Provably Efficient Exploration in Inverse Constrained Reinforcement Learning

Bo Yue[1], Jian Li[2], Guiliang Liu[1]*

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

[2]Stony Brook University

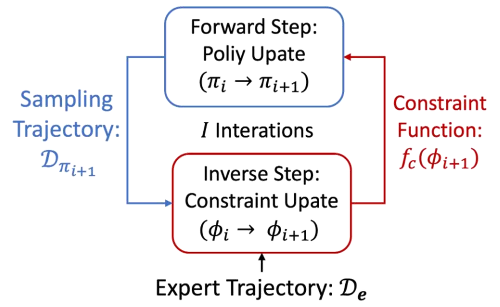
香港中文大學(深圳) The Chinese University of Hong Kong, Shenzhen

Stony Brook University

Abstract

Background: Optimizing objective functions subject to *constraints* is fundamental in many real-world applications. However, ground-truth constraints are often *hard to specify, timevarying and context-dependent*.

Literature: *Inverse Constrained Reinforcement Learning (ICRL)*, recovers feasible constraints via training samples collected from *strategic exploration* in interactive environments.



Challenges: *The efficacy and efficiency of current sampling strategies remain unclear.*

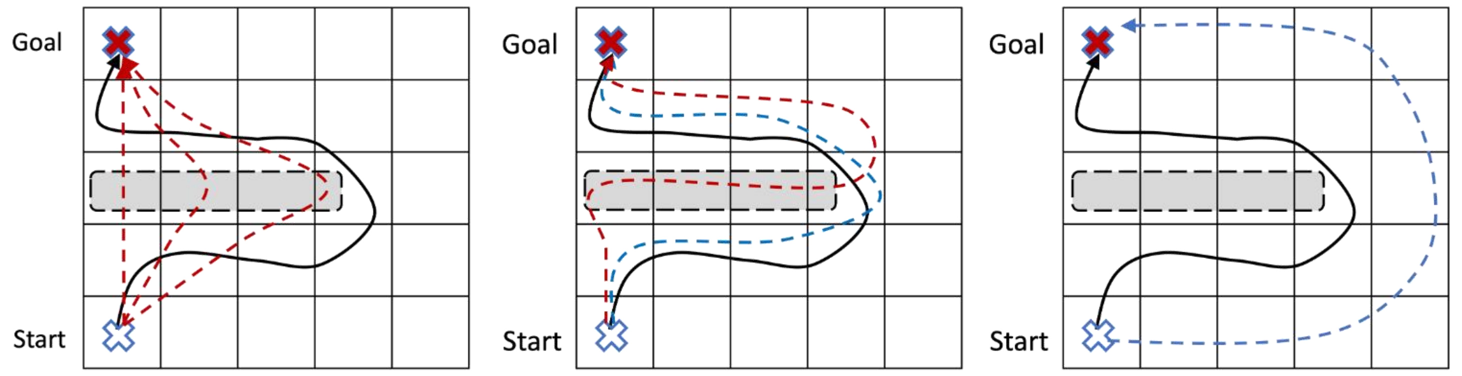
Methodology: Two strategic exploration algorithms: BEAR and PCSE.

Key Takeaways:

- Both BEAR and PCSE are theoretically grounded with tractable sample complexity. (efficiency)
- PCSE outperforms six other baselines in recovering ground-truth constraints. (efficacy)

Recovery of Feasible Constraints

Since the expert policy maximizes rewards within a limited budget, two key insights emerge
1) if a policy achieves higher rewards than the expert policy, **the underlying constraints must be violated**, unsafe state-action pairs by examining these infeasible trajectories (left figure ⬇)
2) if a policy achieves the same or lower rewards than the expert policy, this suggests an absence of notable constraint-violating actions, implying that **the underlying constraints may or may not be violated** (middle & right figure ⬇)



Trajectories of the expert policy (black) and exploratory policies (red and blue) in a Gridworld. The constraint (gray) is not observable.

