

CS-7648-A Interactive Robo Learn

Spring 2025

Uncertainty-Aware IRL from Suboptimal Demonstrations

Final Presentation

Presenters: Mengying Lin, Woo Chul Shin

Instructor: Matthew Gombolay

Introduction

- **Background**

- Most IRL methods assume demonstrations are near-optimal
- In real-world settings, demos are noisy and suboptimal

- **Challenge**

- Existing methods like SSRR use fixed ranking margins, assuming clear separation between demonstrations

- **Our approach**

- Model rewards as a Gaussian distribution
 - When behaviors overlap, the model should reflect high variance (low confidence), and when differences are clear, it should show low variance (high confidence)
- Allows flexible, data-driven confidence in ranking decisions

- **Key Idea**

- Introduce soft margins based on learned uncertainty

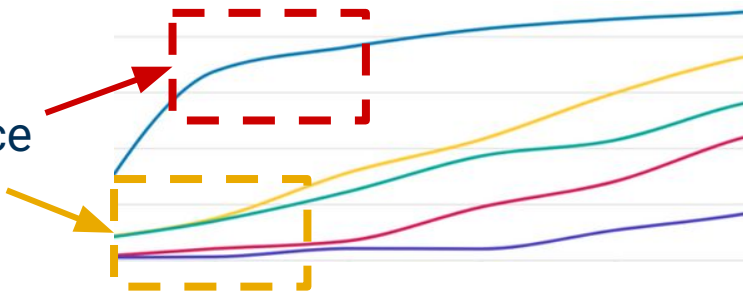
Methods

Reward Modelling

$$R(s, a) \sim \mathcal{N}(\mu_{\theta}(s, a), \sigma_{\phi}^2(s, a))$$

Intuitions behind: When demonstrations provide

- **strong, consistent ranking signals:** low variance
- **ambiguous, overlapping signals:** high variance



Training Objective

When demonstrator i is better than j (τ : trajectory),

$$\mathcal{L}_{\text{rank}} = - \sum_{(i,j) \in \mathcal{D}} \log P(\tau_i \succ \tau_j).$$

Methods

Probability of sampling a trajectory

Shepard-Luce rule in MaxEnt IRL:

$$P(\tau_i) = \frac{\exp\left(\sum_{(s,a) \in \tau_i} \hat{R}(s,a)\right)}{\sum_{k=1}^n \exp\left(\sum_{(s,a) \in \tau_k} \hat{R}(s,a)\right)} \quad (3)$$

In probability:

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

$$\mathbb{E}[\exp(X)] = \exp\left(\mu + \frac{1}{2}\sigma^2\right) \quad (4)$$

Substitute (4) into (3):

$$P(\tau_i) = \frac{\exp\left(\sum_{(s,a) \in \tau_i} \mu_\theta(s,a) + \frac{1}{2} \sum_{(s,a) \in \tau_i} \sigma_\phi^2(s,a)\right)}{\sum_{k=1}^n \exp\left(\sum_{(s,a) \in \tau_k} \mu_\theta(s,a) + \frac{1}{2} \sum_{(s,a) \in \tau_k} \sigma_\phi^2(s,a)\right)}$$

Experiment: Data Collection & Setup

- **Env:** Hopper-v3, Half Cheetah-v3, Ant-v3
- **Data Collection (T-REX-style):**
 - Train a TRPO agent in simulation
 - Periodically checkpoint the policy at different training steps
 - This yields a sequence of suboptimal trajectories that gradually improve over time
- **Checkpoints Used:**
 - Hopper-v3: 9 checkpoints
 - Ant-v3: 12 checkpoints
 - Half Cheetah-v3: 12 checkpoints
- **Train Policies with Learned Rewards:**
 - Actor Critic Network
 - Use reward mean to replace reward signals from environments

