ROIL – Robust Offline Imitation Learning

Gersi Doko

Department of Computer Science
University of New Hampshire

Imitation Learning

Imitation Learning

Objective: Learn from expert demonstrations

- Health care: automating and improving ER care
- Robotics: self-driving cars, manufacturing, etc.
- Retail: recommendation systems, customer service

Offline IL: Given fixed dataset of expert demonstrations

- No interaction with the environment

Imitation Learning

- RL requires rewards
- Rewards are hard to specify
- Often have access to expert demonstrations
- *Key Idea*: Supervised learning from expert demonstrations

Imitation Learning

Behavioral Cloning (BC): Supervised learning from expert demonstrations

$$\min_{\theta} \sum_{(s,a)}^{D_e} \text{Loss}\left(\pi_\theta(s) - a\right)$$

Benefits

- Simple
- Natural
- Easy to implement

Imitation Learning Difficulties

$$\min_{\theta} \sum_{(s,a)}^{D_e} \text{Loss}\left(\pi_\theta(s) - a\right)$$

Central Issues

- Sample inefficient
- Expert demonstrations may not be optimal
- Sensitive to dataset collection

Inverse Reinforcement Learning

Objective: Learn from expert demonstrations

- Leverage model dynamics to reduce sample complexity
- Aims to match experts state-action distribution
- Known model dynamics allow for generalization

Key Idea: Model dynamics allow for generalization

Inverse Reinforcement Learning

Objective: Learn from expert demonstrations

Our Focus: Demonstrations may not be a set of trajectories

- On-Policy: Demonstrations are generated by the expert's policy
- Off-Policy: Demonstrations are generated by a different behavior policy

Off-Policy Inverse Reinforcement Learning

Off-Policy: Demonstrations are generated by a different behavior policy

When would off-policy demonstrations happen?

- Selecting exemplar states
- Non-stationary expert
- Non-stationary environment
- Different inital state dist. p_0
- Different discount factor γ
- ...

Markov Decision Process

Model (tabular in this talk)

States $\mathcal{S}$: $s_1, s_2, s_3, \ldots$

Actions $\mathcal{A}$: $a_1, a_2, \ldots$

Transition probabilities $\mathcal{P} \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}\times\mathcal{S}}$

Initial state distribution $p_0 \in \Delta^{\mathcal{S}}$

Discount factor $\gamma \in \mathbb{R}$

Features $\Phi \in \mathbb{R}^{\mathcal{S}\mathcal{A}\times k}$

~~Rewards $r \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$~~

Solution: Policy $\pi\colon \mathcal{S} \to \Delta^{\mathcal{A}}$

Return: Discounted expected infinite horizon return (expectation over trajectories):

$$\tilde{\rho}(\pi) = \lim_{T\to\infty} \sum_{t=0}^{T} \gamma^t r(\tilde{s}_t^\pi, \tilde{a}_t^\pi)$$

Random variables: $\tilde{\rho}, \tilde{s}, \tilde{a}, \tilde{x}, \ldots$ adorned with tilde

Occupancy Frequencies

$$\mathcal{U} = \left\{ u \in \mathbb{R}_{+}^{SA} \mid u_\pi(s,a) \propto \sum_{t=0}^{\infty} \mathbb{P}(\tilde{s}_t = s, \tilde{a}_t = a \mid \tilde{s}_{t+1} \sim \mathcal{P}(s_t, \pi(s_t))) \right\}$$

$u_\pi(s,a)$ is the long-run probability of agent π being in state s *and* taking action a.

Consistent Occupancy Frequencies

We are given a dataset $D_e = (s_t, \pi_e(s_t))_{t=1}^{T}$.

