

ROIL: Robust Offline Imitation Learning

Gersi Doko¹, Guang Yang², Daniel S. Brown², Marek Petrik¹

¹University of New Hampshire, ²University of Utah

Summary

Motivation

- **Need better offline IRL methods**
- Learning from data in a robust offline way is important in many fields, like health care, robotics or finance
- Existing methods are not robust to covariate shift

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

IRL

- Methods that learn a policy from expert demonstrations and a model of the environment
- **Goal:** Learn a policy that is close to the expert's
- **On-policy:** State visitation frequency is the same as the expert's
- **Off-policy:** State visitation frequency is *different* from the expert's

Not Occupancy Frequency Matching

- Many methods rely on matching the occupancy frequencies of the expert and the learned policy
- LPAL, GAIL, MILO, etc.
- When off-policy, \hat{u}_e is not close to u_e
- ROIL avoids this by not relying on \hat{u}_e

Inverse Reinforcement Learning (IRL)

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

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

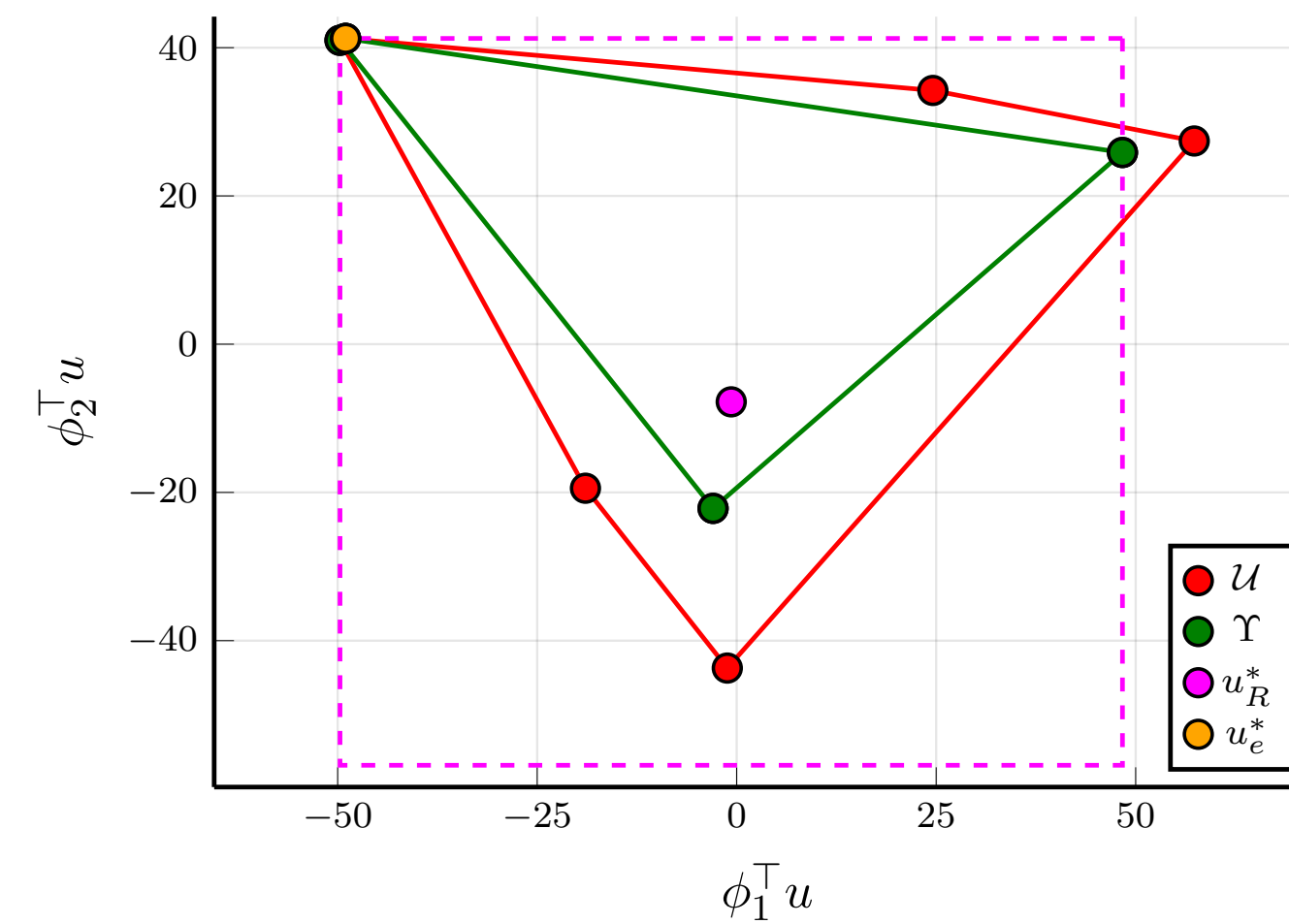
Occupancy Frequencies

$$\mathcal{U} = \{u^\pi \mid \pi \in \Pi\} = \left\{ u \in \mathbb{R}_+^{SA} \mid \sum_{a \in \mathcal{A}} (I - \gamma \cdot P_a^\top) \cdot u(\cdot, a) = p_0 \right\}.$$