(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 $A^{r,\pi^E}_{\mathcal{M}\cup c}(s,a) \le 0$, $Q^{c,\pi^E}_{\mathcal{M}\cup c}(s,a) - V^{c,\pi^E}_{\mathcal{M}\cup c}(s) \le 0$.

Feasible cost function

$$c = A^{r,\pi^E}_{\mathcal{M}}\zeta + (E - \gamma P_{\mathcal{T}})V^c$$

Sample Complexity Analysis

Estimation of Expert Policy and Transitions

Use visitation counts: $\widehat{P_{\mathcal{T}}}_k(s'|s,a) = \frac{N_k(s,a,s')}{N_k^+(s,a)}$, $\widehat{\pi}_k^E(a|s) = \frac{N_k^E(s,a)}{N_k^{E+}(s)}$

Error Propagation to Constraint Estimation

$$|c - \widehat{c}|(s,a) \le \frac{2(\chi(s,a)+\chi)}{1+(\chi(s,a)+\chi)/C_{\max}}, \quad \chi = \max\chi(s,a)$$

$$\chi(s,a) = \gamma \left| \underset{①}{(P_{\mathcal{T}} - \widehat{P_{\mathcal{T}}})} V^c \right|(s,a) + \left| A^{r,\pi^E}_{\mathcal{M}} - A^{r,\widehat{\pi}^E}_{\widehat{\mathcal{M}}} \right| \zeta(s,a).$$

$$\left| A^{r,\pi}_{\mathcal{M}} - A^{r,\widehat{\pi}}_{\widehat{\mathcal{M}}} \right| \le \frac{2\gamma}{1-\gamma} \left\| \underset{②}{(\widehat{P}_{\mathcal{T}} - P_{\mathcal{T}})} V^{r,\widehat{\pi}^E}_{\widehat{\mathcal{M}}} \right\| + \frac{\gamma(1+\gamma)}{1-\gamma} \left\| \underset{③}{(\pi - \widehat{\pi})} P_{\mathcal{T}} V^{r,\pi^E}_{\mathcal{M}} \right\|.$$

Strategic Exploration (BEAR and PCSE)

BEAR: guides the exploration policy to visit (s,a) to **minimize the upper bound of constraint estimation errors;**
PCSE: further **restricts exploration over plausibly optimal policies**

Sample complexity of BEAR:

$$n \le \widetilde{\mathcal{O}}\left(\frac{\check{\sigma}^2(2C_{\max} - \varepsilon_K(1-\gamma))^2}{(1-\gamma)^2\varepsilon_K^2 C_{\max}^2}\right)$$

Sample complexity of PCSE:

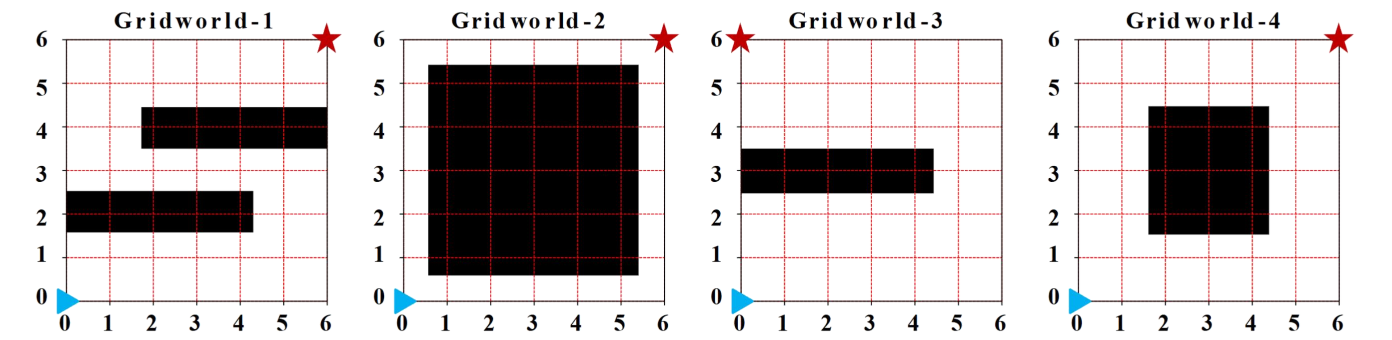
$$n \le \widetilde{\mathcal{O}}\left(\min\left\{\widetilde{\mathcal{O}}\left(\frac{\check{\sigma}^2(2C_{\max} - \varepsilon_K(1-\gamma))^2}{(1-\gamma)^2\varepsilon_K^2 C_{\max}^2}\right), \frac{\sigma^2(6\varepsilon_{K-1}+\epsilon)^2 SA}{\min_{(s,a)}\left(A^{c,*}_{\widehat{\mathcal{M}}\cup\tilde{c}}(s,a)\right)^2 \varepsilon_K^2}\right\}\right)$$