Definition: The set of occupancy frequencies consistent with D_e is

$$\Upsilon = \left\{ u \in \mathcal{U} \mid u(s,a) = 0 \iff (s,a) \notin D_e \text{ and } (s,a') \in D_e \right\},$$

Inverse Reinforcement Learning

Objective: Learn from expert data D_e

$$\min_{\pi \in \Pi} \max_{r \in \mathcal{R}} \rho(\hat{\pi}_{D_e}, r) - \rho(\pi, r)$$

Benefits

- Able to generalize to unseen states
- Can learn from suboptimal demonstrations

Central Issue

- Estimating the expert's policy $\hat{\pi}_{D_e}$

Inverse Reinforcement Learning

Objective: Learn from expert data D_e

$$\min_{\pi\in\Pi}\max_{r\in\mathcal{R}} \rho(\hat{\pi}_{D_e}, r) - \rho(\pi, r) \tag{1}$$

Not convex!

$$\min_{u\in\mathcal{U}}\max_{r\in\mathcal{R}} \rho(\hat{u}_{D_e}, r) - \rho(u, r) \tag{2}$$

$$\hat{u}_{D_e}(s,a) = \sum_{(t,s',a')}^{D_e} \gamma^t \mathbb{1}\left\{s = s' \wedge\ a = a'\right\}$$

The Central Issue

$$\hat{\pi}_{D_e}(s,a) \approx \pi_e(s,a)$$

$$\not\Rightarrow$$

$$\hat{u}_{D_e}(s,a) \approx u_{\pi_e}(s,a)$$

The Central Issue

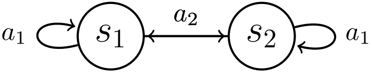
$$\hat{\pi}_{D_e}(s,a) \approx \pi_e(s,a) = \mathbb{P}(\tilde{a} = a \mid \tilde{s} = s)$$

$$\not\Rightarrow$$

$$\hat{u}_{D_e}(s,a) \approx u_{\pi_e}(s,a) = \mathbb{P}(\tilde{s} = s \text{ and } \tilde{a} = a)$$

$$\hat{\pi}_{D_e}(s,a) \cdot \mathbb{P}(\tilde{s} = s) = \hat{u}_{D_e}(s,a)$$

Off-Policy IRL



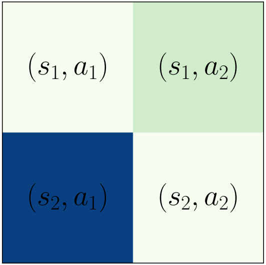
$$D_e = \{(s_1, a_2), (s_2, a_1), (s_2, a_1), \ldots\}$$