Experiment: Ranking accuracy

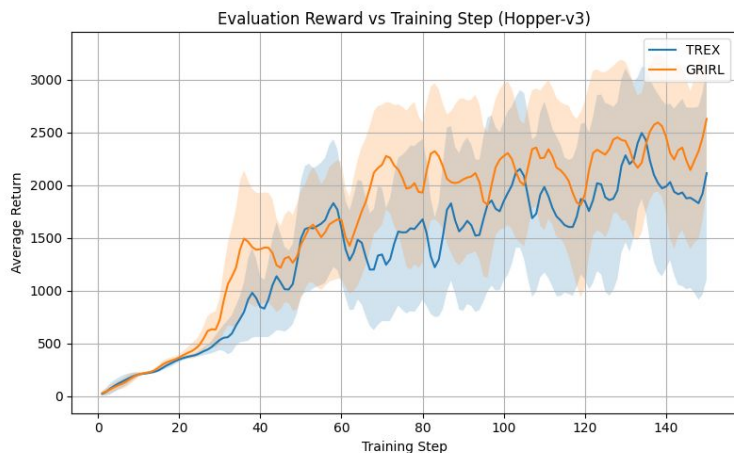
- Both methods achieve comparable ranking accuracy across environments
- GR-IRL consistently yielding slightly better results than T-REX

TABLE I: Trajectory ranking accuracy (%) across different environments. GR-IRL consistently outperforms T-REX.

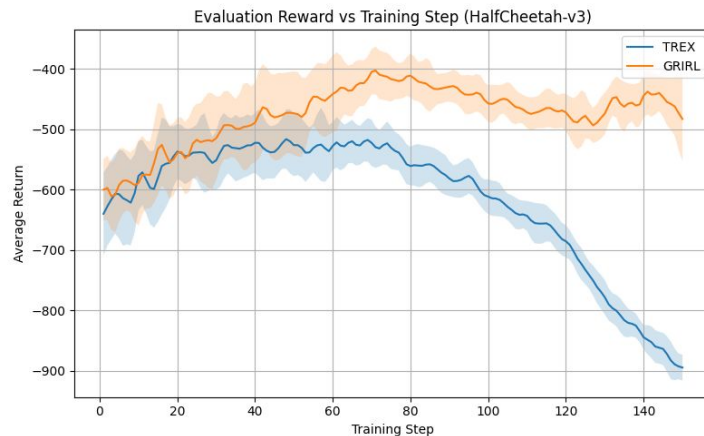
Environment	T-REX	GR-IRL
Hopper-v3	99.78	99.88
Ant-v3	97.35	98.42
HalfCheetah-v3	98.42	98.67

Experiment: Train policy with learned rewards

- Hopper-v3
 - GRIRL consistently outperforms TREX in terms of average return across training steps.



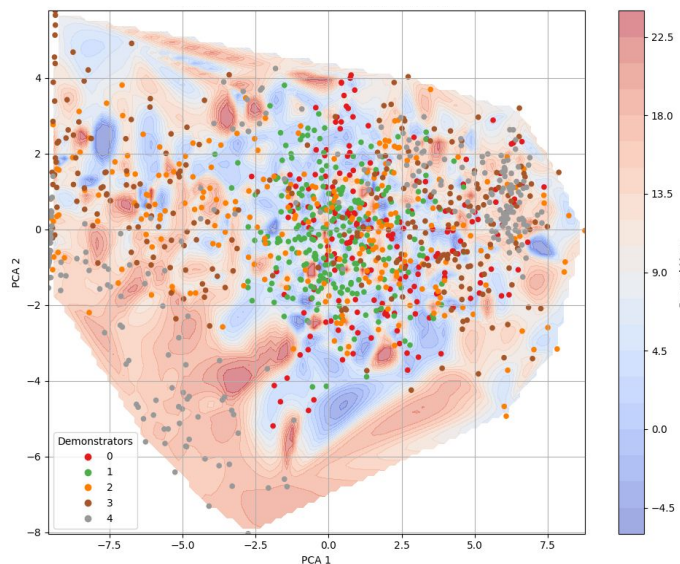
- HalfCheetah-v3
 - GRIRL achieves better stability and higher average return than TREX, which declines over time.



