.