ROIL: Robust Offline Imitation Learning

Gersi Doko[1], Guang Yang[2], Daniel S. Brown[2], Marek Petrik[1]

[1]University of New Hampshire, [2]University of Utah

Summary

Motivation

Learning from data in a robust offline way is important in many fields, like health care, robotics or finance.

Limitations of existing methods

Reliance on $\hat{u}_e$ leads to covariate shift for off-policy datasets.

Inability to specify reliance on $\hat{u}_e$.

No guarantees of policy convergence to u_e even when every state is visited.

Our contributions

New algorithm for robust offline imitation learning.

Guaranteed convergence to the optimal policy for tabular domains.

Flexibility to define the reliance on $\hat{u}_e$.

Inverse Reinforcement Learning (IRL)

$$\rho(\pi, r) = \lim_{T\to\infty} \mathbb{E}^{\pi,p_0}[\sum_{t=0}^{T} \gamma^t r(\tilde{s}_t, \pi(\tilde{s}_t))]$$

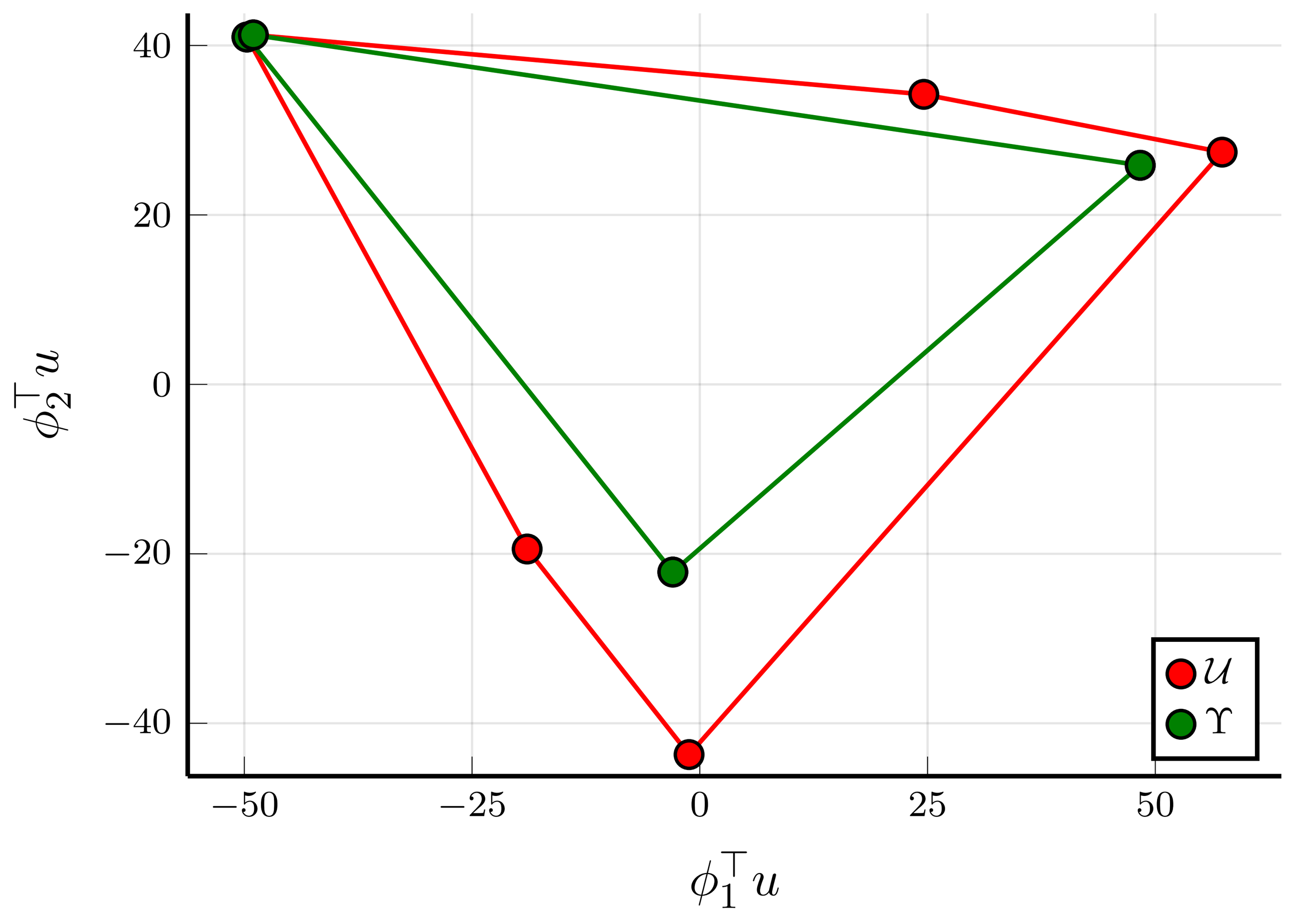
$$\pi^*_{IRL} = \arg\min_{\pi\in\Pi} \max_{r\in\mathcal{R}} \rho(\hat{\pi}_e, r) - \rho(\pi, r)$$

