Bayesian Inverse Reinforcement Learning for strategy extraction in the iterated Prisoner's Dilemma game

Matthias Cami[1], *Supervisors*: Inês Terrucha[1,2], Yara Khaluf[3], and Pieter Simoens[1]

[1] Dept. of Information Technology - IDLab, Ghent University - imec, Belgium
[2] AI Lab, Vrije Universiteit Brussel, 1050 Brussels, Belgium
[3] Information Technology Group, Wageningen University and Research, The Netherlands

1 Introduction and Methodology

As Artificial Intelligence (AI) becomes more relevant in various fields, interactions between humans and artificial agents will become more and more common. In order to be trustworthy and acceptable as assistive agents, such agents should be able to predict and account for human preferences. In strategic situations as modelled in game theory, humans will often deviate from predicted equilibrium models. While it has been shown that AI agents can be trained to reach superhuman performance in zero-sum games like poker, their learned policies do not reflect typical human strategies and therefore are not suited to predict the actions of humans. However, this does not mean that humans act in an unpredictable manner, they follow their own preferences, that take into account both their opponent's actions and the context of the interaction. While a big part of AI research has been focusing on beating the opponent in zero-sum games, most interactions are actually "mixed motive", which means that the interests of the players are not completely aligned, but also not solely competitive. Since the goal of AI is to aid human decision-making, the problem then becomes: how can AI optimize for human social preferences that are not easily hard-coded but change according to different parameters?

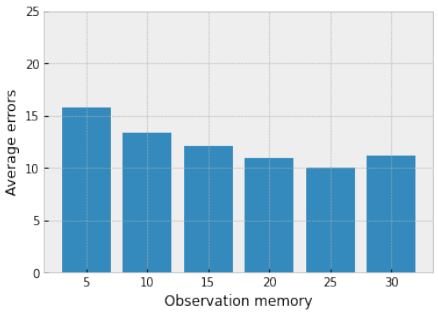
To tackle this question we use a classical mixed-motive game: the Prisoner's Dilemma (PD). Specifically, we focus on repeated interactions as they provide more diverse insights on how human preferences might change in accordance to the actions of the opponent. For this purpose we will use empirical data from the two-player iterated PD to illustrate the aforementioned dynamics [3]. To help AI infer human preferences in such setting we turn to Imitation Learning techniques, specifically the Bayesian Inverse Reinforcement Learning (BIRL) method [7]. As in Inverse Reinforcement Learning (IRL) [6], BIRL extracts the reward function from any given set of demonstrations with the added value that BIRL takes a probabilistic view of the reward. This means that, with BIRL, we can incorporate domain knowledge to choose the prior that selects from the (infinitely) many possible rewards for which the observed actions would be optimal.

2 Results and Discussion

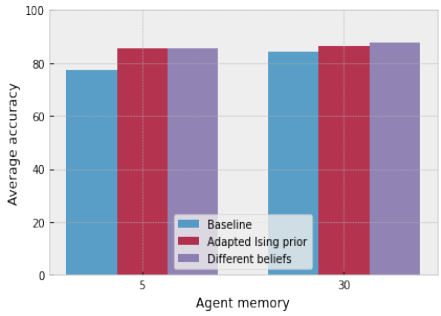
First, we tested whether the rewards humans follow are stationary. For this purpose we tested agents with different memory capacity about previous rounds of play. In Fig. a we show the results: it is clear that higher memory agents outperform the lower memory agents. Even larger memories were tested, but showed a decline in performance, suggesting that general human behavior is only stationary for a certain time frame.

To incorporate domain knowledge, in opposition to keeping the Uniform prior used in the baseline, we used an adapted version of the Ising prior [1]. We chose that for two reasons: to give more relevance to the more recent rounds when predicting the next [2, 5] and to emphasize a clear choice between the available actions by the expert. The results are seen in Fig. b where the 5-memory agent has the most significant increase in performance. This shows that a well constructed prior, using domain knowledge, can significantly help agents in their performance, even when there is not much data available to them.

The transition probabilities in the baseline model, which define the beliefs of the expert about their opponent's action, assumed equal chance of either action. To test different transition probabilities, we follow the principles of theory of the mind [4] and test whether assuming that the expert is able to correctly predict their opponent every period will influence accuracy. From the results in Fig. b we see that the accuracy increases very slightly, suggesting that even though perfect prediction is not realistic, randomness seems to perform worse.



(a) Average amount of errors for the baseline agent comparing different observation memories



(b) Average accuracy comparing 3 different agent setups on agents with memory 5 and 30

In conclusion we show that incorporating domain knowledge and probing the expert's beliefs increase the accuracy of the imitation technique used. The first proving to greatly increase the performance of the agent even for situations with small amount of available data (lower memory setups).

Acknowledgements

This research received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme.

References

1. Cipra, B.A.: An introduction to the ising model. The American Mathematical Monthly **94**(10), 937–959 (1987)
2. Gracia-Lázaro, C., Ferrer, A., Ruiz, G., Tarancón, A., Cuesta, J.A., Sánchez, A., Moreno, Y.: Heterogeneous networks do not promote cooperation when humans play a prisoner's dilemma. Proceedings of the National Academy of Sciences **109**(32), 12922–12926 (2012)
3. Grujić, J., Eke, B., Cabrales, A., Cuesta, J.A., Sánchez, A.: Three is a crowd in iterated prisoner's dilemmas: experimental evidence on reciprocal behavior. Scientific reports **2**(1), 1–7 (2012)
4. Jara-Ettinger, J.: Theory of mind as inverse reinforcement learning. Current Opinion in Behavioral Sciences **29**, 105–110 (2019)
5. Nay, J.J., Vorobeychik, Y.: Predicting human cooperation. PloS one **11**(5), e0155656 (2016)
6. Ng, A.Y., Russell, S.J., et al.: Algorithms for inverse reinforcement learning. In: Icml. vol. 1, p. 2 (2000)
7. Ramachandran, D., Amir, E.: Bayesian inverse reinforcement learning. In: IJCAI. vol. 7, pp. 2586–2591 (2007)