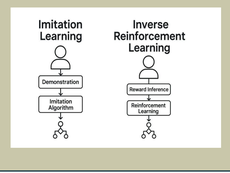
From Demonstrations to Adaptation: Assessing Imitation Learning Robustness and Learned Reward Transferability

Nathan Van Utrecht

Department of Mechanical Engineering, Iowa State University

Introduction

- **What is Reinforcement Learning (RL)?:** AI learns tasks by trial and error, getting rewards for good actions to maximize points.
- **The Reward Design Problem:** Designing perfect rewards is hard for complex real-world tasks (think driving, surgery, etc.).
- **Learning from Examples:** Instead of a reward function, we can show the AI what an expert does. **Imitation Learning (IL)** copies expert actions but often fails if the world changes. **Inverse Reinforcement Learning (IRL)** finds the expert's reward function instead of just copying.



Methods

Algorithms
This study used **Behavioral Cloning (BC)** which is a baseline IL method, **Generative Adversarial Imitation Learning (GAIL)** which is an advanced IL method, and **Adversarial Inverse Reinforcement Learning (AIRL)** which is an advanced IRL method.

Testing Environments
Used original environments and created **modified versions** (changed physics, shifted goals) of the environments shown in Figure 1. The training pipeline is shown in Figure 2.

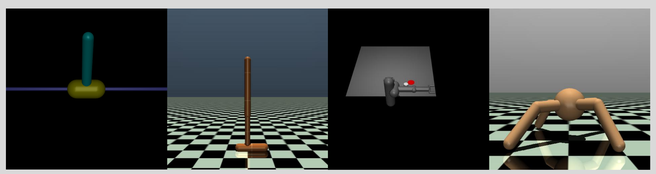


Figure 1: Environment names from left to right in ascending difficulty: Hopper, Pusher, Ant, Inverted Pendulum

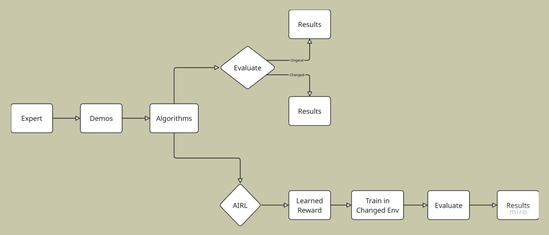


Figure 2: Experiment training/evaluation pipeline

Results

Training Data
The training data was collected by training a Soft Actor-Critic (SAC) agent until it reached expert level performance. One million timesteps of data were collected with deterministic actions.

Evaluation Methodology
All results were averaged over 100 randomly generated seeds. For all environments, the larger the number, the better the performance. The top algorithm is in bold for each environment, and any algorithm within 5% of the top one is also bolded.

Table 1: Algorithm Performance on Standard Environments (Mean ± Std Return)

	Ant	Hopper	Inverted Pendulum	Pusher
AIRL	5233.60 ± 989.55	3410.27 ± 18.06	**1000.00 ± 0.00**	-45.09 ± 5.98
BC	5204.70 ± 1386.33	3233.93 ± 416.85	**1000.00 ± 0.00**	**-30.49 ± 11.30**
Expert SAC	**5655.26 ± 882.56**	**4070.47 ± 307.83**	**1000.00 ± 0.00**	**-30.53 ± 7.83**
GAIL	4576.90 ± 1649.49	3476.49 ± 1.49	**1000.00 ± 0.00**	-35.51 ± 6.41
Modified SAC	2865.51 ± 1029.50	3363.64 ± 378.97	**1000.00 ± 0.00**	-52.08 ± 5.58

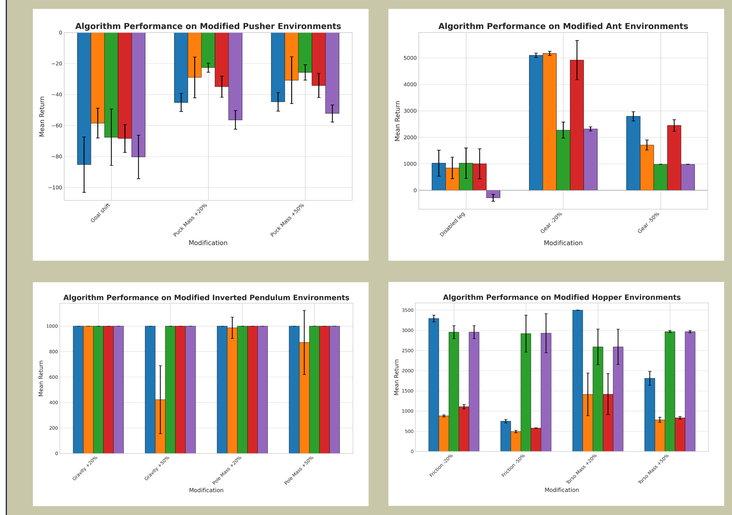


Figure 3: Results for each modified environment

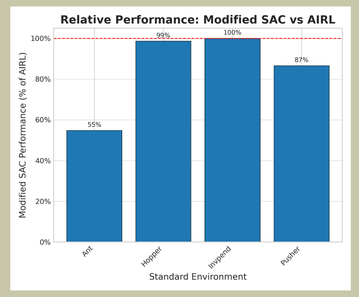


Figure 4: Performance recovery of SAC agent trained on AIRL reward

Conclusions

- **Fidelity Ceiling:** AIRL's learned reward can drive SAC to solve and even match expert performance on simple or moderately perturbed tasks (inverted pendulum, low friction hopper, etc.), but falls well short on complex locomotion (Ant) and goal-oriented tasks (Pusher)
- **Robustness Transfer:** The robustness baked into the expert AIRL policy only partially carries over via the learned reward – the larger the structural change, the more distortion the surrogate reward suffers.
- **Goal Transfer vs Policy Robustness:** These results suggest that while AIRL has the *potential* to learn a transferable goal that allow retraining in a new environment, the policy learned by AIRL in the original environment might sometimes be more robust when deployed zero-shot than the agent trained with the transferred AIRL reward in the modified environment.
- **Thoughts on Why AIRL Goal Transfer Struggled:** For starters, RL is sensitive to hyperparameter tuning, and IRL exacerbates this problem. It could also be that the AIRL discriminator only learns the behavior for local policies and forgets the state-action visitation distribution for state-action pairs explored during the initial training.

Future Work

- **Test on More Tasks:** This study used several environments, but a broader range of tasks and environments must be used (especially more complex real-world-like scenarios) before drawing any definitive conclusions.
- **Analyze What Makes Policies Robust:** Advanced methods were more robust zero-shot than simple copying. Future research would analyze *why* these policies were better at handling new situations.
- **Improve AIRL's Goal Learning:** These results showed that AIRL's goal learning struggled in complex tasks. Future work should focus on broader
- **Altered Data Quality:** This study was not testing the efficiency of these algorithms. All models were given perfect expert demonstrations and allowed to train for as long as possible. A future research direction could be rerunning these experiments with poorer data quality and for a limited amount of time.

Acknowledgements: Professor Cody Fleming for advising this project