On-the-fly Strategy Adaptation for ad-hoc Agent Coordination

Extended Abstract

Jaleh Zand
University of Oxford
Oxford, UK
jz@robots.ox.ac.uk

Jack Parker-Holder
University of Oxford
Oxford, UK
jackph@robots.ox.ac.uk

Stephen J. Roberts
University of Oxford
Oxford, UK
sjrob@robots.ox.ac.uk

ABSTRACT

Training agents in cooperative settings offers the promise of AI agents able to interact effectively with humans (and other agents) in the real world. Multi-agent reinforcement learning (MARL) has the potential to achieve this goal, demonstrating success in a series of challenging problems. However, whilst these advances are significant, the vast majority of focus has been on the *self-play* paradigm. This often results in a *coordination problem*, caused by agents learning to make use of arbitrary conventions when playing with themselves. This means that even the strongest self-play agents may have very low *cross-play* with other agents, including other initializations of the same algorithm. In this paper we propose to solve this problem by adapting agent strategies *on the fly*, using a posterior belief over the other agents' strategy. Concretely, we consider the problem of selecting a strategy from a finite set of previously trained agents, to play with an unknown partner. We propose an extension of the classic statistical technique, Gibbs sampling, to update beliefs about other agents and obtain close to optimal ad-hoc performance. Despite its simplicity, our method is able to achieve strong cross-play with unseen partners in the challenging card game of Hanabi, achieving successful ad-hoc coordination without knowledge of the partner's strategy a priori.

KEYWORDS

Multi-Agent Systems; Cooperative Multi-Agent Systems; Multi-Agent Reinforcement Learning; Bayesian Inference; Gibbs Sampling

ACM Reference Format:
Jaleh Zand, Jack Parker-Holder, and Stephen J. Roberts. 2022. On-the-fly Strategy Adaptation for ad-hoc Agent Coordination: Extended Abstract. In *Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), Online, May 9–13, 2022*, IFAAMAS, 3 pages.

1 INTRODUCTION

Many of the most prominent successes in MARL have come in the *self-play* paradigm [13]. It is shown that in a two-player game, a self-play strategy could converge to a Nash equilibrium [10] and could even lead to super-human performance in certain domains [1, 4, 5, 15]. However the self-play framework seems more beneficial in an adversarial MARL setting compared to a cooperative one. In this work we consider the *ad-hoc coordination problem* [12], whereby an agent has to coordinate with an unknown partner in either an ad-hoc paradigm or with just a few successive trials.

Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), P. Faliszewski, V. Mascardi, C. Pelachaud, M.E. Taylor (eds.), May 9–13, 2022, Online.

2 ON-THE-FLY STRATEGY ADAPTATION

In line with the *theory of mind*, one possible approach for coordination in multi-agent ad-hoc problems is the formation of beliefs regarding strategies that are played by the other agents in the system. We consider fully-cooperative Markov games and we model this setting with a Decentralized Partially-Observable Markov Decision Process (Dec-POMDP [3, 9]). Further in our setting the Markov state s_t, consists of discrete features, f^{s_t}, which are themselves composed of public features, f^{pub,s_t}, and private features, f^{pri,s_t}. In addition, we assume two types of agents in the system; simple agents, A^S, that play a fixed policy, π^S, and complex agents, A^C, which coordinate with the set of simple agents and further have access to the set of policies, $\Pi = \{\pi^1, ..., \pi^n\}$. In our ad-hoc coordination settings, the complex agents, A^C, need to estimate the joint probability distribution $P(\pi^S, f^{S,s_t} \mid u_t^S, s_t)$ at every step of the game, where π^S is the simple agent policy, and $f^{S,s_t} \in F^{S,s_t} = \{f^{1,s_t}, ..., f^{n,s_t}\}$. Here, F^{S,s_t} denotes a set of features that A^S could be privy to but are hidden from A^C in state s_t.

In order to estimate this joint distribution we utilize an extension of the Gibbs sampling algorithm [6], on-the-fly Strategy Adaptation (OSA). The full OSA procedure is shown in Algorithm 1.