$$D_e = \{(s_1, a_2), (s_2, a_1)\}$$

Off-Policy IRL

On-Policy

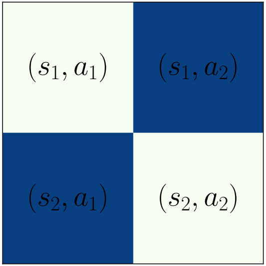
True Expert u_e



LPAL Return $= 86/87$

Off-Policy

Estimated Expert $\hat{u}_{D_e}$

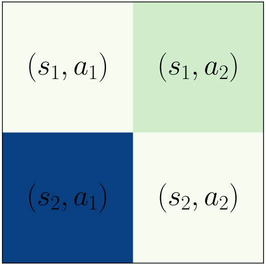


LPAL Return $= 38/87$

Off-Policy IRL

On-Policy

True Expert u_e

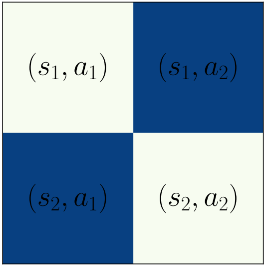


LPAL Return $= 86/87$

ROIL Return $= 79/87$

Off-Policy

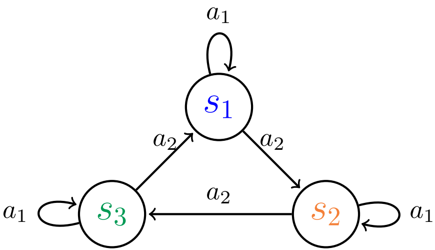
Estimated Expert $\hat{u}_{D_e}$



LPAL Return $= 38/87$

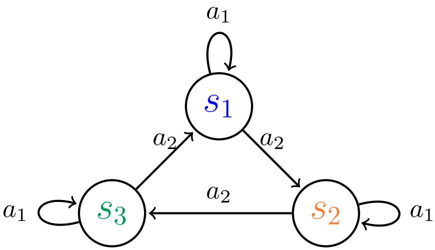
ROIL Return $= 82/87$

Full State Coverage



$$D_e = \{(s_1, a_2), (s_2, a_2), (s_3, a_1)\}$$

Full State Coverage



$$D_e = \{(s_1, a_2), (s_2, a_2), (s_3, a_1)\}$$

ROIL Return $= 100\%$
LPAL Return $= 50\%$
GAIL Return $= 50\%$

This Talk

Objective: Don't estimate the expert's policy

$$\min_{\pi \in \Pi} \max_{r \in \mathcal{R}} \rho(\hat{\pi}_{D_e}, r) - \rho(\pi, r)$$

Key Idea: Minimize worst case regret

$$\min_{\pi \in \Pi} \max_{\pi_e \in \Pi_{D_e}} \max_{r \in \mathcal{R}} \rho(\pi_e, r) - \rho(\pi, r)$$

ROIL: Robust Offline Imitation Learning

$$\min_{\pi \in \Pi} \max_{\pi_e \in \Pi_{D_e}} \max_{r \in \mathcal{R}} \rho(\pi_e, r) - \rho(\pi, r)$$

ROIL: Robust Offline Imitation Learning

$$\min_{\pi \in \Pi} \max_{\pi_e \in \Pi_{D_e}} \max_{r \in \mathcal{R}} \rho(\pi_e, r) - \rho(\pi, r)$$

$$\min_{u \in \mathcal{U}} \max_{u_e \in \Upsilon} \max_{r \in \mathcal{R}} \rho(u_e, r) - \rho(u, r)$$

ROIL: Robust Offline Imitation Learning

$$\min_{u \in \mathcal{U}} \max_{u_e \in \Upsilon} \max_{r \in \mathcal{R}} \rho(u_e, r) - \rho(u, r)$$

$$\begin{aligned}
\underset{t \in \mathbb{R}, u \in \mathbb{R}^{\mathcal{S}\mathcal{A}}}{\text{minimize}} \quad & t \\
\text{subject to} \quad & t \geq \max_{u_e \in \Upsilon} {u_e}^{\mathsf{T}} r - u^{\mathsf{T}} r, \quad \forall\, r \in ext(\mathcal{R}), \\
& u \in \Upsilon
\end{aligned}$$

- $ext(\mathcal{R})$ is the set of extreme points of $\mathcal{R}$
- u is the occupancy frequency of our policy
- t is the worst case regret

ROIL-P

- **Key Strength**: ROIL does not estimate the expert's policy $\hat{\pi}_e$
- **Problem**: In on-policy domains, estimates of $\hat{\pi}_e$ are close to the true expert
- **Solution**: ROIL-P, a variant of ROIL that estimates $\hat{\pi}_e$, and prunes the set of reward functions

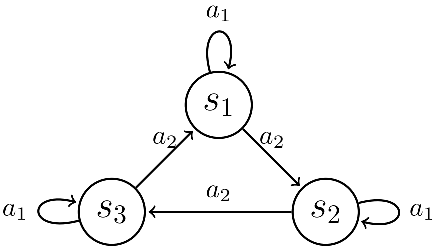
ROIL-P

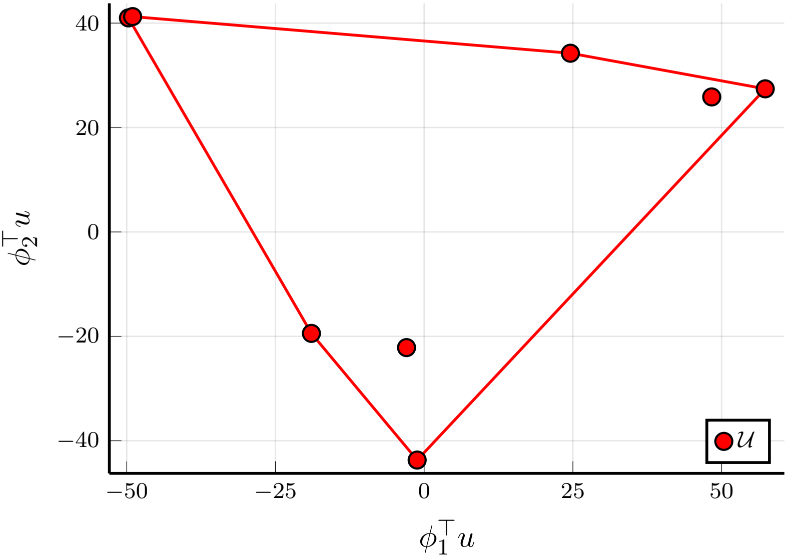
Solution: ROIL-P, a variant of ROIL that estimates the expert's occupancy frequency, and prunes the set of reward functions