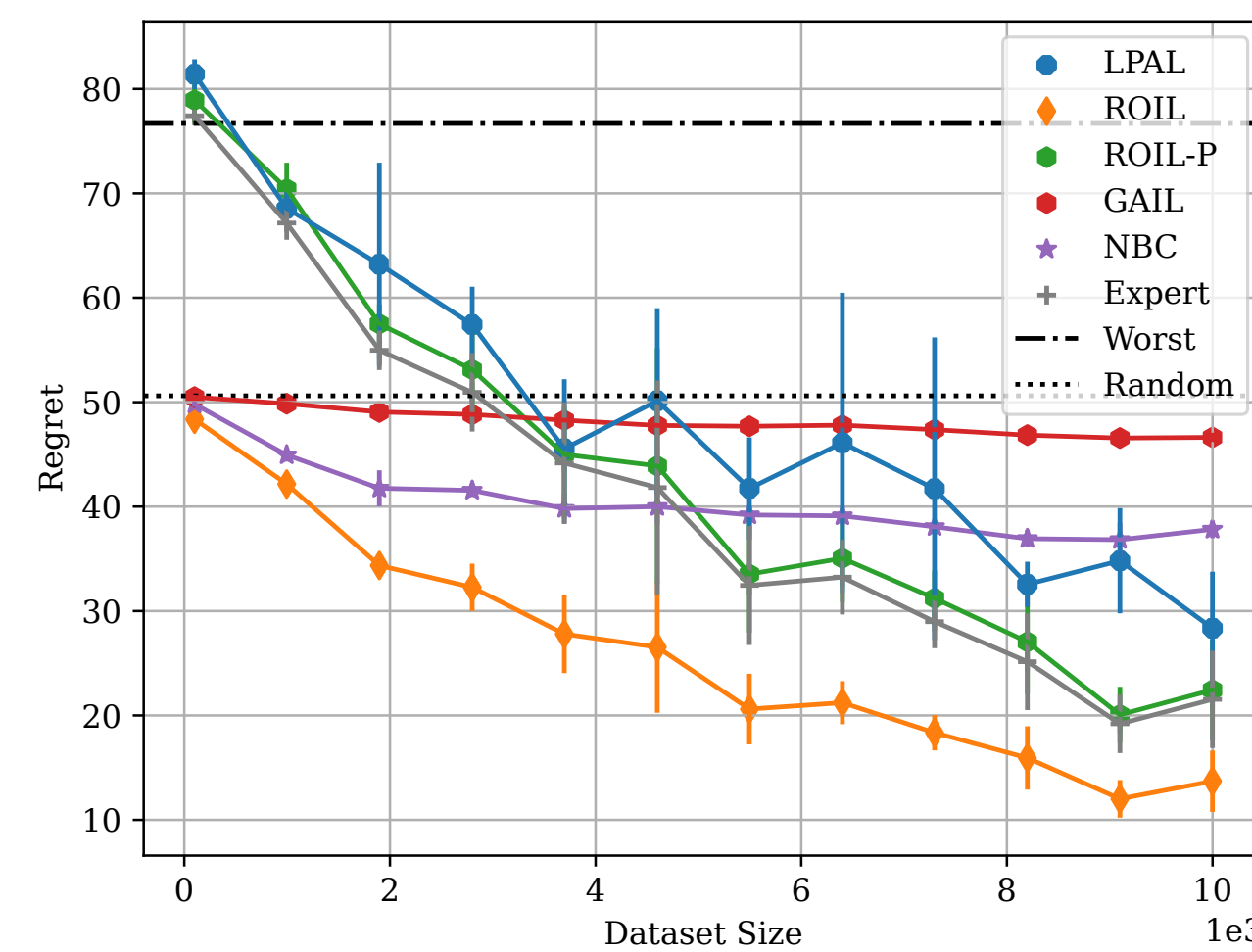
ROIL LP

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

