Inverse Learning from Gradient-Based Multi-Agent Learners (I-LOLA)

Zhongjian Yu, Yutian Lei, Zhengxun Yin, Chengrui Gao, Maorong Lin

1. Motivation & Problem

Inverse Reinforcement Learning (IRL) aims to recover an underlying reward function from observed agent behavior.

Most existing **multi-agent IRL** methods assume agents follow **Multi-Agent Soft Policy Iteration (MA-SPI)**, which:

A.works well in small, discrete environments;

B.is incompatible with modern **gradient-based learners** (e.g.policy gradient &PPO)

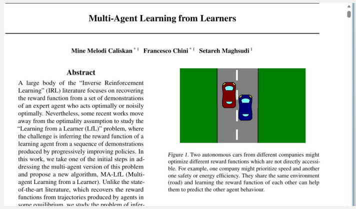
However, in realistic multi-agent systems:

A.agents are *learning*, not optimal;

B.policies evolve through gradient updates;

C.learning dynamics themselves contain reward information

Can we perform inverse learning directly from gradient-based multi-agent learners?



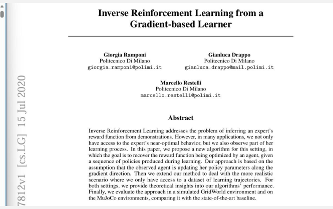
Multi-Agent Learning from Learners

Mine Melodi Caliskan *[1] Francesco Chini *[1] Setareh Maghsudi [1]

Abstract

A large body of the "Inverse Reinforcement Learning" (IRL) literature focuses on recovering the reward function from a set of demonstrations of an expert agent who acts optimally or noisily optimally. Nevertheless, some recent works move away from the optimality assumption to study the "Learning from a Learner (LfL)" problem, where the challenge is inferring the reward function of a learning agent from a sequence of demonstrations produced by progressively improving policies. In this work, we take one of the initial steps in addressing the multi-agent version of this problem and propose a new algorithm, MA-LfL (Multi-agent Learning from a Learner). Unlike the state-of-the-art literature, which recovers the reward functions from trajectories produced by agents in some equilibrium, we study the problem of infer-

Figure 1. Two autonomous cars from different companies might optimize different reward functions which are not directly accessible. For example, one company might prioritize speed and another one safety or energy efficiency. They share the same environment (road) and learning the reward function of each other can help them to predict the other agent behaviour.



7812v1 [cs.LG] 15 Jul 2020

Inverse Reinforcement Learning from a Gradient-based Learner

Giorgia Ramponi
Politecnico Di Milano
giorgia.ramponi@polimi.it

Gianluca Drappo
Politecnico Di Milano
gianluca.drappo@mail.polimi.it

Marcello Restelli
Politecnico Di Milano
marcello.restelli@polimi.it

Abstract

Inverse Reinforcement Learning addresses the problem of inferring an expert's reward function from demonstrations. However, in many applications, we not only have access to the expert's near-optimal behavior, but we also observe part of her learning process. In this paper, we propose a new algorithm for this setting, in which the goal is to recover the reward function being optimized by an agent, given a sequence of policies produced during learning. Our approach is based on the assumption that the observed agent is updating her policy parameters along the gradient direction. Then we extend our method to deal with the more realistic scenario where we only have access to a dataset of learning trajectories. For both settings, we provide theoretical insights into our algorithms' performance. Finally, we evaluate the approach in a simulated GridWorld environment and on the MuJoCo environments, comparing it with the state-of-the-art baseline.

2. Challenges

- **Model mismatch:** MA-SPI assumptions vs PPO learners
- **Coupled learning:** agents adapt while anticipating others
- **Reward ambiguity:** different rewards may induce identical behavior

3. Method Overview

We extend **LOGEL** (Inverse RL from a Gradient-Based Learner) to multi-agent settings by explicitly modeling **learning interactions**.

Key ideas:

Treat agents as **learners**, not equilibria

Infer rewards from **policy updates**, not final policies

Model opponent learning via **LOLA**

Two stages:

Stage A (I-LOGEL): independent learners

Stage B (I-LOLA): coupled learning dynamics

Optimization uses **black-box CMA-ES**, avoiding second-order gradients.