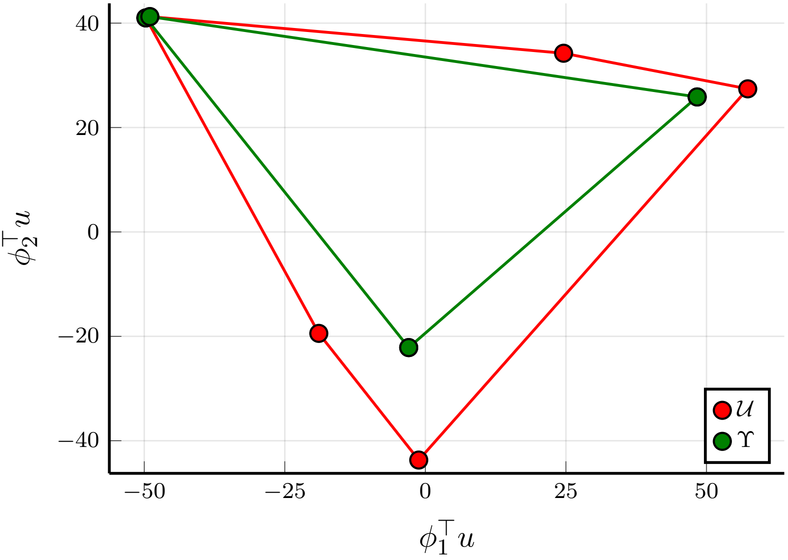
$$\begin{aligned} \underset{t\in\mathbb{R},u\in\mathbb{R}^{\mathcal{S}\times\mathcal{A}}}{\text{minimize}} \quad & t \\ \text{subject to} \quad & t \geq \max_{u_e\in\Upsilon} u_e^{\mathsf{T}} r - u^{\mathsf{T}} r, \quad \forall\, r \in ext(\{r\in\mathcal{R} \mid \hat{u}_e^{\mathsf{T}} r \geq 0\}), \\ & u \in \Upsilon \end{aligned}$$