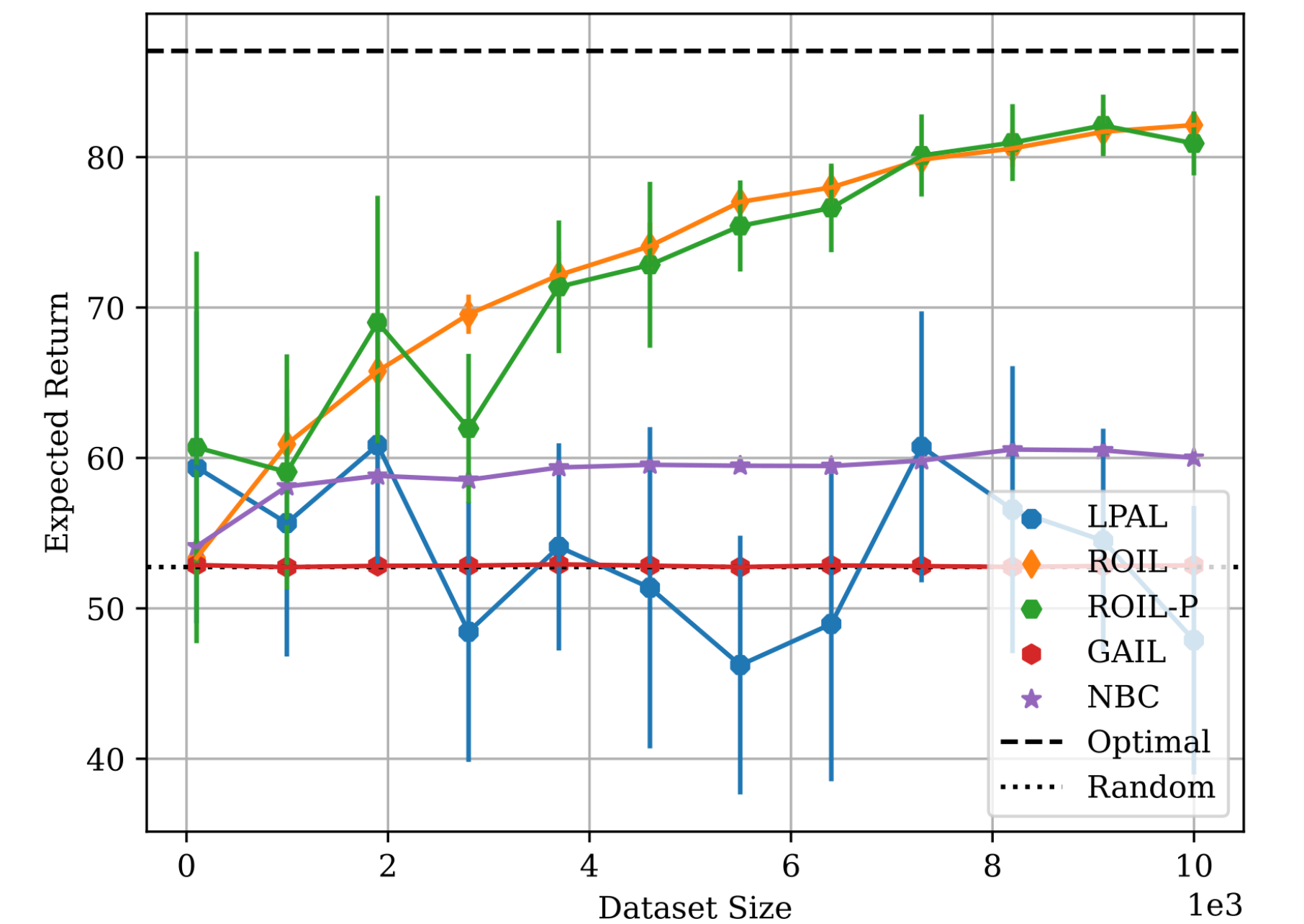
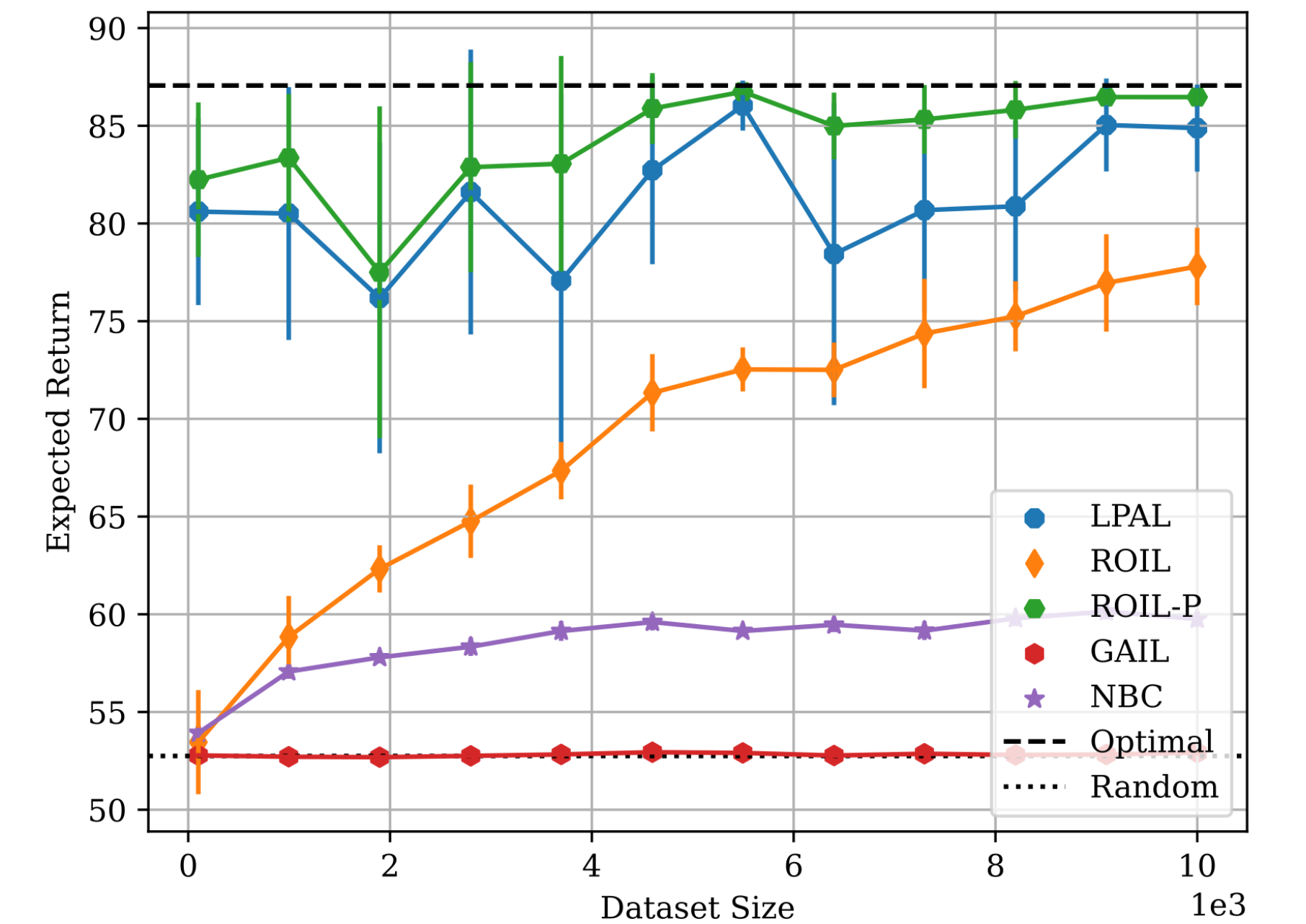
Visual Representation of ROIL



Regret Results



Gridworld Results



Contact

Gersi Doko
Gersi.Doko@unh.edu