Literature Review:
Inverse Reinforcement Learning for Video Games

Noa JELSCH
CESI École d'Ingénieurs

June 2025

Summary

This literature review summarizes the 2018 paper titled **Inverse Reinforcement Learning for Video Games**, authored by George Tucker, Rowan Gleave, and Stuart Russell. The work addresses a central problem in AI-driven behavior modeling: how to learn realistic, human-like agent policies in complex environments without explicit reward functions. The paper proposes applying Inverse Reinforcement Learning (IRL) to infer underlying reward structures from human gameplay demonstrations in video games. It builds on Adversarial IRL (AIRL) and adapts it to high-dimensional environments like Atari.

Objectives and Motivation

Traditional Reinforcement Learning (RL) can produce highly effective agents but often leads to inhuman or unnatural behavior, particularly when rewards are sparse or manually engineered. IRL shifts the focus to recovering the reward function from human demonstrations, thereby enabling agents to learn behavior that aligns more closely with human expectations and style. This is especially useful in video games where believability and immersion matter more than optimal performance.

Methodology

The method consists of:

- Collecting demonstration data from human players in two games: *Catcher* and *Atari Enduro*.
- Compressing high-dimensional visual inputs using an autoencoder to a latent space.
- Using Adversarial IRL to train a discriminator that distinguishes expert behavior from policy-generated behavior.
- Using the learned reward function to guide policy learning via Trust Region Policy Optimization (TRPO).

Results

- In Catcher (simple game), IRL reproduces expert-like behavior with high fidelity.
- In Atari Enduro, the learned reward function generalizes well and produces more natural driving behavior, though not yet achieving expert-level performance.
- The inferred reward function performs better than manually engineered ones on generalization to new states.

Discussion

This study validates IRL as a powerful alternative to hand-crafted reward design, particularly when aiming for realism and human-likeness. Key insights include:

- Latent space embeddings improve efficiency in high-dimensional visual environments.
- IRL helps avoid the reward hacking common in traditional RL.

Limitations include:

- Computational cost is significant.
- Performance depends heavily on the quality/diversity of demonstrations.
- The reward function remains difficult to interpret or verify.

Conclusion and Future Work

This paper is an important step toward training human-like NPCs using data-efficient and interpretable methods. Future directions may include:

- Combining IRL with preference learning.
- Extending the model to multiplayer or collaborative environments.
- Evaluating via human-player perception studies.

Reference

Tucker, G., Gleave, R., & Russell, S. (2018). *Inverse Reinforcement Learning for Video Games*. arXiv preprint arXiv:1810.10593.