Understanding Constraint Inference in Safety-Critical Inverse Reinforcement Learning

Bo Yue[1], Shufan Wang[2], Ashish Gaurav[3,4], Jian Li[2], Pascal Poupart[3,4], Guiliang Liu[1*]

[1]School of Data Science, The Chinese University of Hong Kong, Shenzhen,

[2]Stony Brook University, [3]University of Waterloo, [4]Vector Institute

香港中文大學(深圳) The Chinese University of Hong Kong, Shenzhen

Stony Brook University

UNIVERSITY OF WATERLOO

VECTOR INSTITUTE | INSTITUT VECTEUR

Abstract

Background: Constraint inference is crucial in safety-critical decision-making processes.

Literature: Existing methods, ***Inverse Constrained Reinforcement Learning (ICRL)***, characterizes constraint learning as a inherently complex tri-level optimization problem.

Challenges: ***Can we implicitly embed constraint signals into reward functions and effectively solve this problem using a classic reward inference algorithm?***

Methodology: ***Inverse Reward Correction (IRC)*** *VS.* ***ICRL***

- IRC infers **a reward correction term**, which, when added to the reward function, ensures the optimality of the expert.
- ICRL infers **a cost function**, which, when serving as a constraint condition, ensures the optimality of the expert.

Takeaways:

- *Training Efficiency:* ***IRC*** > ***ICRL*** (***IRC*** learns constraint knowledge faster!)
- *Cross-Environment Transferability:* ***IRC*** < ***ICRL*** (***IRC*** fail to guarantee safety in target envs!)

Methods

Inverse Constraint Inference: Infer **constraint knowledge** followed by **expert policy** π^E.

IRC solver: $\pi^E = \max_{\pi} \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t [r + \Delta r](s_t, a_t)]$

(i) if $\pi^E(a|s) > 0$, then $Q^{r+\Delta r, \pi^E}_{\mathcal{M}\cup(r+\Delta r)}(s,a) = V^{r+\Delta r, \pi^E}_{\mathcal{M}\cup(r+\Delta r)}(s)$,

(ii) if $\pi^E(a|s) = 0$, then $Q^{r+\Delta r, \pi^E}_{\mathcal{M}\cup(r+\Delta r)}(s,a) \leq V^{r+\Delta r, \pi^E}_{\mathcal{M}\cup(r+\Delta r)}(s)$.

Implicit Model

π^E should be **optimal** regarding $r + \Delta r$

ICRL solver: $\pi^E = \max_{\pi} \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t [r - \lambda^* c](s_t, a_t)]$

(i) if $\pi^E(a|s) > 0$, $Q^{c,\pi^E}_{\mathcal{M}\cup c}(s,a) - V^{c,\pi^E}_{\mathcal{M}\cup c}(s) = 0$;

(ii) if $\pi^E(a|s) = 0$ and $A^{r,\pi^E}_{\mathcal{M}\cup c}(s,a) > 0$, $Q^{c,\pi^E}_{\mathcal{M}\cup c}(s,a) - V^{c,\pi^E}_{\mathcal{M}\cup c}(s) > 0$;

(iii) if $\pi^E(a|s) = 0$ and $A^{r,\pi^E}_{\mathcal{M}\cup c}(s,a) \leq 0$, $Q^{c,\pi^E}_{\mathcal{M}\cup c}(s,a) - V^{c,\pi^E}_{\mathcal{M}\cup c}(s) \leq 0$.

