Search-Improved Game-Theoretic Multiagent Reinforcement Learning in General and Negotiation Games

Extended Abstract

Zun Li
DeepMind, U. of Michigan, Ann Arbor
USA
lizun@umich.edu

Marc Lanctot
DeepMind
Canada
lanctot@deepmind.com

Kevin R. McKee
DeepMind
United Kingdom
kevinrmckee@deepmind.com

Luke Marris
DeepMind
United Kingdom
marris@deepmind.com

Ian Gemp
DeepMind
United Kingdom
imgemp@deepmind.com

Daniel Hennes
DeepMind
France
hennes@deepmind.com

Kate Larson
University of Waterloo, DeepMind
Canada
katelarson@deepmind.com

Yoram Bachrach
DeepMind
United Kingdom
yorambac@deepmind.com

Michael P. Wellman
University of Michigan, Ann Arbor
USA
wellman@umich.edu

Paul Muller
DeepMind
France
pmuller@deepmind.com

ABSTRACT

Multiagent reinforcement learning (MARL) has benefited significantly from population-based and game-theoretic training regimes. One approach, Policy-Space Response Oracles (PSRO), employs standard reinforcement learning to compute response policies via approximate best responses and combines them via meta-strategy selection. We augment PSRO by adding a novel search procedure with generative sampling of world states, and introduce two new meta-strategy solvers based on the Nash bargaining solution. We evaluate PSRO's ability to compute approximate Nash equilibrium, and its performance in negotiation games: Colored Trails and Deal-or-no-Deal. We conduct behavioral studies where human participants negotiate with our agents ($N = 346$). Search with generative modeling finds stronger policies during both training time and test time, enables online Bayesian co-player prediction, and can produce agents that achieve comparable social welfare negotiating with humans as humans trading among themselves.

KEYWORDS

Policy-Space Response Oracles, AlphaZero, Nash Bargaining Solution, Negotiation Games, Multiagent, Reinforcement Learning

ACM Reference Format:
Zun Li, Marc Lanctot, Kevin R. McKee, Luke Marris, Ian Gemp, Daniel Hennes, Kate Larson, Yoram Bachrach, Michael P. Wellman, and Paul Muller. 2023. Search-Improved Game-Theoretic Multiagent Reinforcement Learning in General and Negotiation Games: Extended Abstract. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 3 pages.

1 INTRODUCTION

Learning to act from experience in an environment with multiple learning agents is a difficult problem.

One class of MARL algorithms, policy-space response oracles (PSRO), follows the spirit of fictitious play and its generalizations [7, 8, 10, 15], decomposing the goal into two steps via empirical game-theoretic analysis (EGTA) [21]: *i)* the *meta-strategy solver (MSS) step*, a process which chooses which policy a player should play from a library, and *ii)* the *best response (BR) step* which computes approximate best response policies to the distribution over opponents' policies, adding them to the library.

We propose a training regime for multiagent (partially observable) general sum, n-player, and negotiation games using game-theoretic RL. We extend PSRO as follows: (i) We integrate an Monte Carlo tree search (MCTS) AlphaZero-style approximate best response into the *best-response step*, incorporating deep-generative models into the training loop, which allows us to tractably represent belief-states during search in large imperfect information games). (ii) We introduce and evaluate several new *meta-strategy solvers*, including those based on bargaining theory, which are particularly well-suited for negotiation games. (iii) We conduct an extensive evaluation across a variety of benchmark games and in two negotiation games, including one with human participants.

Due to the space limitations, we present only an overview of our algorithm and subset of our results. The full paper is found at [12].

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

2 SEARCH-IMPROVED GENERATIVE PSRO

Empirical game-theoretic analysis (EGTA) [21] is an approach to reasoning about large sequential games through normal-form **empirical game** models, induced by simulating enumerated subsets of the players' full policies in the sequential game. Policy-Space Response Oracles (PSRO) [10] uses EGTA to incrementally build up each player's set of policies ("oracles") through repeated applications of approximate best response using RL.

We integrate search into the best response step based on Approximate Best Response (ABR) [20], which uses a variant of Information Set Monte Carlo tree search [2] called IS-MCTS-BR. At the root of the IS-MCTS-BR search (starting at information set s), the posterior distribution over world states, $\Pr(h \mid s, \pi_{-i})$ is computed explicitly, which requires both (i) enumerating every history in s, and (ii) computing the opponents' reach probabilities for each history in s. Then, during each simulation step, a world state is sampled from this belief distribution, then the game-tree regions are explored in a similar way as in the vanilla MCTS, and finally the statistics are aggregated on the information-set level. Steps (i) and (ii) are prohibitively expensive in games with large belief spaces. Hence, we propose learning a generative model online during the BR step.

