Effect of Demonstrations with Temporal Biases on Learning Rewards using Inverse Reinforcement Learning

Author Mateja Zatezalo (M.Zatezalo@student.tudelft.nl)
Professor Luciano Cavalcante Siebert
Supervisor Angelo Caregnato Neto

Introduction

- **Inverse Reinforcement Learning** to learn from expert demonstrations in order to obtain the maximized reward function in Markov Decision Process (MDP)
- **Cognitive biases** present a form of deviation from rationality that affects human decision-making
- **Temporal biases**: time consistent (present) and time inconsistent (temptation, pre-commitment)

Objective

"To what extent can IRL learn rewards from demonstrations that contain some form of temporal cognitive bias?"

- We perform this using **Maximum Entropy IRL (MEIRL)** algorithm [1]

Methodology

Environment:
6x6 Grid-World MDP inspired by [1]

Agents with biases:
- **Time consistent**: trained with adapted value iteration (1a) to involve bias with exponential discounting (1b)
- **Time inconsistent (Naive and Sophisticated)**: trained with value iteration (2a) adapted to [2], to involve biases with hyperbolic discounting (2b)

Experiment:
- Train agents to have biases and generate trajectories using policy
- Learn (recover) rewards from generated trajectories and original reward using MEIRL
- Evaluate performance of learning rewards
- Compare performance to that of unbiased(optimal) agent