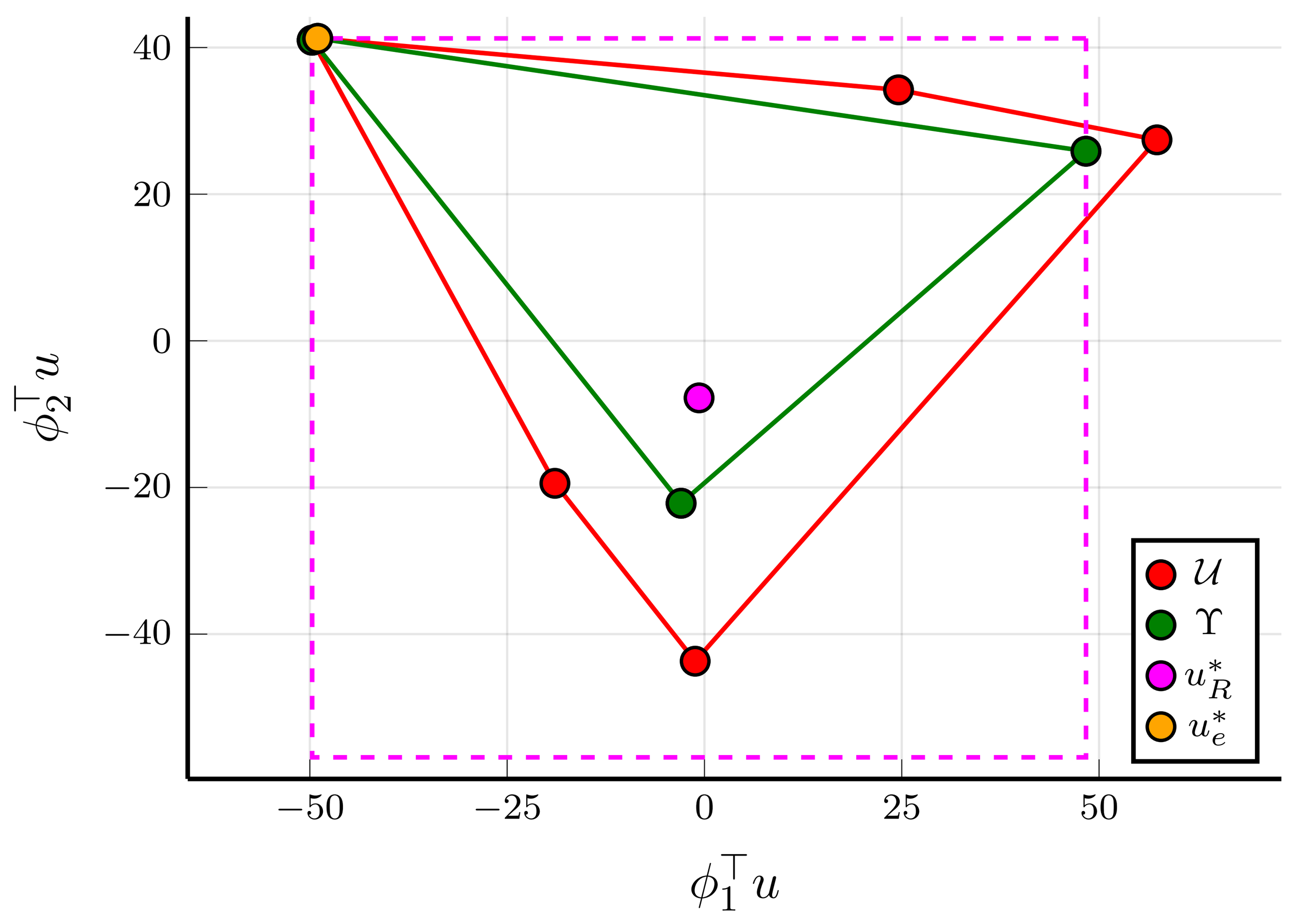
$$\pi^*_{ROIL} = \arg\min_{\pi\in\Pi} \max_{\pi_e\in\Pi} \max_{r\in\mathcal{R}} \rho(\pi_e, r) - \rho(\pi, r)$$