We introduce new meta-strategy solvers based on the Nash Bargaining Solution (NBS) [16]. Define the set of achievable payoffs as all expected utilities $u_i(\mu)$ under a joint-policy profile μ [6, 14]. Denote the disagreement outcome of player i, which is the payoff it gets if no agreement is achieved, as d_i. The **Nash bargaining score** is: $\max_{\mu \in \Delta(\Pi)} \Pi_{i \in \mathcal{N}} (u_i(\mu) - d_i)$; the NBS is the joint policy that maximizes this score. When $n = 2$ this leads to a quadratic program (QP) with the constraints derived from the policy space structure [5]. Even in this simplest case, the objective is non-concave posing a problem for most QP solvers. Scaling to n players requires higher-order polynomial solvers. Instead, we solve for the NBS using projected gradient ascent, and use it at an optimization criteria for other solution concepts (correlated equilibria) [13].

3 EXPERIMENTS IN A NEGOTIATION GAME

"Deal or No Deal" (DoND) is a simple alternating-offer bargaining game with incomplete information, which has been used in many AI studies [1, 3, 9, 11]. Our focus is to train RL agents to play against humans *without human data*, similar to previous work [19]. Two players are assigned *private* preferences $\mathbf{v}_1 \geq \mathbf{0}, \mathbf{v}_2 \geq \mathbf{0}$ for three different items (books, hats, and basketballs). At the start of the game, there is a pool $\mathbf{p}$ of 5 to 7 items drawn randomly such that: (i) the total value for a player of all items is 10: $\mathbf{v}_1 \cdot \mathbf{p} = \mathbf{v}_2 \cdot \mathbf{p} = 10$, (ii) each item has non-zero value for at least one player: $\mathbf{v}_1 + \mathbf{v}_2 > \mathbf{0}$, (iii) some items have non-zero value for both players, $\mathbf{v}_1 \odot \mathbf{v}_2 \neq \mathbf{0}$, where $\odot$ represents element-wise multiplication. The players take turns proposing how to split the pool of items, for up to 10 turns (5 turns each). If an agreement is not reached, the negotiation ends and players both receive 0. Otherwise, the agreement represents a split of the items to each player, $\mathbf{p}_1 + \mathbf{p}_2 = \mathbf{p}$, and player i receives a utility of $\mathbf{v}_i \cdot \mathbf{p}_i$. DoND is an imperfect information game because the other player's preferences are private. We use a database of 6796 bargaining instances made publicly available in [11].

We recruited participants from Prolific [17, 18] to evaluate the performance of our agents in DoND (overall $N = 346$; 41.4% female, 56.9% male, 0.9% trans or nonbinary; median age range: 30–40). Crucially, participants played DoND for real monetary stakes, with an additional payout for each point they earned in the game.

Agent	$\bar{u}_{\text{Humans}}$	$\bar{u}_{\text{Agent}}$	$\bar{u}_{\text{Comb}}$	D%	NBS
IndRL	5.86 [5.37, 6.40]	**6.50 [5.93, 7.06]**	6.18 [5.82, 6.56]	0.96	38.12
Com1	5.14 [4.56, 5.63]	5.49 [4.87, 6.11]	5.30 [4.93, 5.76]	0.90	28.10
Com2	6.00 [5.49, 6.55]	5.54 [4.96, 6.10]	5.76 [5.33, 6.12]	0.92	33.13
Coop	6.71 [6.23, 7.20]	6.17 [5.66, 6.64]	6.44 [6.11, 6.75]	**1**	41.35
Fair	**7.39 [6.89, 7.87]**	5.98 [5.44, 6.49]	**6.69 [6.34, 7.01]**	**1**	**44.23**

Table 1: Humans versus Agents performance with $N = 129$ human participants, 547 games total. $\bar{u}_X$ refers to the average utility to group X (for the humans when playing the agent, or for the agent when playing the humans), Comb refers to Combined, D% is the proportion of deals accepted. Square brackets indicate 95% confidence intervals.

We trained 112 agents using search-augmented PSRO with generative world state sampling for 15-20 iterations.

As detailed in [12], these agents vary in terms of MSS, backpropagation type, and final extraction technique. We then ran tournaments to rank and select from four representative categories: (i) the most competitive agents (maximizing utility), (ii) the most cooperative agents (maximizing social welfare), the (iii) the fairest agent (minimizing social inequity [4]); (iv) we add a separate category of the top-performing independent RL agent trained in self-play (DQN).

We collect data under two conditions: human vs. human (HvH), and human vs. agent (HvA). In the HvH condition, we collect 483 games: 482 end in deals made (99.8%), and achieve a return of 6.93 (95% c.i. [6.72, 7.14]), on expectation. We collect 547 games in the HvA condition: 526 end in deals made (96.2%; see Table 1). There are several observations: first, DQN achieves the highest individual return. By looking at the combined reward, it achieves this by aggressively reducing the human reward (down to 5.86)–possibly by playing a policy that is less human-compatible. The competitive PSRO agents seem to do the same, but without overly exploiting the humans, resulting in the lowest social welfare overall. The cooperative agent achieves significantly higher combined utility playing with humans. Better yet is Human/Fair, the only Human vs. Agent combination to achieve social welfare comparable to the Human vs. Human social welfare.