Algorithm 1 OSA algorithm

1: **Initialize:**
   $\hat{\pi}_0^S \leftarrow \pi^i, \pi^i \in \Pi = \{\pi^1, ..., \pi^n\}$
2: **while** the game is ongoing **do**
3: **if** it is A^C turn to play **then**
4: $f_{t+1}^{S,s_t} \sim P(f^{S,s_t} \mid \hat{\pi}_t^S, u_t^S, s_t)$
5: **for** $\pi^i \in \{\pi^1, \ldots, \pi^n\}$ **do**
6: **if** $P(u_t^S \mid \pi^i, s_t) \approx 0$ **then**
7: $\Pi = \Pi \setminus \{\pi^i\}$ ▷ Remove redundant policies
8: **end if**
9: **end for**
10: $\pi_{t+1}^S \sim P(\pi^S \mid f_{t+1}^{S,s_t}, u_t^S, s_t)$
11: $\hat{\pi}_{t+1}^S = \text{Mode}(\pi_1^S, ..., \pi_{t+1}^S)$ ▷ $\hat{\pi}_t^S$ is the most frequent π_t^S
12: $\pi_{t+1}^C = B(\hat{\pi}_{t+1}^S)$ ▷ B: optimal response policy function
13: **end if**
14: **end while**

Noting that all agents observe u_t^S at every step, we may use Bayes' theorem to estimate the distributions in steps 4 and 10 of Algorithm 1 as follows:

$$P(f^{S,s_t}|\pi_t^S, u_t^S, s_t) = \frac{P(u_t^S|f^{S,s_t}, \pi_t^S, s_t)P(f^{S,s_t} \mid \pi_t^S, s_t)}{\sum_{f^{i,s_t}} P(u_t^S|f^{i,s_t}, \pi_t^S, s_t)P(f^{i,s_t} \mid \pi_t^S, s_t)} \quad (1)$$

$$P(\pi^S|f_{t+1}^{S,s_t}, u_t^S, s_t) = \frac{P(u_t^S|\pi^S, f_{t+1}^{S,s_t}, s_t)P(\pi^S \mid f_{t+1}^{S,s_t}, s_t)}{\sum_{\pi^i} P(u_t^S|\pi^i, f_{t+1}^{S,s_t}, s_t)P(\pi^i \mid f_{t+1}^{S,s_t}, s_t)} \quad (2)$$

Table 1: Mean ± standard error of rewards for A^C playing each of the 7 policies in the two Hanabi experiments.

	MAPPO-1	MAPPO-2	Holmesbot	Iggi	Piers	Rainbow	Valuebot
$\pi^S \in \Pi$:							
$\mathbb{E}_\tau R(\tau)$ no OSA	8.98 ± 0.12	10.15 ± 0.12	5.13 ± 0.10	7.48 ± 0.09	8.10 ± 0.10	8.87 ± 0.11	6.01 ± 0.10
$\mathbb{E}_\tau R(\tau)$ with OSA	22.96 ± 0.05	21.74 ± 0.06	13.79 ± 0.10	16.09 ± 0.04	16.35 ± 0.04	19.98 ± 0.07	16.10 ± 0.07
$\pi^S \notin \Pi$:							
$\mathbb{E}_\tau R(\tau)$ no OSA	6.47 ± 0.11	7.94 ± 0.12	3.36 ± 0.09	5.90 ± 0.09	6.59 ± 0.10	6.69 ± 0.10	3.93 ± 0.09
$\mathbb{E}_\tau R(\tau)$ with OSA	10.44 ± 0.12	11.90 ± 0.13	13.31 ± 0.10	10.03 ± 0.10	9.92 ± 0.11	11.36 ± 0.11	14.28 ± 0.09