Previous Work

LPAL
MILO
GAIL
BROIL

Visual Representation of ROIL

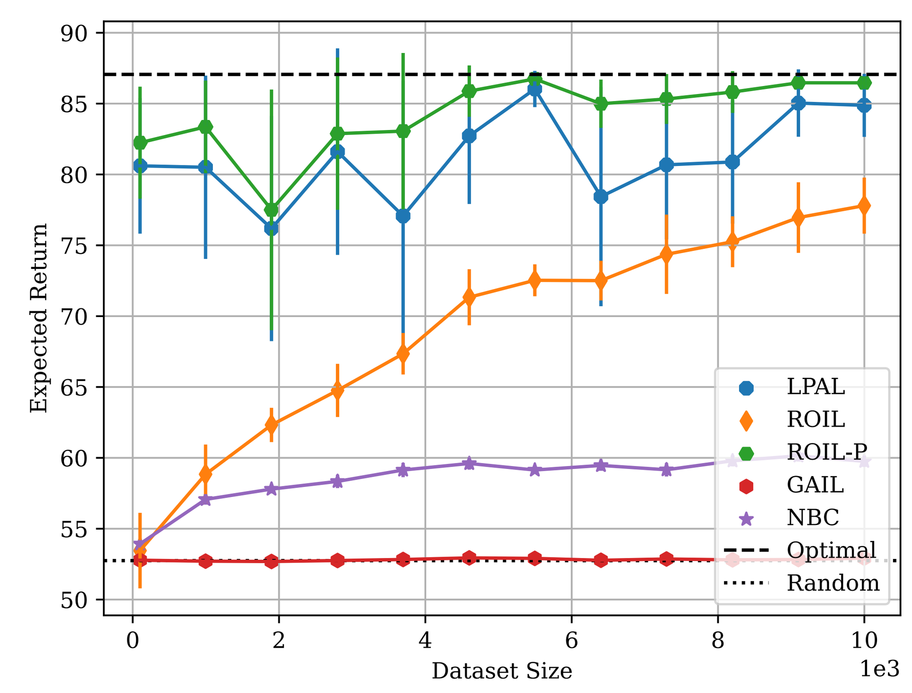




ROIL LP

$$\begin{aligned} \min_{t\in\mathbb{R}, u\in\mathbb{R}^{\mathcal{S}\times\mathcal{A}}} \quad & t \\ \text{s.t.} \quad & t \ge -u^\top\Phi w + \max_{v\in\Upsilon} v^\top\Phi w, \quad \forall\, w \in ext(\mathcal{W}), \\ & u \in \Upsilon, \end{aligned}$$