Algorithm 1 BEAR and PCSE for ICRL in an unknown environment

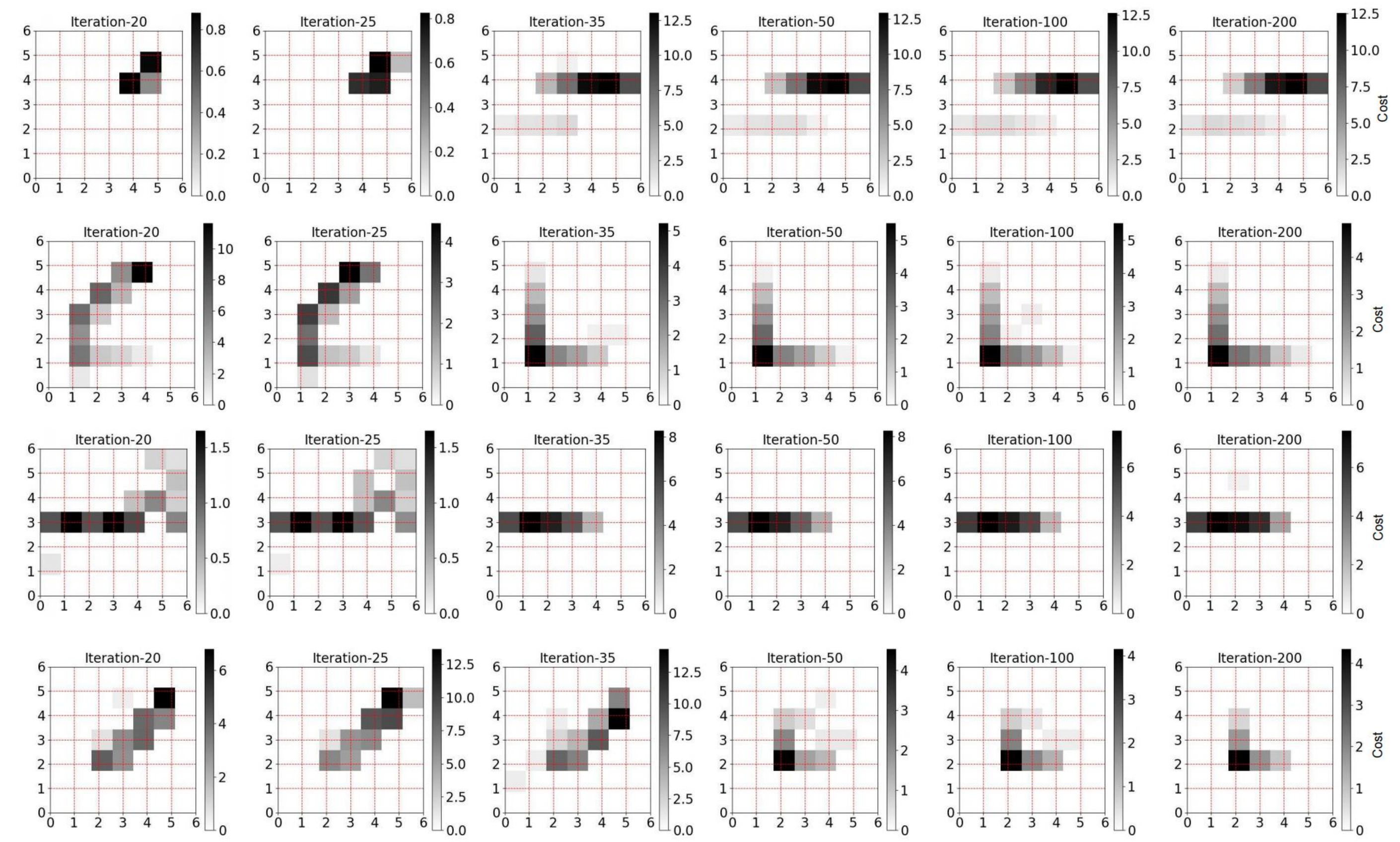
Input: significance $\delta \in (0,1)$, target accuracy ε, maximum number of samples per iteration $n_{\max}$;
Initialize $k \leftarrow 0$, $\varepsilon_0 = \frac{1}{1-\gamma}$;
while $\varepsilon_k > \varepsilon$ **do**
  Solve RL problem defined by $\mathcal{M}^{\mathcal{C}_k}$ to obtain the exploration policy π_k;
  Solve optimization problem in (15) to obtain the exploration policy π_k;
  Explore with π_k for n_e episodes;
  For each episode, collect $n_{\max}$ samples from $\mathcal{S} \times \mathcal{A}$;
  Update accuracy $\varepsilon_{k+1} = \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mathcal{C}_{k+1}(s,a)/(1-\gamma)$;
  Update accuracy $\varepsilon_{k+1} = \|\mu_0^T(I_{\mathcal{S}\times\mathcal{A}} - \gamma P_{\mathcal{T}}\pi)^{-1}\mathcal{C}_k\|_\infty$;
  Update $\widehat{\pi}^E_{k+1}$ and $\widehat{P_{\mathcal{T}}}_{k+1}$ in (7);
  $k \leftarrow k+1$.
end while