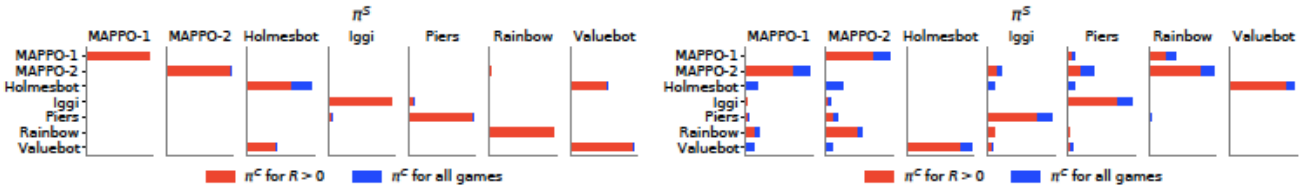

Figure 1: Distribution of π_T^C for each π^S in the first Hanabi experiment (left) and the second Hanabi experiment (right).

Table 2: Mean ± standard error of rewards in k-shot games. A^C plays each of the 7 policies using OSA and $\pi^S \notin \Pi$.

	MAPPO-1	MAPPO-2	Holmesbot	Iggi	Piers	Rainbow	Valuebot
$\mathbb{E}_\tau(R(\tau) \mid k)$:							
$k = 0$	10.44 ± 0.12	11.90 ± 0.13	13.31 ± 0.10	10.03 ± 0.10	9.92 ± 0.11	11.36 ± 0.11	14.28 ± 0.09
$k = 1$	12.84 ± 0.12	13.76 ± 0.12	16.45 ± 0.08	10.76 ± 0.10	10.86 ± 0.11	13.45 ± 0.11	16.72 ± 0.08
$k = 4$	14.52 ± 0.11	14.86 ± 0.11	17.33 ± 0.07	11.05 ± 0.10	11.43 ± 0.10	14.21 ± 0.10	17.24 ± 0.07
$\max_{i=0}^{n}(\mathbb{E}_\tau R_i)$	16.41 ± 0.11	15.66 ± 0.11	17.49 ± 0.07	13.24 ± 0.09	12.75 ± 0.09	16.30 ± 0.10	17.44 ± 0.07

3 EXPERIMENTS

We examine ad-hoc coordination in the challenging Hanabi learning environment [2]. We consider a wide range of agents: Rainbow [8], hand coded bots from the Hanabi Open Agent Dataset (HOAD, [11]), Valuebot, Holmesbot, Iggi and Piers, and Multi Agent PPO (MAPPO, [14]). Notably, this includes both on policy (MAPPO) and off policy (Rainbow) deep RL approaches, as well as scripted bots, providing a diverse range of agents. For the MAPPO model, we use two separate trained policies with different seeds.

We focus on a 2-player version of Hanabi, consisting of agents A^C and A^S, and conduct two sets of experiments. We further assume that self-play is the optimal response policy in our experiments, given that in a two-player non zero-sum-game setting where regret is minimised, a self-play strategy could still converge to a Nash equilibrium and perform very well [7]. In the first Hanabi experiment agent A^S's policy is included in the A^C's policy set, therefore $\pi^S \in \Pi$. In this experiment, OSA is able to recover performance close to self-play. In the second experiment, we consider a more challenging setting, where A^C does not have A^S's policy, $\pi^S \notin \Pi$. Despite not containing the optimal policy, A^C still successfully plays with A^S using the policy in its set that is most correlated with π^S for majority of the games. The results are shown in Table 1.

In order to better understand the strength and weakness of the method, we assess the distribution of the policies that A^C plays in coordination with A^S for each Hanabi experiment. Figure 1 depicts the distribution for π_T^C in the first Hanabi experiment ($\pi^S \in \Pi$) and the second Hanabi experiment ($\pi^S \notin \Pi$), in which π_T^C is agent A^C's policy at the end of each game. The plots show that in the first Hanabi experiment, π^S is the dominant policy selected by A^C to play A^S, and in the second Hanabi experiment, where $\pi^S \notin \Pi$, the policy with the highest cross-play score with π^S is the dominant strategy for A^C to play with A^S.

Next we define a k-shot ad-hoc game whereby A^C plays with A^S a total of k times in the second Hanabi game ($\pi^S \notin \Pi$). Table 2 shows the results for k-shot games. These results are compared with the maximum cross-play scores of A^C policies against π^S.