Gridworld Results



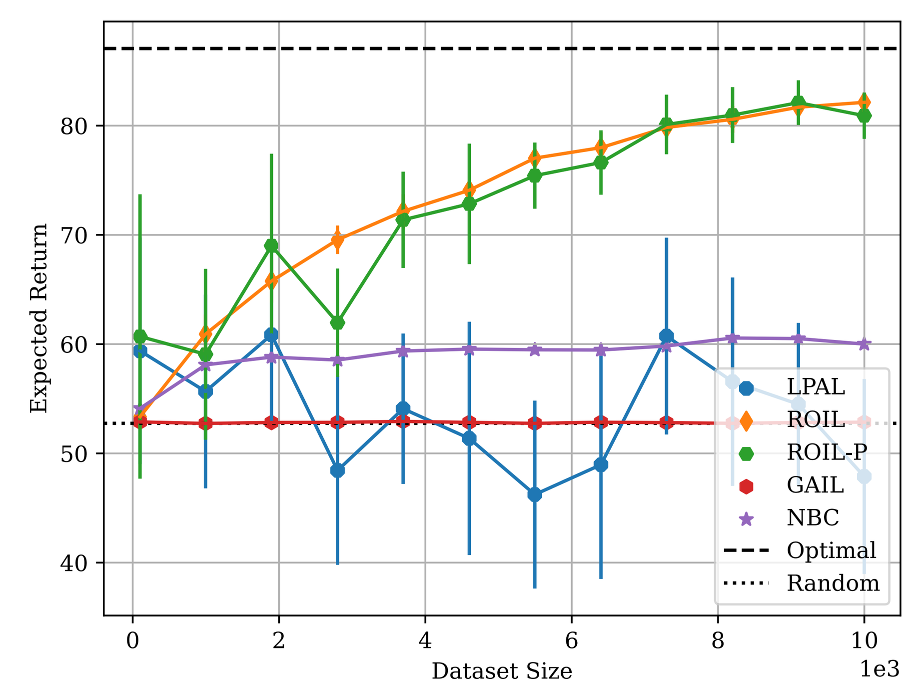


Figure: On-policy and off-policy returns (respectively) for 40x40 Gridworld.

Regret Results

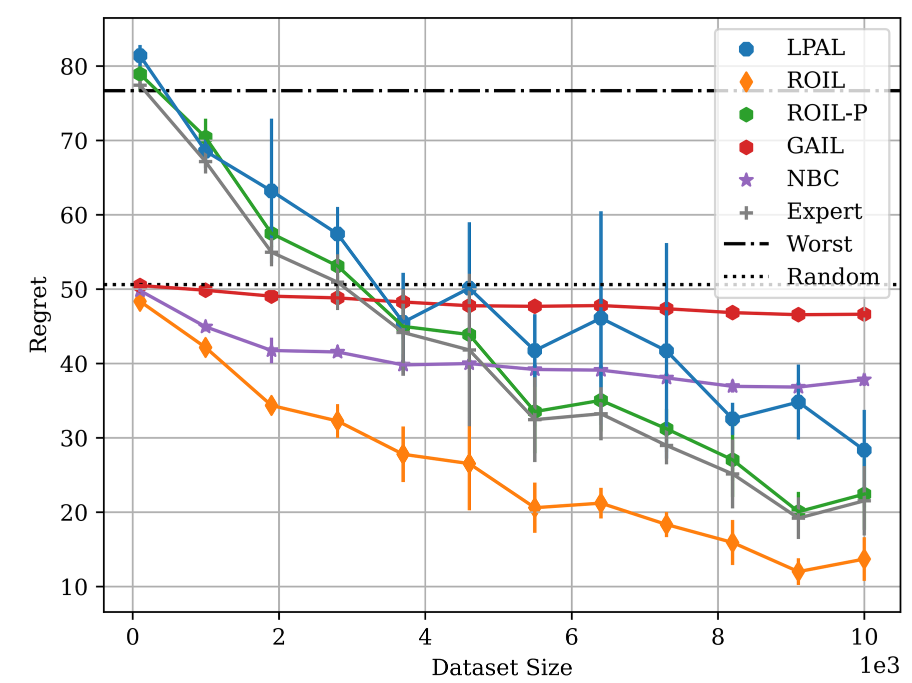


Figure: On-policy regret for 40x40 Gridworld.