Empirical Results

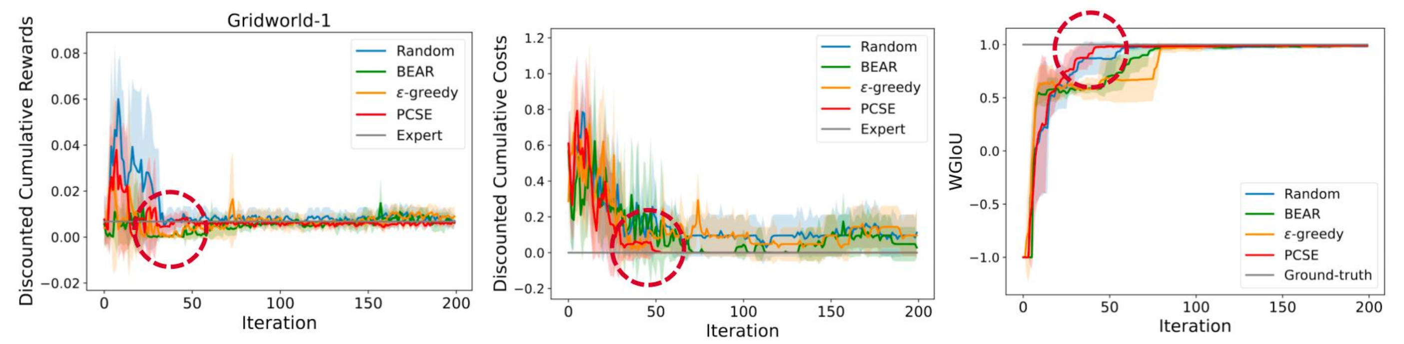
Gridworld Envs



Constraint Recovery Visualization (PCSE)



Learning Curves



More details

Group Info