Explicit Model

π^E should be **optimal** regarding r under constraint condition $\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t)] \leq \epsilon$

Theoretical Findings

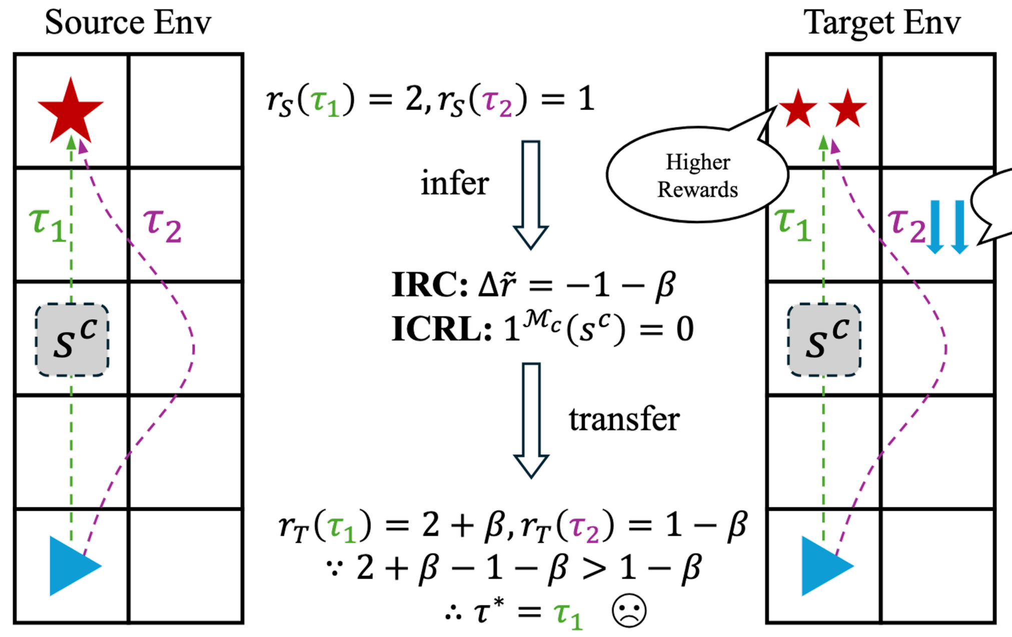
Sample Complexity

IRC solves $\max_{\Delta r} \min_{\pi} \mathcal{J}(\pi^E, r + \Delta r) - \mathcal{J}(\pi, r + \Delta r)$.

ICRL solves $\max_{c} \max_{\lambda} \min_{\pi} \mathcal{J}(\pi^E, r - \lambda c) - \mathcal{J}(\pi, r - \lambda c)$.

Bi-level to tri-level
A higher complexity of $1/(1-\gamma)^2$

Safety



Safety	IRC	ICRL
Hard Constraint	✗	✓
Soft Constraint	influenced by different rewards and transitions	influenced by different transitions

Source: $\mathcal{M}_c = \mathcal{M} \cup c^E = (\mathcal{S}, \mathcal{A}, P_{\mathcal{T}}, r, c^E)$ Target: $\mathcal{M}'_c = \mathcal{M}' \cup (c')^E = (\mathcal{S}, \mathcal{A}, P'_{\mathcal{T}}, r', (c')^E)$

$$Q^{r'}(s,a) - Q^{r}(s,a) = \Big[\underbrace{(Y')^{-1}(r' - r)}_{\text{reward transfer shift}} + \underbrace{(Y')^{-1}AQ^r}_{\text{transition transfer shift}} + \underbrace{(Y')^{-1}BQ^r}_{\text{expert policy transfer shift}}\Big](s,a).$$

$$Q^{c'}(s,a) - Q^{c}(s,a) = \Big[\underbrace{(Y')^{-1}AQ^c}_{\text{transition transfer shift}} + \underbrace{(Y')^{-1}BQ^c}_{\text{expert policy transfer shift}}\Big](s,a)$$