$$V_s = \max_a \{R(s) + \gamma * p(s, s', a)V_{s'}\}$$

1a) Value of reward at state **s** for time consistent agent

$$D = \gamma^t$$

1b) Exponential discounting of reward

$$V_s = \max_{a,d} \left\{\frac{1}{1+kd} R(s,a) + p(s,s',a)V_{s'}\right\}$$

2a) Value of reward at state **s** for time inconsistent agents

$$D = \frac{1}{1+kd}$$

2b) Hyperbolic discounting of reward

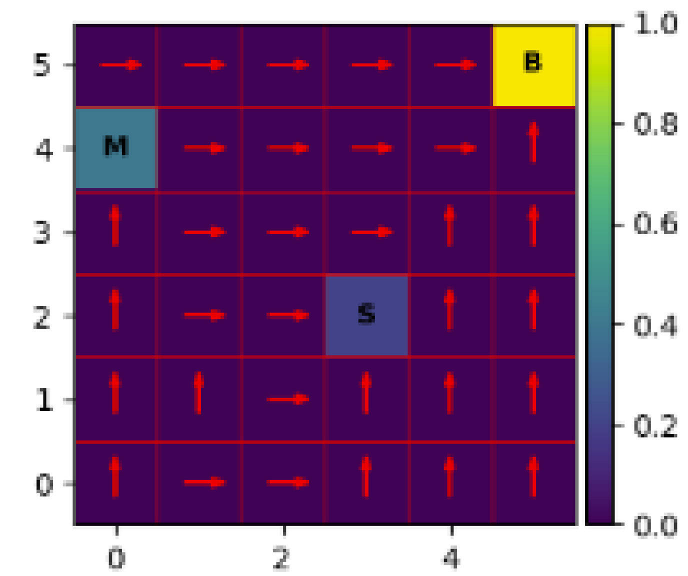

Figure 1a: Optimal policy of unbiased agent

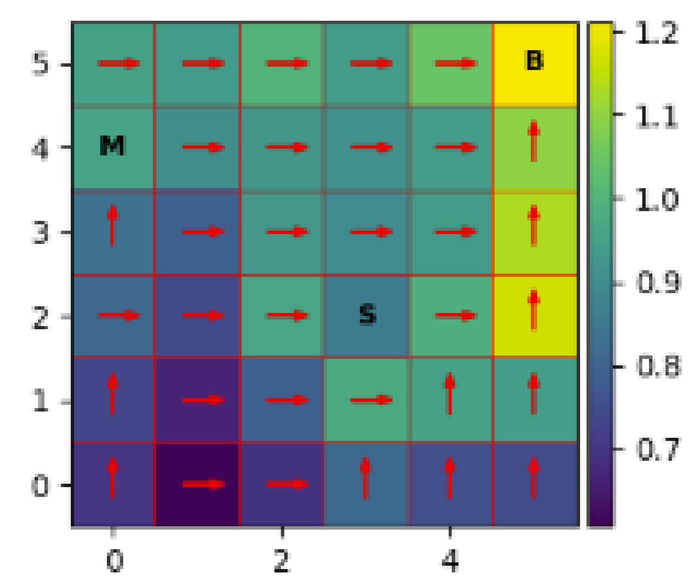

Figure 1b: Recovered reward using MEIRL

Figure 1: Example of optimal policy of agent and recovered reward using MEIRL

Results

- Time consistent:

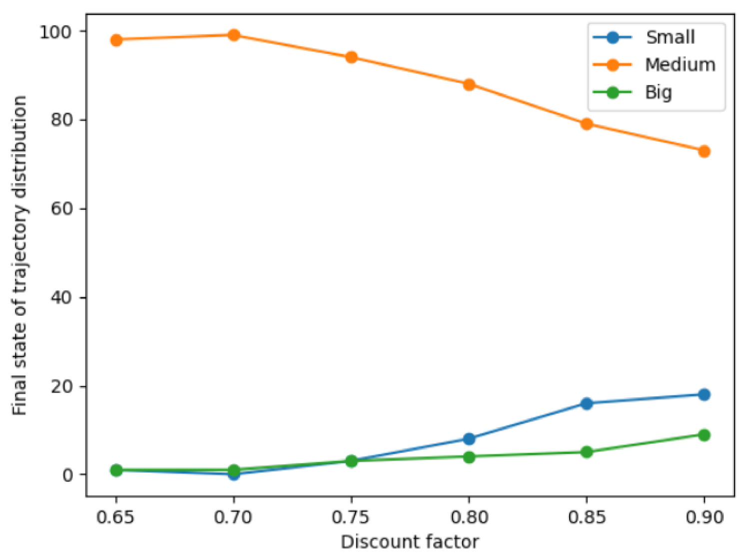

Figure 2: Trajectory distribution of the biased agent

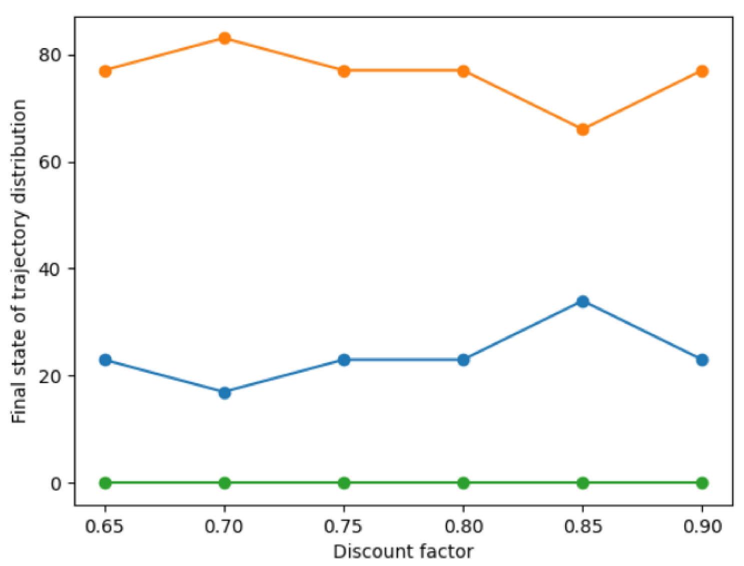

Figure 3: Trajectory distribution with recovered reward

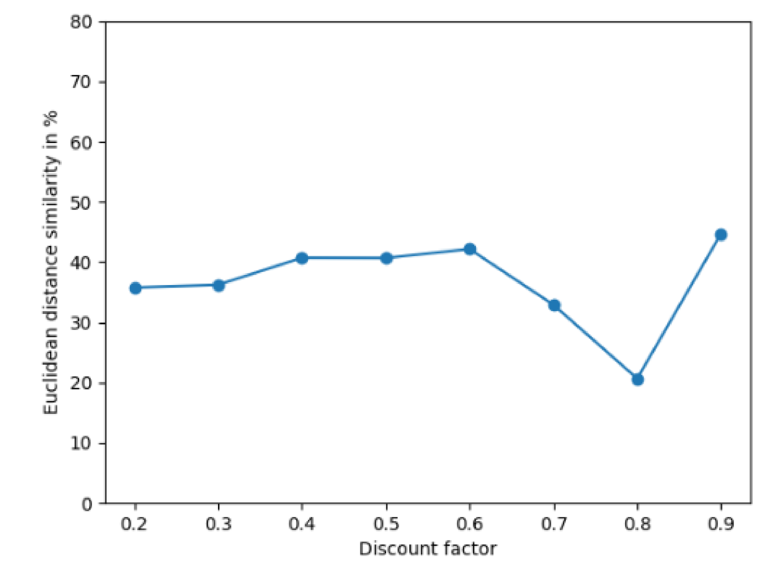

Figure 4: Euclidean similarity of trajectories of biased agent and unbiased agent with recovered reward

- Time inconsistent:
  Sophisticated
  Naive

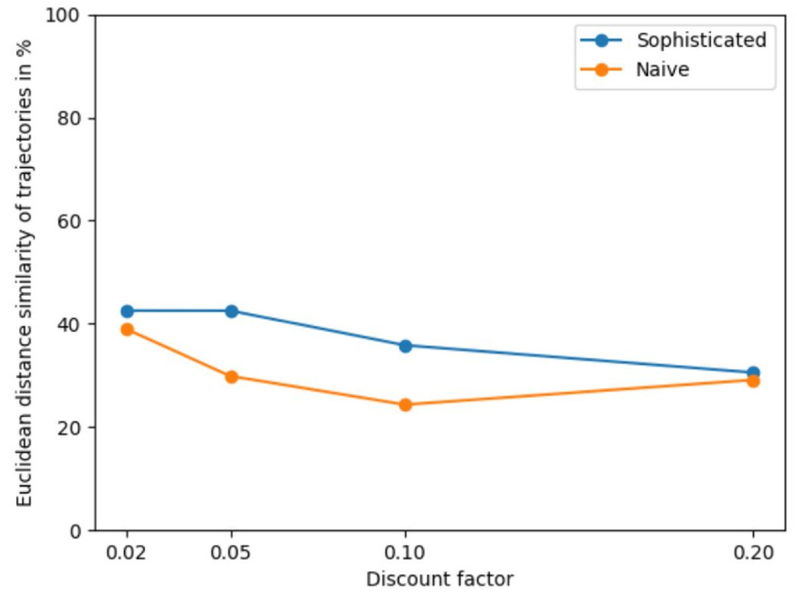

Figure 5: Euclidean similarity of trajectories of biased agents and unbiased agent with recovered reward

Limitations

- Creating synthesized data for demonstrations(agents) instead of real-life human data
- Implementation of Naive and Sophisticated agent may not replicate the supposed behavior to perfection due to adaptation in implementation
- Environment limited to 6x6 Grid-World MDP, can be expanded and enriched with features (e.g. walls)

Conclusion

- All our agents with temporal biases show a substantial effect on learning rewards, especially when compared to learning from unbiased agent
- Reward is recovered solidly from agents with time consistent bias
- Sophisticated and agent with time consistent bias learn rewards better than Naive
- Potential improvements and other temporal biases can expand this research topic

References
[1] B. D. Ziebart, A. I. R. Maas, J. A. Bagnell, and A. K. Dey, *Maximum entropy inverse reinforcement learning*. 2008, pp. 1433–1438. [Online]. Available: http://ai.stanford.edu/~amaas/papers/amaas_aaai.pdf
[2] O. Evans, A. Stuhlmueller, J. Salvatier and D. Filan, *Modeling Agents with Probabilistic Programs*. 2017. [Online]. Available: http://agentmodels.org