Emergent Correlated Equilibrium through Synchronized Exploration

Mark Beliaev*[1], Woodrow Z. Wang*[2], Daniel A. Lazar[1], Erdem Bıyık[3], Dorsa Sadigh[2,3] and Ramtin Pedarsani[1]
[1] Electrical and Computer Engineering, UC Santa Barbara, Emails: {mbeliaev,dlazar,ramtin}@ucsb.edu
[2] Computer Science, Stanford University, Emails: {wwang153,dorsa}@stanford.edu
[3] Electrical Engineering, Stanford University, Email: ebiyik@stanford.edu

I. Introduction

As Reinforcement Learning (RL) methods become more successful in training agents to solve human-relevant tasks, they become more common in real world applications. However, in a multi-agent setting, it is challenging to train the agents in a decentralized way to learn prosocial behaviours. Deep-Q-Networks (DQNs) and policy gradient methods such as REINFORCE have proven successful in single agent RL [1]–[3]. Unfortunately these methods do not succeed when applied naively to Multi-Agent Reinforcement Learning (MARL) problems [4]. The introduction of a non-stationary environment caused by multiple agents simultaneously changing their policies breaks the Markov assumption required for convergence of Q-learning, and exacerbates the problem of high variance gradient estimates in policy gradient methods. To address this, Lowe *et al.* extend policy gradient methods by introducing a centralized critic as well as policy ensembles, allowing multiple learning agents to cooperate and compete in partially observable Markov games [5]. We first try to extend these methods in a game-theoretic paradigm, directly applying them to the Iterated Game of Chicken.

	Dare	**Chicken**
Dare	0, 0	7, 2
Chicken	2, 7	6, 6

TABLE I: Payoff table for the Game of Chicken. The first and second numbers are the rewards for the row and column players, respectively.

Table I shows the payoff matrix for the well-known Game of Chicken. In Iterated Game of Chicken (IGC), each turn the row and column players simultaneously select their actions and receive rewards accordingly. This game is played out for multiple turns, with each agent wishing to maximize its reward summed over the turns. One solution concept is the *Pure Nash Equilibrium* (PNE), in which each agent has no incentive to deviate from a deterministic strategy, given the strategy of their opponent. In this game the two PNE are (Chicken, Dare) and (Dare, Chicken). Although this is a good predictor of rational behaviour, there exists a more general notion of equilibrium that still maintains rationality and allows for better social outcomes: *Correlated Equilibrium* (CE). A CE strategy is any randomized assignment of potentially correlated *action recommendations* that no party wants to deviate from [6]. If one player chooses to follow the recommendation, the other player's best response would be to follow their respective recommendation as well. A game can have more than one CE, and PNE are a subset of CE.

In the IGC setting, we consider an oracle which at each turn sends each agent a signal that encodes the recommendation. Each agent only observes their own signal, but the oracle correlates the two revealed signals each turn. One example of CE in this game is when the oracle draws one of three pairs of numbers and gives them to the agents - $(0, 0)$, $(0, 1)$, and $(1, 0)$ with probabilities 0.5, 0.25 and 0.25, respectively. Note that this alone is not a CE unless the agents associate these signals accordingly, choosing to Chicken when receiving a 0, and to Dare when receiving a 1. This strategy constitutes a CE since neither party would benefit from deviating if the other party chose to follow the signal. Unlike the PNE mentioned earlier, this CE strategy is prosocial as it achieves a larger expected sum of rewards when considering both agents.

We observe that vanilla DQNs, REINFORCE, and MADPPG fail to reach CE in the IGC setting, converging to PNE instead. Although these methods were not designed with this goal in mind, other works considering a similar setting have significant drawbacks. Borowski *et al.* provide a method to provably reach CE in repeated matrix games, but it does not generalize beyond that setting [7]. Greenwald *et al.* design learning algorithms to reach correlated equilibrium in general stochastic games, however their method requires agents to either directly share their Q-values with one another, or have a rich enough observation to be able to estimate each others' Q-values [8]. In our work, we seek to drop this interdependence and instead reach CE by coordinating *when* agents explore.