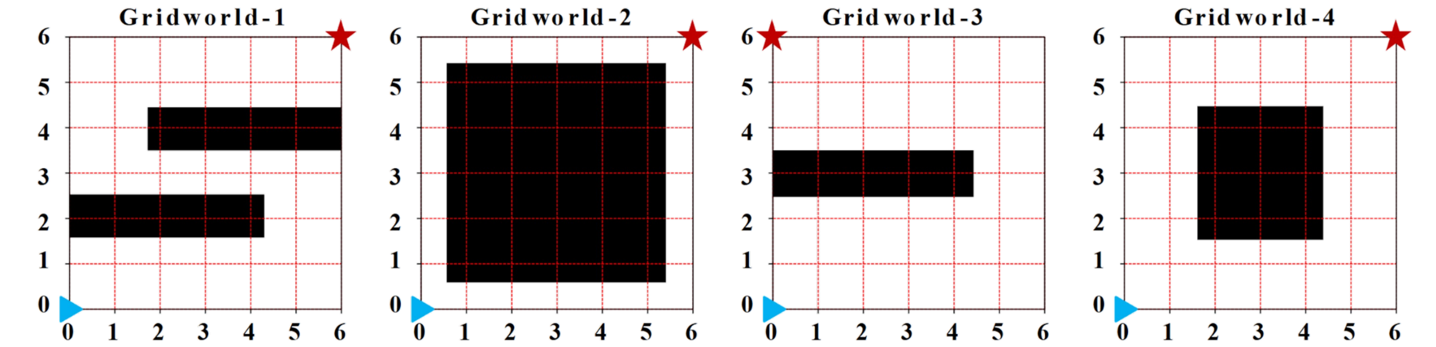
Optimality

$$\varepsilon = 2\max\left\{d_1^2 \sin\left(\theta_{\max}(P_{\mathcal{T}}', P_{\mathcal{T}})\right)^2/2, 2\varepsilon_1/\sigma_{\mathcal{R}}\right\}/\eta, \quad d_1 = \|[c^E - \hat{c}]_{\mathcal{U}_{P'_{\mathcal{T}}}}\|_2$$

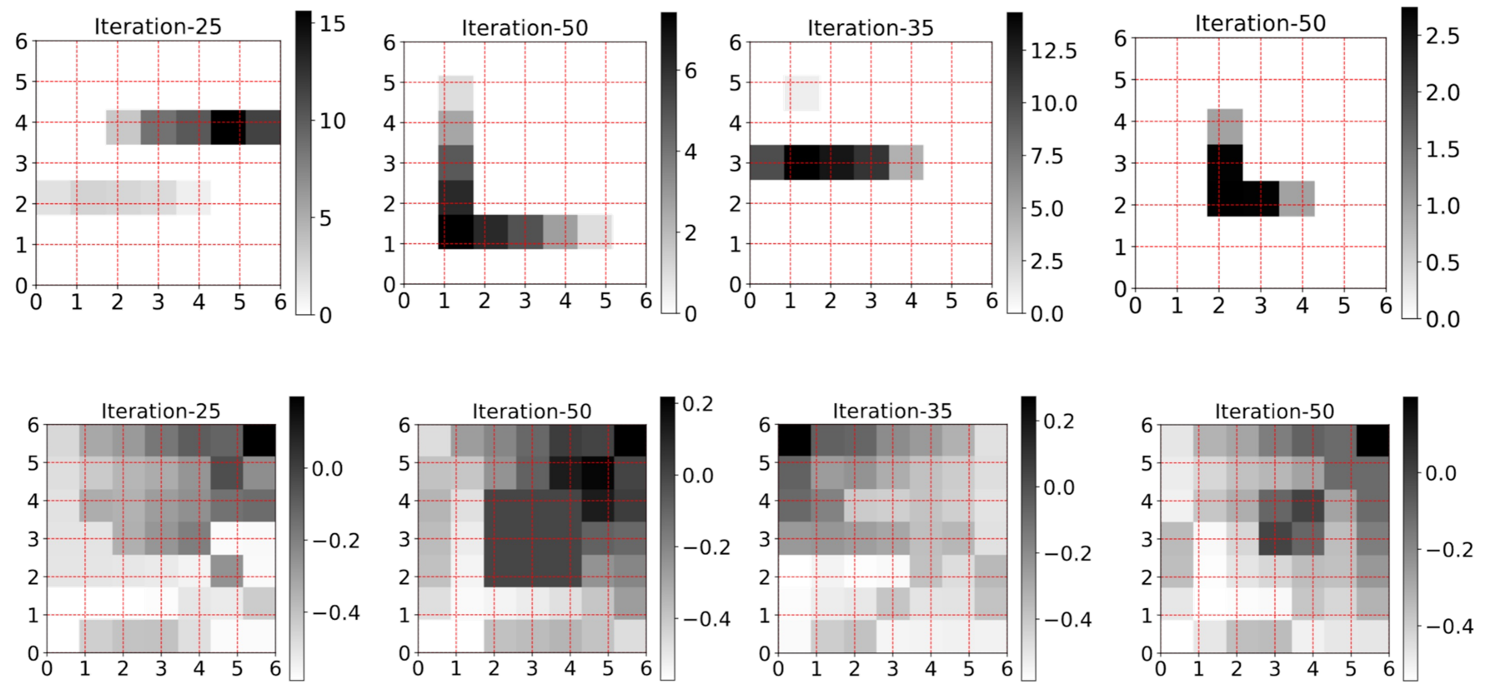
If the two transition laws are close and the recovered cost has a small suboptimality gap in the target environment, then ε-optimality of the recovered cost is guaranteed