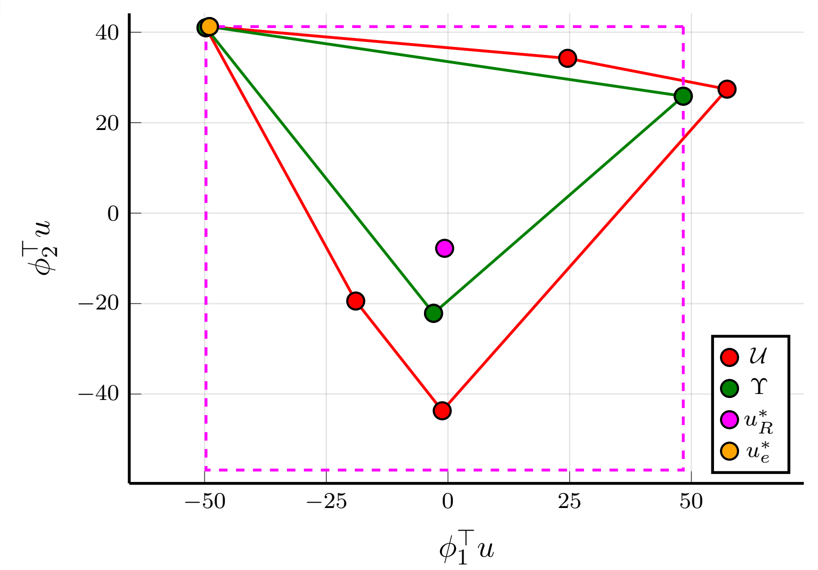
Assume the expert's policy is good

Example



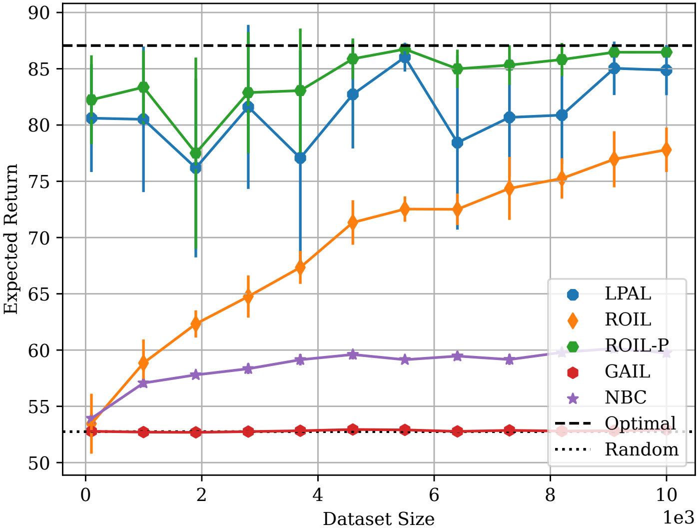

40
20
0
−20
−40
−50
−25
0
25
50
$\phi_2^\top u$
$\phi_1^\top u$
$\mathcal{U}$


40
20
0
−20
−40
−50
−25
0
25
50
$\phi_2^\top u$
$\phi_1^\top u$
$\mathcal{U}$
Υ

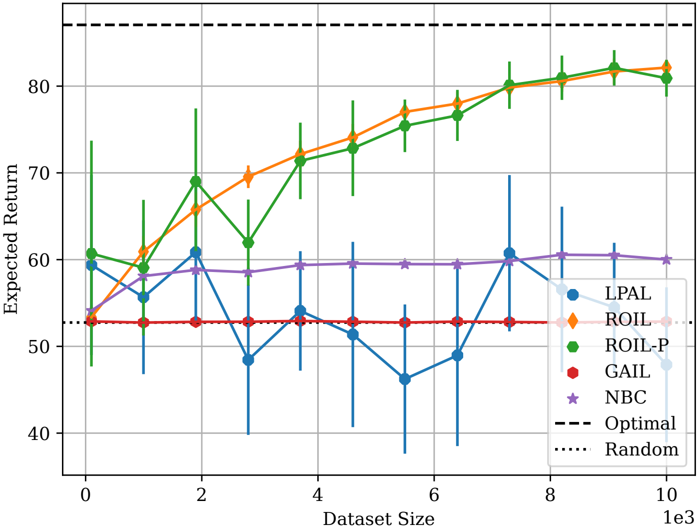


40
20
0
−20
−40
−50
−25
0
25
50
$\phi_2^\top u$
$\phi_1^\top u$
$\mathcal{U}$
Υ
u_R^*
u_e^*

40x40 Gridworld - On-Policy

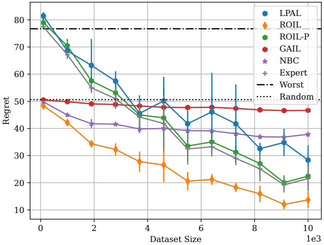


40x40 Gridworld - Off Policy

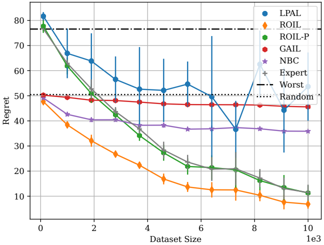


Regret Comparison

$$\text{Regret}(\pi) = \max_{\pi_e \in \Pi_{D_e}} \max_{r \in \mathcal{R}} \rho(\pi_e, r) - \rho(\pi, r)$$



(a) On-Policy



(b) Off-Policy

Conclusion

- Need offline IRL methods that are robust to off-policy data
- Existing methods fail to learn a robust policy
- ROIL is a principled approach to solving the robust offline IRL problem