4. Experimental Setup

We evaluate across three increasingly realistic environments:

Task	Environment	Learner	Purpose
T1	GridWorld	MA-SPI	Assumption-matched baseline
T2	MPE simple_spread	PPO	Model mismatch test
T3	MultiWalker	PPO	Continuous-control feasibility

Metrics:

KL(1a): prediction error using true reward

KL(1b): prediction error using recovered reward

Induced return: usefulness of recovered reward

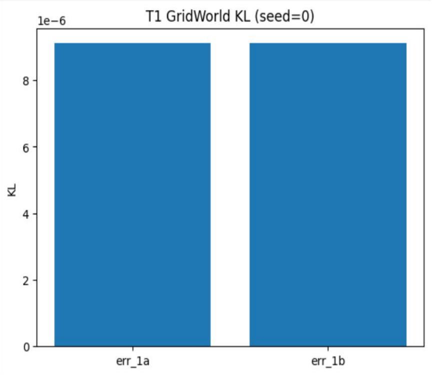


Figure 1: T1 KL comparison

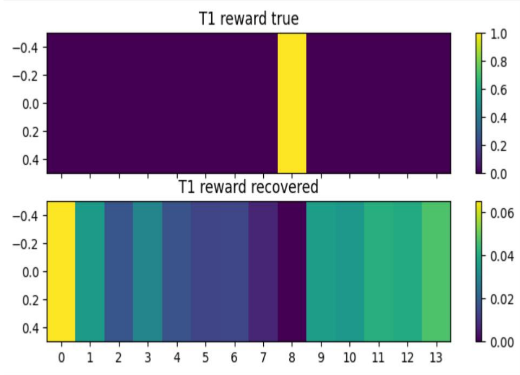


Figure 2: T1 reward heatmap

5.1 Results — T1 GridWorld

- KL(1a) ≈ KL(1b) ≈ 9×10^{-6}
- MA-LfL accurately predicts policy updates
- Recovered reward differs numerically but induces identical behavior

Interpretation:

MA-LfL works well when its assumptions hold.

5.2 Results — T2 MPE (PPO)

Learners trained with PPO (violating MA-SPI assumptions)

I-LOLA achieves:

KL(1a) ≈ 1.03×10^{-6}

KL(1b) ≈ 1.09×10^{-6}

Key result:

Recovered rewards predict future policy updates almost as accurately as true rewards.

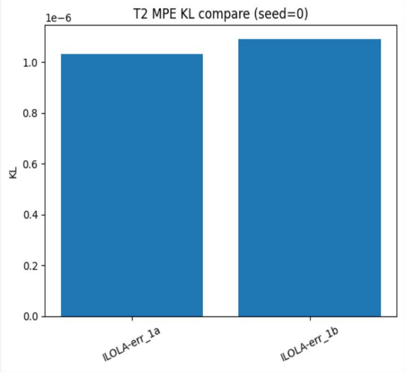


Figure 3: T2 PPO KL comparison

5.3Results — T3 MultiWalker

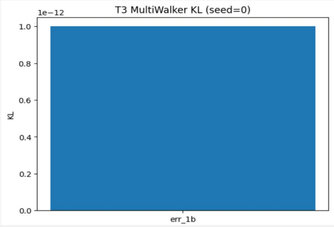


Figure 4: T3 KL (Monte Carlo)

Continuous-action, multi-agent physical environment

MA-LfL not applicable

I-LOLA remains numerically stable: Monte Carlo KL ≈ 5×10^{-4}

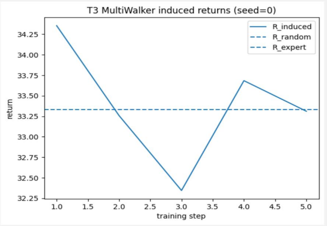


Figure 5: T3 induced returns

Induced policies achieve performance comparable to random baseline

Indicates recovered reward contains usable signal

Further optimization required

6. Takeaways & Limitations

Takeaways:

MA-LfL works when assumptions hold (T1)

I-LOLA remains accurate under PPO learners (T2)

Inverse learning is feasible in continuous control (T3)

Limitations

Single-seed experiments

No MA-LfL baseline in T2/T3

Induced performance not yet expert-level