Overall, the fair agent is both adaptive to many different types of agents, and cooperative, increasing the social welfare in all the groups it negotiated with. This could be due to its MSS (MGCE) putting significant weight on many policies leading to Bayesian prior with high support (similarly to the uniform distribution over self-play checkpoints method in Fictitious Co-Play, which collaborated well with humans in Overcooked [19]), and/or its backpropagation of the product of utilities rather than individual return.

REFERENCES

[1] Kris Cao, Angeliki Lazaridou, Marc Lanctot, Joel Z. Leibo, Karl Tuyls, and Stephen Clark. 2018. Emergent communication through negotiation. In *Sixth International Conference on Learning Representations.*

[2] Peter I. Cowling, Edward J. Powley, and Daniel Whitehouse. 2012. Information set Monte Carlo tree search. *IEEE Transactions on Computational Intelligence and AI in Games* 4 (2012), 120–143. Issue 2.

[3] David DeVault, Johnathan Mell, and Jonathan Gratch. 2015. Toward natural turn-taking in a virtual human negotiation agent. In *AAAI Spring Symposium on Turn-taking and Coordination in Human-Machine Interaction.*

[4] E. Fehr and K. Schmidt. 1999. A theory of fairness, competition and cooperation. *Quarterly Journal of Economics* 114 (1999), 817–868.

[5] Christopher Griffin. 2010. Quadratic programs and general-sum games. In *Game Theory: Penn State Math 486 Lecture Notes.* 138–144. https://docs.ufpr.br/~volmir/Math486.pdf.

[6] John C Harsanyi and Reinhard Selten. 1972. A generalized Nash solution for two-person bargaining games with incomplete information. *Management science* 18 (1972), 80–106.

[7] Johannes Heinrich, Marc Lanctot, and David Silver. 2015. Fictitious self-play in extensive-form games. In *Thirty-Second International Conference on Machine Learning.*

[8] Johannes Heinrich and David Silver. 2016. Deep reinforcement learning from self-play in imperfect-information games. *CoRR* abs/1603.01121 (2016).

[9] Minae Kwon, Siddharth Karamcheti, Mariano-Florentino Cuellar, and Dorsa Sadigh. 2021. Targeted data acquisition for evolving negotiation agents. In *Thirty-Eighth International Conference on Machine Learning*, Vol. 139. 5894–5904.

[10] Marc Lanctot, Vinicius Zambaldi, Audrunas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Perolat, David Silver, and Thore Graepel. 2017. A unified game-theoretic approach to multiagent reinforcement learning. In *Thirtieth International Conference on Neural Information Processing Systems.*

[11] Mike Lewis, Denis Yarats, Yann N. Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or no deal? End-to-end learning for negotiation dialogues. In *2017 Conference on Empirical Methods in Natural Language Processing.*

[12] Zun Li, Marc Lanctot, Kevin R. McKee, Luke Marris, Ian Gemp, Daniel Hennes, Paul Muller, Kate Larson, Yoram Bachrach, and Michael P. Wellman. 2023. Combining Tree-Search, Generative Models, and Nash Bargaining Concepts in Game-Theoretic Reinforcement Learning. https://doi.org/10.48550/ARXIV.2302.00797

[13] Luke Marris, Paul Muller, Marc Lanctot, Karl Tuyls, and Thore Graepel. 2021. Multi-agent training beyond zero-sum with correlated equilibrium meta-solvers. In *Twenty-Eighth International Conference on Machine Learning.*

[14] Peter Morris. 2012. *Introduction to game theory.* Springer Science & Business Media.

[15] Paul Muller, Shayegan Omidshafiei, Mark Rowland, Karl Tuyls, Julien Pérolat, Siqi Liu, Daniel Hennes, Luke Marris, Marc Lanctot, Edward Hughes, Zhe Wang, Guy Lever, Nicolas Heess, Thore Graepel, and Rémi Munos. 2019. A generalized training approach for multiagent learning. In *Eighth International Conference on Learning Representations.*

[16] John Nash. 1950. The bargaining problem. *Econometrica* 18, 2 (1950), 155–162.

[17] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. 2017. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology* 70 (2017), 153–163.

[18] Eyal Pe'er, David Rothschild, Andrew Gordon, Zak Evernden, and Ekaterina Damer. 2021. Data quality of platforms and panels for online behavioral research. *Behavior Research Methods* (2021), 1–20.

[19] DJ Strouse, Kevin McKee, Matt Botvinick, Edward Hughes, and Richard Everett. 2021. Collaborating with Humans without Human Data. In *Thirty-Fifth Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. 14502–14515.

[20] Finbarr Timbers, Nolan Bard, Edward Lockhart, Marc Lanctot, Martin Schmid, Neil Burch, Julian Schrittwieser, Thomas Hubert, and Michael Bowling. 2022. Approximate exploitability: Learning a best response in large games. In *Thirty-First International Conference on Artificial Intelligence.*

[21] Michael P. Wellman. 2006. Methods for empirical game-theoretic analysis. In *Twenty-First National Conference on Artificial Intelligence.*