"Adversarial Inverse Reinforcement Learning: A Study in Lunar Lander Simulations"

Tate Reynolds
u0578264@utah.edu

Tanmay Sharma
u1472860@utah.edu

1 Problem

Abstract Problem Definition: In the domain of Reinforcement Learning (RL), defining reward functions that accurately encapsulate the desired outcomes is a central challenge. Adversarial Inverse Reinforcement Learning (AIRL) posits a solution by learning reward functions from expert demonstrations. This approach is not without its complexities, but it offers a promising avenue for developing RL systems that are more aligned with human intentions and expertise.

Contextual Problem within the Experimental Domain: The Lunar Lander simulation presents a well-defined but complex problem space for applying and evaluating AIRL. It requires an RL agent to successfully land on a simulated lunar surface, managing finite resources and ensuring a safe descent. This task is an ideal testbed for AIRL, as it demands a nuanced understanding of the lander's dynamics and the environment, much like what would be required in real-world applications.

2 Justification

This investigation is pertinent to the broader goals of AI safety and alignment, as successful lunar landings are critical for advancing space exploration. By applying AIRL to the Lunar Lander simulation, we can assess the efficacy of learned reward functions in a controlled, yet meaningful and complex environment. The insights gained could inform the development of AI systems for actual space missions, where understanding and aligning with human objectives are crucial.

3 Literature Review

3.1 References

- Fu, J., Luo, K., & Levine, S. (2018). Learning Robust Rewards with Adversarial Inverse Reinforcement Learning.
- Wulfmeier, M., Ondruska, P., & Posner, I. (2016). Maximum Entropy Deep Inverse Reinforcement Learning.
- Ng, A. Y., & Russell, S. (2000). Algorithms for Inverse Reinforcement Learning.
- Abbeel, P., & Ng, A. Y. (2004). Apprenticeship learning via inverse reinforcement learning.
- Mnih, V. et al. (2015). Human-level control through deep reinforcement learning.

Brief Summaries Fu et al. (2018) present AIRL as an approach that formulates inverse RL as a game between adversaries, aiming for robust reward functions. Wulfmeier et al. (2016) delve into the entropic principles within inverse RL, crucial for understanding the range of strategies AIRL can encapsulate. The foundational works of Ng and Russell (2000) and Abbeel and Ng (2004) introduce methodologies that underpin AIRL's approach to learning from expert demonstrations. Mnih et al. (2015) illustrate the potential of deep learning to achieve expert-level control in RL tasks.

Class Project for CS 6960: Human-AI Alignment at the University of Utah, Fall 2022.

3.2 Identification of Gaps

Despite the depth of existing research, there is a notable gap in applying AIRL to the complex and dynamic environment of the Lunar Lander simulation. Current literature primarily focuses on more simplistic or terrestrial applications, leaving the potential of AIRL in space-related tasks underexplored. This gap represents an opportunity for significant contributions to the field, especially in understanding how AIRL can adapt to the diverse and unpredictable conditions of space exploration.

4 Method

Problem: Landing multiple lunar landers on the moon's surface without collisions using AIRL

4.1 Key Insight

In tackling the challenge of safely landing two lunar modules on the moon's surface without collisions, our key insight revolves around the application of Adversarial Inverse Reinforcement Learning (AIRL). This approach will allow us to derive a reward function that inherently balances the dual objectives of individual landing precision and inter-lander avoidance strategies.

4.2 High-Level Plan

Our approach will leverage the OpenAI Gym's Lunar Lander environment, modified to incorporate two landers within the simulation. The AIRL framework will enable us to not only deduce the optimal behaviors for safe landing but also to encode the nuances of spatial coordination between the two landers. A communication function will be developed to facilitate the sharing of positional and velocity data between the landers. The reward function will be engineered to incentivize safe, solo landings while imposing penalties for near-miss incidents or collisions. It will also integrate aspects of the landers' state space, including current position, velocity, and proximity to the other lander, while the action space will be composed of orientation and thrust decisions.

5 Experimental Design

5.1 Testable Hypothesis

Our hypothesis asserts that AIRL can effectively train two lunar landers to autonomously and simultaneously land on the lunar surface without collisions. By employing a shared reward function that reflects both individual landing success and mutual non-interference, the landers will learn to navigate and land in harmony.

5.2 Concrete Domain For Testing

The simulation will run within a customized version of the Lunar Lander environment that now includes two landers. This environment will serve as the testing domain where data on the landers' movements and interactions will be collected for analysis.

5.3 Experiment Description

The experimental process will be sequential. Initially, we will focus on training a single lander to perfect the landing process. Once proficiency is attained, a second lander will be introduced. Training will proceed with both landers, emphasizing the importance of collision avoidance as much as landing success. Evaluation criteria will be clearly defined: each lander must land vertically with legs in contact with the ground, neither lander should collide with the other, and all maneuvers must be completed within a set time frame. Rewards will be assigned for successful communication, landing execution, and collision avoidance. The experiment's success will be measured by the consistency with which both landers meet the evaluation criteria across multiple trials.