Our main contributions in this paper are two-fold:

- We propose *Synchronized ϵ-Greedy Exploration*, which builds on the commonly-used ϵ-greedy exploration, and therefore can be generalized to stochastic games and used in any off-policy learning algorithm.
- We test our method on two different games: IGC and Grandmas' Cookies. Our method is the only one that successfully reaches CE in IGC. None of the methods reach CE in Grandmas' Cookies, but we hope that discussing the difficulty of this problem and outlining possible solutions will spur further research into this interesting field.

II. Methods

We present brief descriptions of methods that we hypothesized could reach CE, but failed in our experiments. We then present our most successful proposed approach, which supplements DQN with a synchronized exploration strategy. A modified

* Equal contribution.

exploration strategy is desirable as it allows agents to retain self-interest in the long run and can be applied to other algorithms that allow off-policy learning.

REINFORCE. We adapt REINFORCE to the multi-agent setting. In contrast to DQN, REINFORCE and policy gradient methods directly optimize the policy, which can be helpful when it is particularly difficult to estimate the Q-function [2].

MADDPG. Lowe *et al.* propose this as a multi-agent adaptation of actor-critic policy gradient methods that considers the policies of other agents [5]. They show its efficacy in mixed cooperative-competitive environments.

Learning with Opponent-Learning Awareness (LOLA). Foerster *et al.* have shown success with LOLA on matrix games such as Iterated Prisoner's Dilemma [9]. LOLA accounts for the impact of one agent's policy on the anticipated parameter update of the other agents.

Deep-Q-Network (DQN). Due to difficulties reaching CE using policy gradient approaches, we propose extensions of the DQN. Since much of the difficulty of reaching correlated equilibria in a multi-agent setting is that agents have to explore a certain combination of actions together, we experiment with the following proposed exploration strategies during training.

Independent ϵ-Greedy Exploration: Each agent independently follows an ϵ-greedy strategy: With probability $1-\epsilon$, the agent follow its policy, and with probability ϵ, the agent selects an action randomly with uniform probability.

Synchronized ϵ-Greedy Exploration: With probability $1-\epsilon$, all agents follow their policies, and with probability ϵ, all agents select actions randomly with uniform probability.

III. Experiments

We perform experiments on two environments of increasing complexity: IGC and Grandmas' Cookies. With our proposed exploration method, we reach the desired CE in the simpler IGC environment. However, our method fails to reach CE in the more complex Grandmas' Cookies environment.

A. Iterated Game of Chicken (IGC)

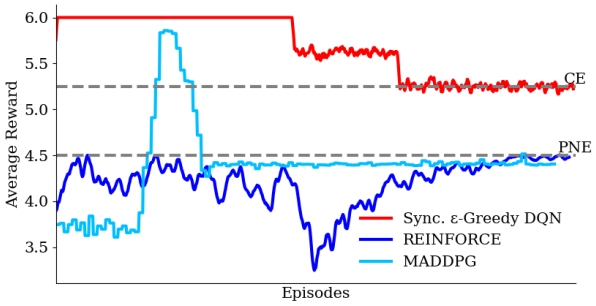

Fig. 1: Results from our proposed method and previous methods on IGC. The x-axis is scaled to normalize for convergence time. The two dashed lines correspond to average rewards for the PNE and CE.

Figure 1 shows the results. MADPPG, REINFORCE, and Independent ϵ-Greedy DQN all converge to a PNE in IGC and fail to converge to the desired CE. Although LOLA reaches higher average reward than all other algorithms, it does not converge to any Nash Equilibrium, violating the basic precept of myopic self-interest. Because of this, we exclude it from comparison of methods achieving game-theoretic equilibria, which we consider as good predictors of rational behavior. With *Synchronized ϵ-Greedy Exploration*, we are able to reach CE with the DQN. By synchronizing the agents' exploration, we increase the probability of the agents simultaneously performing the actions required for CE.

B. Grandmas' Cookies

We develop a new game, Grandmas' Cookies, that extends IGC to be a stochastic game with a larger state-action space, and temporally separating actions and their rewards. We explain a 2-agent version of this game, as difficulties in this setting prevented our consideration of more players.

