Opponent Modelling for Reinforcement Learning in Multi-Objective Normal Form Games*

Extended Abstract

Yijie Zhang
Universiteit van Amsterdam
The Netherlands
yijie.zhang@student.uva.nl

Roxana Rădulescu
Vrije Universiteit Brussel
Belgium
roxana.radulescu@vub.be

Patrick Mannion
National University of Ireland Galway
Ireland
patrick.mannion@nuigalway.ie

Diederik M. Roijers
HU University of Applied Sciences
Utrecht, The Netherlands
diederik.yamamoto-roijers@hu.nl

Ann Nowé
Vrije Universiteit Brussel
Belgium
ann.nowe@vub.be

ABSTRACT

In this paper, we investigate the effects of opponent modelling on multi-objective multi-agent interactions with non-linear utilities. Specifically, we consider multi-objective normal form games (MONFGs) with non-linear utility functions under the scalarised expected returns optimisation criterion. We contribute a novel actor-critic formulation to allow reinforcement learning of mixed strategies in this setting, along with an extension that incorporates opponent policy reconstruction using conditional action frequencies. Our empirical results demonstrate that opponent modelling can drastically alter the learning dynamics in this setting.

KEYWORDS

Multi-agent systems; multi-objective decision making; reinforcement learning; opponent modelling; game theory; Nash equilibrium

ACM Reference Format:
Yijie Zhang, Roxana Rădulescu, Patrick Mannion, Diederik M. Roijers, and Ann Nowé. 2020. Opponent Modelling for Reinforcement Learning in Multi-Objective Normal Form Games. In *Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), Auckland, New Zealand, May 9–13, 2020,* IFAAMAS, 3 pages.

1 OPPONENT MODELLING IN MONFGS

In many multi-agent interactions in the real world, agents receive payoffs over multiple distinct criteria; i.e. the payoffs are multi-objective in nature. However, the same multi-objective payoff vector may lead to different utilities for each participant. Therefore, it is essential for agents to learn about the behaviour of other agents.

We present the first study of the effects of opponent modelling on multi-objective multi-agent interactions with non-linear utilities. Specifically, we consider MONFGs [3, 6, 15, 18] with non-linear utility functions under the *scalarised expected returns (SER)* optimisation criterion [8, 12]. I.e., we are interested in the utility, $p_{u,i}$, an agent i derives from the expected payoff over multiple episodes:

$$p_{u,i} = u_i(\mathbb{E}[\mathbf{p}_i^{\boldsymbol{\pi}}]), \tag{1}$$

where $\mathbf{p}_i^{\boldsymbol{\pi}}$ is the payoff agent i receives after executing a joint (possibly mixed) strategy $\boldsymbol{\pi}$, and u_i is the utility function of agent i that maps the expected payoff vector to a scalar utility.

SER stands in contrast to expected scalarised results (ESR) [10], which is more common in game theory [8]. However, we argue that there are many settings in which interaction is repeated, and it is the expected payoff vector that induces the utility, leading to the SER criterion. SER is the typically employed criterion in multi-objective planning and reinforcement learning [11].

A mixed strategy profile π^{NE} is a *Nash equilibrium (NE)* [7] in a MONFG under SER if for all $i \in \{1, ..., N\}$ and all $\pi_i \in \Pi_i$, with Π_i the set of mixed strategies for agent i:

$$u_i\left[\mathbb{E}\,\mathbf{p}_i(\pi_i^{NE}, \pi_{-i}^{NE})\right] \geq u_i\left[\mathbb{E}\,\mathbf{p}_i(\pi_i, \pi_{-i}^{NE})\right] \tag{2}$$

i.e. π^{NE} is an NE under SER if no agent can increase the *utility of her expected payoffs* by deviating unilaterally from π^{NE}. Recent work [13, 14] has demonstrated that NE need not exist in MONFGs under SER with non-linear utility functions. For this paper, we study both MONFGs with (Section 2) and without NEs [20]. As the agents do not know each other's utility functions, it becomes key to explicitly learn about the other agents to reach favourable NEs.

For such opponent modelling, we employ policy reconstruction using conditional action frequencies [1, 17], i.e., an agent i maintains a set of beliefs regarding the strategy of the opponents, using empirical distributions derived from observing the actions of the opponent. These are then used to represent the policy π_{-i} of her opponent and to derive the valuation of her actions (marginalising out π_{-i}).

To exploit the opponent model we developed an actor-critic algorithm [16, 19]. This algorithm has 3 steps. After taking action a chosen from the policy $\pi(a|\boldsymbol{\theta})$, the agent observes a multi-objective payoff $\boldsymbol{p}$ as well as the opponent's action a'. Then, the agent updates its own estimate of the opponent's policy π'. In the second step, the agent updates its multi-objective joint action value estimate:

$$Q\left(a_t, a'_t\right) \leftarrow Q\left(a_t, a'_t\right) + \alpha_Q\left[\boldsymbol{p}_t - Q\left(a_t, a'_t\right)\right] \tag{3}$$

We note that in a many MONFG settings, the payoffs observed by the agent for known joint actions are deterministic. Equation 3

*An extended version of this paper is available [20].

Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), May 9–13, 2020, Auckland, New Zealand.

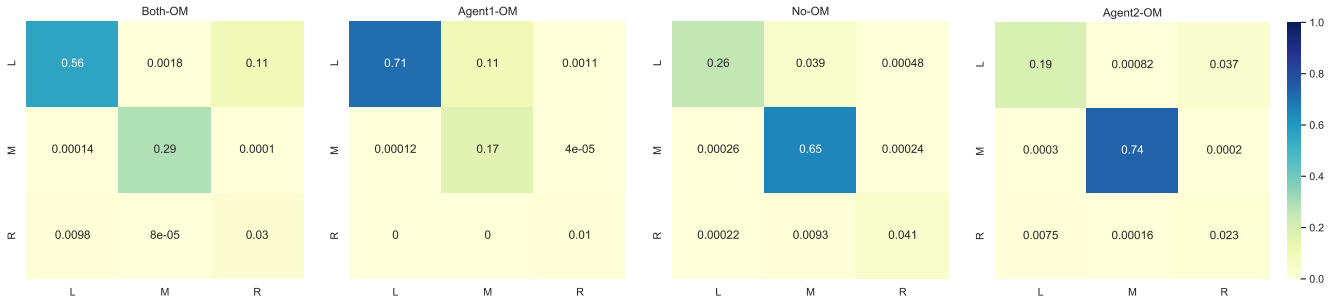

Figure 1: Results for the game in Table 1. The left column shows the estimated SER for Agent 1 (top) and Agent 2 (bottom) under the 4 experiment settings. The middle and right columns show the empirical outcome distributions.

	L	M	R
L	$(4, 1)$	$(1, 2)$	$(2, 1)$
M	$(3, 1)$	$(3, 2)$	$(1, 2)$
R	$(1, 2)$	$(2, 1)$	$(1, 3)$

Table 1: A MONFG with pure strategy NE in (L,L), (M,M), and (R,R), for utility functions $u_1([p^1, p^2]) = p^1 \cdot p^1 + p^2 \cdot p^2$, and $u_2([p^1, p^2]) = p^1 \cdot p^2$, under SER. NB: (L,L) and (M,M) Pareto-dominate (R,R). (L,L) offers the highest utility for the row player, and (M,M) for the column player.

applies to both deterministic and stochastic settings. After updating the joint action values, the third step is to compute the SER objective $J(\boldsymbol{\theta})$, taking derivatives with regard to agent i's strategy parameters, $\boldsymbol{\theta}$, and subsequently update $\boldsymbol{\theta}$ in the direction of the gradient:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha_\theta \nabla_{\boldsymbol{\theta}} J \tag{4}$$

For a full description of this algorithm and its derivation, please refer to the extended version of this paper [20].

2 RESULTS & DISCUSSION

To evaluate the impact of opponent modelling, we use, among others, a 2-objective MONFG (Table 1). For the other games we study, please refer to [20].

We consider four different settings: (1) neither agent performs opponent modelling; (2) both agents perform opponent modelling; (3) only agent 1 performs opponent modelling; (4) only agent 2 performs opponent modelling. For each setting, agents interact for 3000 episodes, averaged over 100 trials. Furthermore, in this experiment, the gradient $\nabla_{\boldsymbol{\theta}}$ is computed analytically w.r.t $J(\boldsymbol{\theta})$. An agent's strategy $\pi(a|\boldsymbol{\theta})$ is represented using a simple softmax function:

$$\pi(a = a_i|\boldsymbol{\theta}) = \frac{e^{\theta_i}}{\sum_{j=1}^{|A_i|} e^{\theta_j}} \tag{5}$$

The actor learning rate for the presented experimental results is $\alpha_\theta = 0.05$, while h, the opponent modelling window size is 100. For the setting without opponent modelling we used a critic learning rate $\alpha_Q = 0.05$. For the Opponent Modelling Actor-Critic approach, because the agents are learning the Q-function for the join-action space in a deterministic setting, we used $\alpha_Q = 1$. We note that we carried out an extensive analysis with respect to all these parameters and we present all the results in [20].

As the results (Figure 1) show, if only one of the agents is using opponent modelling, the agent that does the opponent modelling significantly benefits from doing so, with respect to the setting in which neither agent does OM. When both agents do OM, the distribution over the possible outcomes becomes more balanced. We thus conclude that opponent modelling can significantly benefit agents in MONFGs under SER that have Nash Equilibria.

We have also tested MONFGs in which there are no NEs [20]. For such games, the benefits of opponent modelling are not as good. On the contrary, in most settings, the agent performing OM seemed to be unable to accurately capture information regarding the opponent's strategy, and thus making decisions on the basis of incorrect or outdated information. Implementing OM in these settings does not confer a significant advantage in terms of outcomes, and when the learning parameters are not tuned well, it may even hurt the performance of both agents.