Experiment: PCA on state-action pairs

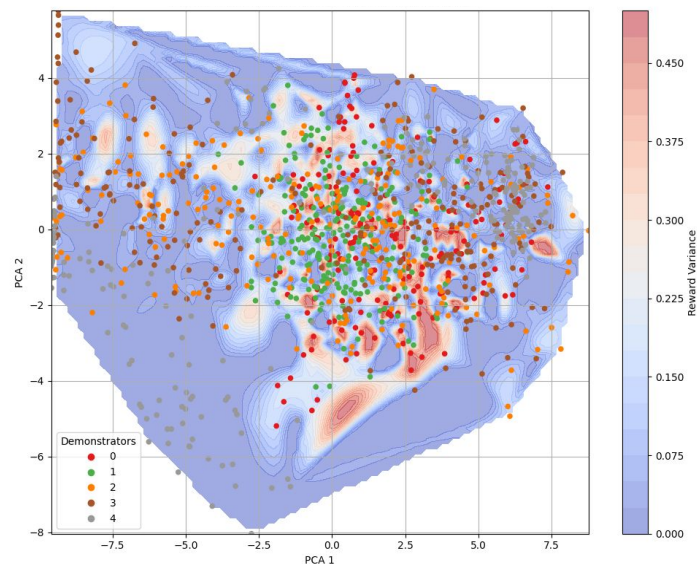
Reward Mean Field

- Higher-skilled demos (e.g., gray = Demo 4) cluster in high-reward regions
- Lower-skilled demos (red, green = Demo 0/1) lie in lower-reward areas



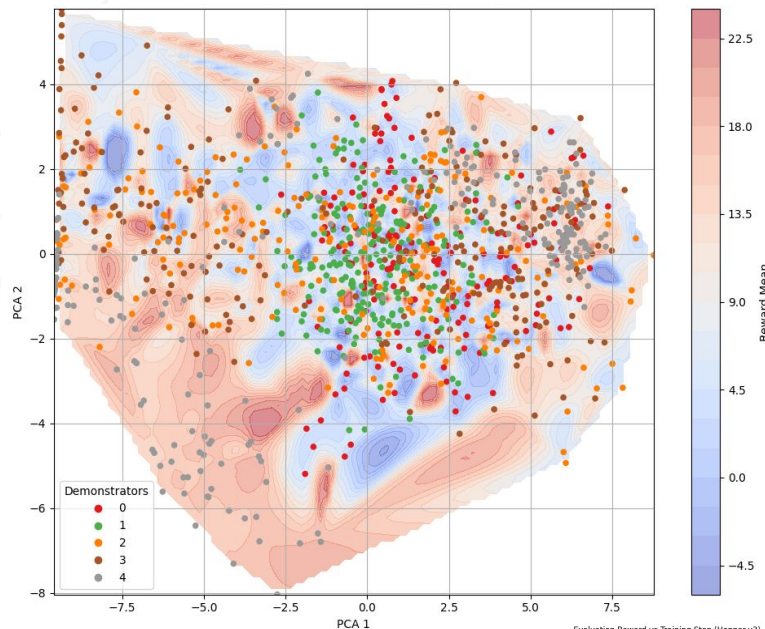
Reward Variance Field

- High variance in overlapping regions → model is uncertain
- Low variance where one demonstrator dominates → model is confident