Grandma's house and two kids are spawned randomly on a grid. Grandma's house is a part of the environment, while the kids are trainable agents. Each kid can move freely in the four cardinal directions throughout the grid, and both kids and Grandma's house can occupy the same tile. In addition, each kid can also choose to eat cookies at Grandma's house or do nothing, giving them a total of 6 actions each. When they are both in the same location as Grandma's house, their rewards are as follows: $(1, -4)$ if one eats cookies while the other does nothing respectively, $(1, 1)$ if they both eat cookies, and $(0, 0)$ otherwise. To make this game an extension of IGC, when both kids eat simultaneously, Grandma's house spawns in a random unoccupied location. If a kid is outside of Grandma's house, the kid receives a reward of -9 associated with being bored. This ensures that kids should prefer to be in the house even if they are not eating. There is an additional movement tax of 0.9378 subtracted as well. Finally we provide both agents with a correlated signal that is distributed identically to the one in IGC. This makes the CE for this game under the grid size of 3×3 identical to that of IGC, where chickening out is analogous to doing nothing, and daring is analogous to eating.

Each agent receives the full information about the state of the world, but only the correlated signal corresponding to their recommendation. The agents then choose one of 6 actions, where eating cookies is regarded the same as doing nothing when the agent is outside of Grandma's house.

Results on Grandmas' Cookies: The algorithms fail to reach CE in the Grandmas' Cookies environment, which would be achieved if agents chose to eat cookies or do nothing based on the recommendation from the oracle's signal. Although providing a correlated signal to the agents introduces a CE strategy profile, there is no direct incentive for agents to correlate their actions with this signal. Even with our synchronized exploration approach, we find agents are indifferent to the signal, *i.e.* the signal does not influence action Q-values.

IV. Future Directions

We have proposed exploration algorithms that succeed in reaching CE in repeated matrix games but do not yet succeed in more general stochastic games. There are a number of promising directions in this regard. Specifically, we consider a variety of ways to explicitly condition the exploration strategies on the signal, as well as combining the exploration strategy with proximal policy methods [10]. We hope that these directions will enable more prosocial MARL agents.

References

[1] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," 2013, cite arxiv:1312.5602Comment: NIPS Deep Learning Workshop 2013. [Online]. Available: http://arxiv.org/abs/1312.5602.

[2] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Mach. Learn.*, vol. 8, no. 3–4, 229–256, May 1992, ISSN: 0885-6125. DOI: 10.1007/BF00992696. [Online]. Available: https://doi.org/10.1007/BF00992696.

[3] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Proceedings of the 12th International Conference on Neural Information Processing Systems*, ser. NIPS'99, Denver, CO: MIT Press, 1999, 1057–1063.

[4] L. Matignon, G. Laurent, and N. Fort-Piat, "Independent reinforcement learners in cooperative markov games: A survey regarding coordination problems," *The Knowledge Engineering Review*, vol. 27, pp. 1 –31, Mar. 2012. DOI: 10.1017/S0269888912000057.

[5] R. Lowe, Y. WU, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Advances in Neural Information Processing Systems 30*, Curran Associates, Inc., 2017, pp. 6379–6390. [Online]. Available: http://papers.nips.cc/paper/7217-multi-agent-actor-critic-for-mixed-cooperative-competitive-environments.pdf.

[6] R. J. Aumann, "Subjectivity and correlation in randomized strategies," *Journal of mathematical Economics*, vol. 1, no. 1, pp. 67–96, 1974.

[7] H. P. Borowski, J. R. Marden, and J. S. Shamma, *Learning efficient correlated equilibria*, 2015. arXiv: 1512.02160 [cs.GT].

[8] A. Greenwald and K. Hall, "Correlated-q learning," in *In AAAI Spring Symposium*, AAAI Press, 2003, pp. 242–249.

[9] J. N. Foerster, R. Y. Chen, M. Al-Shedivat, S. Whiteson, P. Abbeel, and I. Mordatch, "Learning with opponent-learning awareness," *CoRR*, vol. abs/1709.04326, 2017. arXiv: 1709.04326. [Online]. Available: http://arxiv.org/abs/1709.04326.

[10] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, *Proximal policy optimization algorithms*, 2017. arXiv: 1707.06347 [cs.LG].