Conflicting demonstrations in Inverse Reinforcement Learning

Author: Rafaël Labbé (r.m.labbe@student.tudelft.nl) Supervisor: Angelo Caregnato Neto Responsible Professor: Luciano Cavalcante Siebert

1. Background

Reinforcement Learning (RL) is used to train agents to solve problems by maximizing a reward function.

Reward functions are not always easy to define. Inverse reinforcement learning (IRL) can learn the reward function from expert demonstrations (for example, humans when driving cars).

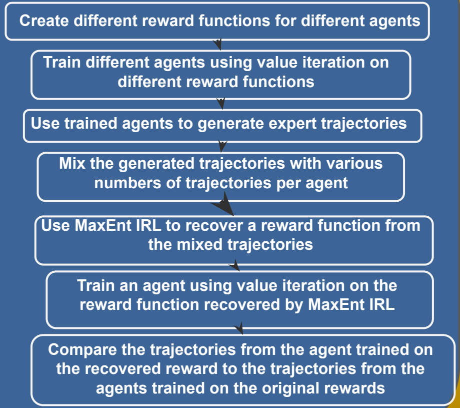
IRL is an ill-posed problem, since there are many reward functions that explain a given set of demonstrations. Maximum Entropy IRL (MaxEnt IRL) maximizes the entropy to select a reward function with minimal bias.

Sometimes different experts maximize different reward functions. In this case, it may be hard for the IRL algorithm to learn from the resulting conflicting demonstrations.

2. Research Question

To what extent can Inverse Reinforcement Learning learn rewards from conflicting demonstrations?

3. Methodology



Create different reward functions for different agents

Train different agents using value iteration on different reward functions

Use trained agents to generate expert trajectories

Mix the generated trajectories with various numbers of trajectories per agent

Use MaxEnt IRL to recover a reward function from the mixed trajectories

Train an agent using value iteration on the reward function recovered by MaxEnt IRL

Compare the trajectories from the agent trained on the recovered reward to the trajectories from the agents trained on the original rewards

4. Results

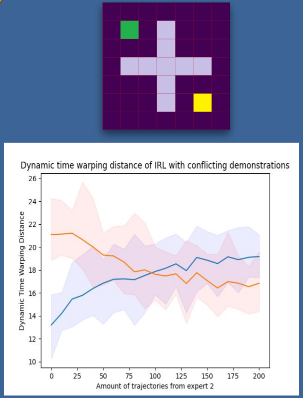


Figure 1: Dynamic Time warping distance in severely conflicting scenario. Blue for agent 1, orange for agent 2

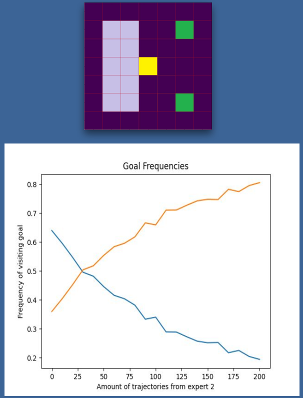


Figure 2: Blue shows goal frequencies of Green goals, orange of yellow goal.

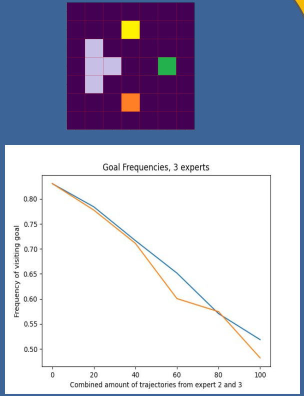


Figure 3: Blue shows goal frequencies of Green goals, orange of yellow goal.

Dark Purple squares are traversable squares without reward
Light purple squares are random starting points
Green, yellow and orange are prioritized goals for agent 1,2 and 3 respectively

5. Conclusion

- We can see conflicting demonstrations significantly impact the ability of MaxEnt IRL to recover the reward, such as in Figure 1.
- Easier rewards in conflicting demonstrations can significantly affect the ability of MaxEnt IRL to recover rewards of harder demonstrations as shown in figure 2.
- Whether the conflicting demonstrations come from 1 agent or from 2 different agents doesn't seem to make a difference as shown in figure 3.

6. References

1. [Ziebart et al., 2008] Ziebart, B. D., Maas, A., Bagnell, J. A., and Dey, A. K. (2008). Maximum entropy inverse reinforcement learning. In Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 3, AAAI'08, page 1433–1438. AAAI Press.