4 CONCLUSION

On-the-fly Strategy Adaptation (OSA), a novel approach for coordination between ad-hoc agents across a set of diverse models, achieves impressive performance yet scales gracefully. Average rewards for both ad-hoc and k-shot ad-hoc games show performance improves in k-shot games.

REFERENCES

[1] Yu Bai and Chi Jin. 2020. Provable self-play algorithms for competitive reinforcement learning. In *37th International Conference on Machine Learning, ICML 2020*. PMLR, 551–560.

[2] Nolan Bard, Jakob N Foerster, Sarath Chandar, Neil Burch, Marc Lanctot, H Francis Song, Emilio Parisotto, Vincent Dumoulin, Subhodeep Moitra, Edward Hughes, et al. 2020. The Hanabi challenge: A new frontier for AI research. *Artificial Intelligence* 280 (2020), 103216.

[3] Daniel S Bernstein, Robert Givan, Neil Immerman, and Shlomo Zilberstein. 2002. The complexity of decentralized control of Markov decision processes. *Mathematics of operations research* 27, 4 (2002), 819–840.

[4] Vincent Conitzer and Tuomas Sandholm. 2007. AWESOME: A general multiagent learning algorithm that converges in self-play and learns a best response against stationary opponents. *Machine Learning* 67, 1-2 (2007), 23–43.

[5] Jacob W Crandall and Michael A Goodrich. 2011. Learning to compete, coordinate, and cooperate in repeated games using reinforcement learning. *Machine Learning* 82, 3 (2011), 281–314.

[6] Stuart Geman and Donald Geman. 1984. Stochastic Relaxation, Gibbs Distributions and Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-6, 6 (1984), 721–741. https://doi.org/10.1109/TPAMI.1984.4767596

[7] Richard Gibson. 2013. Regret Minimization in Non-Zero-Sum Games with Applications to Building Champion Multiplayer Computer Poker Agents. *CoRR* abs/1305.0034 (2013). arXiv:1305.0034 http://arxiv.org/abs/1305.0034

[8] Matteo Hessel, Joseph Modayil, Hado van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Gheshlaghi Azar, and David Silver. 2018. Rainbow: Combining Improvements in Deep Reinforcement Learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, Sheila A. McIlraith and Kilian Q. Weinberger (Eds.). AAAI Press, 3215–3222.

[9] Ranjit Nair, Milind Tambe, Makoto Yokoo, David Pynadath, and Stacy Marsella. 2003. Taming decentralized POMDPs: Towards efficient policy computation for multiagent settings. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence IJCAI-03*, Vol. 3. International Joint Conferences on Artificial Intelligence Organization, 705–711.

[10] John F Nash. 2016. 4. The Bargaining Problem. In *The Essential John Nash*. Princeton University Press, 37–48.

[11] Aron Sarmasi, Timothy Zhang, Chu-Hung Cheng, Huyen Pham, Xuanchen Zhou, Duong Nguyen, Soumil Shekdar, and Joshua McCoy. 2021. HOAD: The Hanabi Open Agent Dataset. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*. 1646–1648.

[12] Peter Stone, Gal A. Kaminka, Sarit Kraus, and Jeffrey S. Rosenschein. 2010. Ad Hoc Autonomous Agent Teams: Collaboration without Pre-Coordination. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*. 1504–1509.

[13] Gerald Tesauro. 1994. TD-Gammon, a self-teaching backgammon program, achieves master-level play. *Neural computation* 6, 2 (1994), 215–219.

[14] Chao Yu, Akash Velu, Eugene Vinitsky, Yu Wang, Alexandre M. Bayen, and Yi Wu. 2021. The Surprising Effectiveness of MAPPO in Cooperative, Multi-Agent Games. *CoRR* abs/2103.01955 (2021). arXiv:2103.01955 https://arxiv.org/abs/2103.01955

[15] Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione. 2007. Regret minimization in games with incomplete information. *Advances in neural information processing systems* 20 (2007), 1729–1736.