In conclusion, our studies of MONFGs under SER with non-linear utility functions demonstrated that opponent modelling can significantly alter the learning dynamics in MONFGs. In cases where NE are present, opponent modelling can confer significant benefits to agents that implement it. However, when there are no NE, we observe that an agent implementing opponent modelling can experience adverse effects on its utility. These adverse effects could be (mostly) mitigated after careful hyper-parameter optimisation of the learning algorithm, but did not contribute to the utility of the agent implementing the opponent modelling. This is highly surprising, and does not occur in the single-objective setting – where there are always NE in mixed strategies. Therefore, in future work, we aim to investigate if more sophisticated schemes for opponent modelling [9], such as explicitly modelling the (properties of) the utility function of the opponent (e.g., using preference elicitation [2, 4, 5, 21]), can make opponent modelling effective in all MONFGs.

Acknowledgments

This research is partially funded by the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" program.

REFERENCES

[1] Stefano V. Albrecht and Peter Stone. 2018. Autonomous agents modelling other agents: A comprehensive survey and open problems. *Artificial Intelligence* 258 (2018), 66 – 95.

[2] Nawal Benabbou, Cassandre Leroy, and Thibaut Lust. 2020. An Interactive Regret-Based Genetic Algorithm for Solving Multi-Objective Combinatorial Optimization Problems. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI'20)*.

[3] David Blackwell et al. 1956. An analog of the minimax theorem for vector payoffs. *Pacific J. Math.* 6, 1 (1956), 1–8.

[4] Urszula Chajewska, Daphne Koller, and Ronald Parr. 2000. Making rational decisions using adaptive utility elicitation. In *AAAI/IAAI*. 363–369.

[5] Shengbo Guo, Scott Sanner, and Edwin V Bonilla. 2010. Gaussian process preference elicitation. In *Advances in neural information processing systems*. 262–270.

[6] Dmitrii Lozovanu, D Solomon, and A Zelikovsky. 2005. Multiobjective games and determining Pareto-nash equilibria. *Buletinul Academiei de Ştiinţe a Republicii Moldova. Matematica* 3 (2005), 115–122.

[7] John Nash. 1951. Non-Cooperative Games. *Annals of Mathematics* 54, 2 (1951), 286–295.

[8] Roxana Rădulescu, Patrick Mannion, Diederik M Roijers, and Ann Nowé. 2020. Multi-objective multi-agent decision making: a utility-based analysis and survey. *Autonomous Agents and Multi-Agent Systems* 34 (2020). https://doi.org/10.1007/s10458-019-09433-x

[9] Roberta Raileanu, Emily Denton, Arthur Szlam, and Rob Fergus. 2018. Modeling Others using Oneself in Multi-Agent Reinforcement Learning. In *International Conference on Machine Learning (ICML)*. 4254–4263.

[10] Diederik M Roijers, Denis Steckelmacher, and Ann Nowé. 2018. Multi-objective Reinforcement Learning for the Expected Utility of the Return. In *Proceedings of the Adaptive and Learning Agents workshop at FAIM*.

[11] Diederik M Roijers, Peter Vamplew, Shimon Whiteson, and Richard Dazeley. 2013. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research* 48 (2013), 67–113.

[12] Diederik M Roijers and Shimon Whiteson. 2017. Multi-objective decision making. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 11, 1 (2017), 1–129.

[13] Roxana Rădulescu, Patrick Mannion, Diederik M Roijers, and Ann Nowé. 2019. Equilibria in Multi-Objective Games: a Utility-Based Perspective. In *Proceedings of the Adaptive and Learning Agents Workshop (ALA-19) at AAMAS*.

[14] Roxana Rădulescu, Patrick Mannion, Yijie Zhang, Diederik Martin Roijers, and Ann Nowé. 2020. A utility-based analysis of equilibria in multi-objective normal form games. *arXiv preprint arXix:2001.08177* (2020). https://arxiv.org/abs/2001.08177

[15] Lloyd S Shapley and Fred D Rigby. 1959. Equilibrium points in games with vector payoffs. *Naval Research Logistics Quarterly* 6, 1 (1959), 57–61.

[16] Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction* (second ed.). The MIT Press. http://incompleteideas.net/book/the-book-2nd.html

[17] William Uther and Manuela Veloso. 1997. *Adversarial reinforcement learning*. Technical Report. Technical report, Carnegie Mellon University, 1997. Unpublished.

[18] Mark Voorneveld, Dries Vermeulen, and Peter Borm. 1999. Axiomatizations of Pareto equilibria in multicriteria games. *Games and economic behavior* 28, 1 (1999), 146–154.

[19] Ronald J. Williams. 1992. Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning. *Mach. Learn.* 8, 3-4 (May 1992), 229–256. https://doi.org/10.1007/BF00992696

[20] Yijie Zhang, Roxana Rădulescu, Patrick Mannion, Diederik Martin Roijers, and Ann Nowé. 2020. Opponent Modelling using Policy Reconstruction for Multi-Objective Normal Form Games. In *Proceedings of the Adaptive and Learning Agents Workshop (ALA-20) at AAMAS*. (under review).

[21] Luisa M Zintgraf, Diederik M Roijers, Sjoerd Linders, Catholijn M Jonker, and Ann Nowé. 2018. Ordered preference elicitation strategies for supporting multi-objective decision making. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 1477–1485.