Results

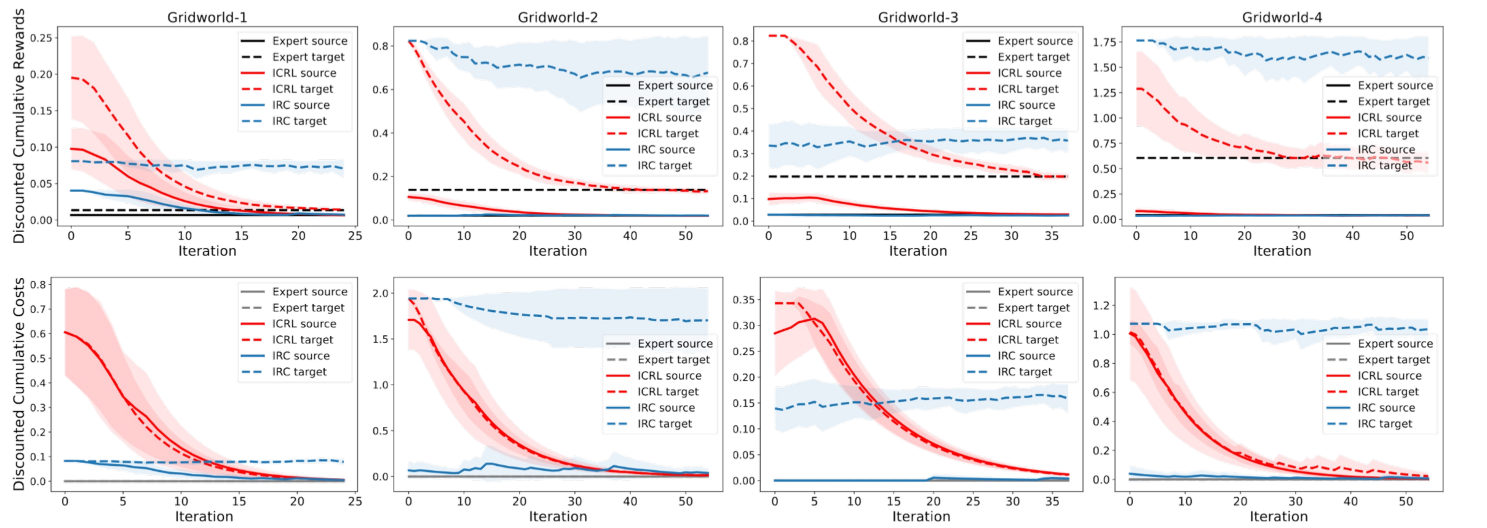
Discrete Envs



Constraint Knowledge (Up: ICRL; Bottom: IRC)



Learning Curves



More details

Group Info