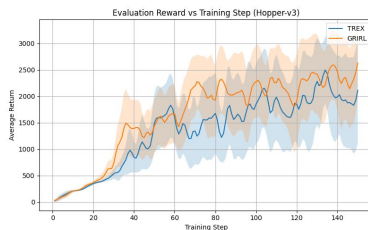


Experiment: PCA on state-action pairs

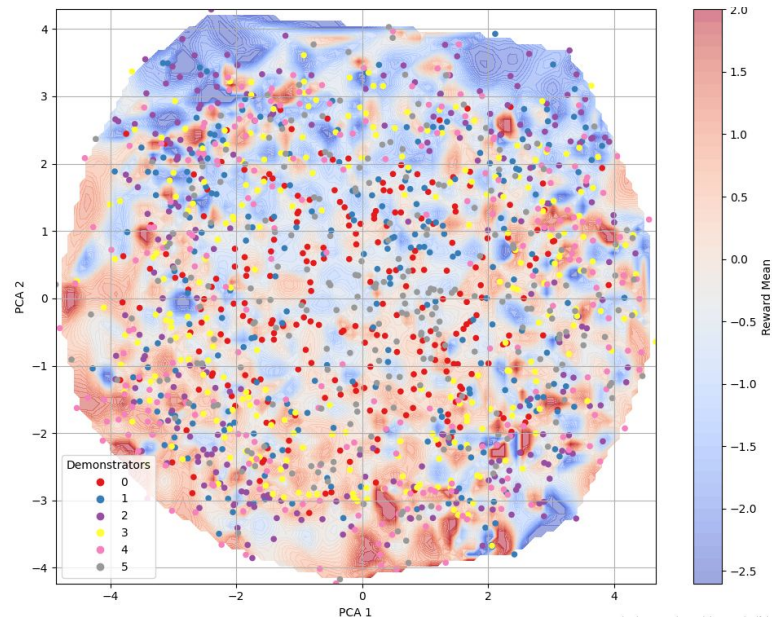
Hopper-v3



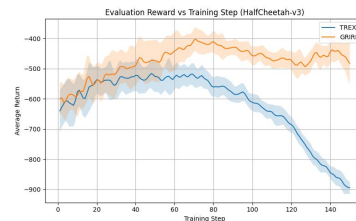
More separable
→ Better results



HalfCheetah-v3

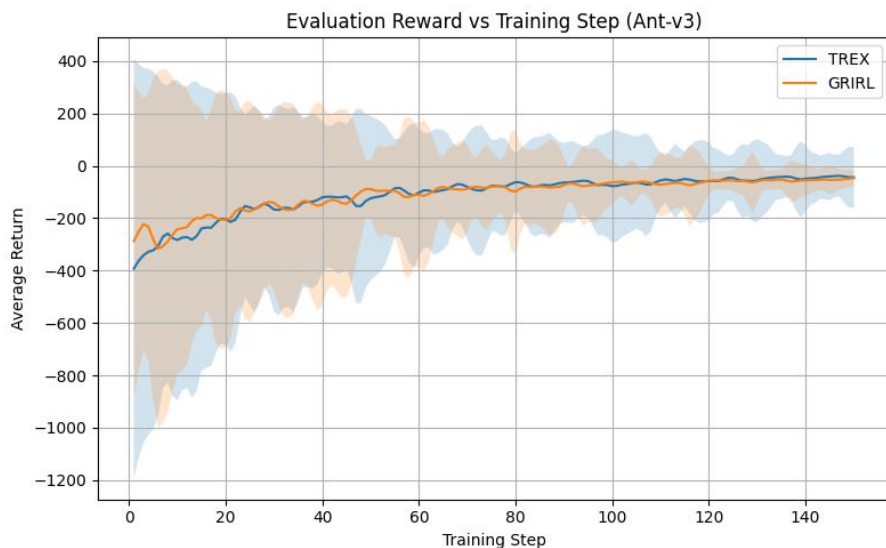


Almost non-separable
→ Worse results



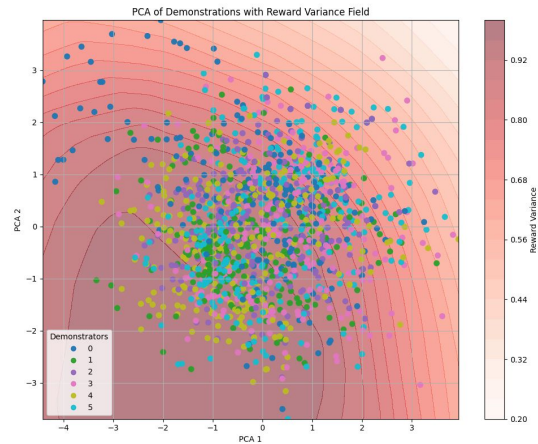
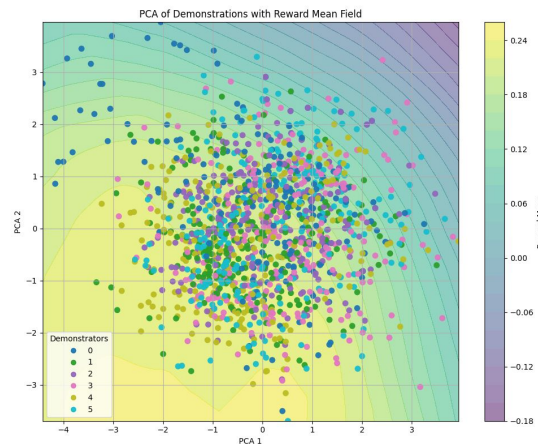
Experiment: PCA on state-action pairs

- Ant-v3



Non-separable demo

→ Worse results



Limitation & Conclusion

- Conclusion

- GR-IRL is a Gaussian reward modeling framework that improves learning from noisy demonstrations by explicitly modeling uncertainty.
- It achieves better generalization and policy performance, as shown across MuJoCo benchmarks.

- Limitation

- The model does not fully exploit uncertainty theory like Bayesian uncertainty, leaving room for more robust approaches.
- The model requires trajectory-level rankings, which can be challenging to acquire in real-world scenarios.

Thanks for your attention!
Q & A