Population-aware Online Mirror Descent for Mean-Field Games by Deep Reinforcement Learning

Extended Abstract

Zida Wu
University of California, Los Angeles
Los Angeles, USA
zdwu@ucla.edu

Mathieu Laurière
New York University Shanghai
Shanghai, China
ml5197@nyu.edu

Samuel Jia Cong Chua
University of California, Los Angeles
Los Angeles, USA
samuelchua2001@gmail.com

Matthieu Geist
Cohere
France
matthieu@cohere.com

Olivier Pietquin
Cohere
France
olivier@cohere.com

Ankur Mehta
University of California, Los Angeles
Los Angeles, USA
mehtank@ucla.edu

ABSTRACT

Mean Field Games (MFGs) have the ability to handle large-scale multi-agent systems, but learning Nash equilibria in MFGs remains a challenging task. In this paper, we propose a deep reinforcement learning (DRL) algorithm that achieves population-dependent Nash equilibrium without the need for averaging or sampling from history, inspired by Munchausen RL and Online Mirror Descent. Through the design of an additional inner-loop replay buffer, the agents can effectively learn to achieve Nash equilibrium from any distribution, mitigating catastrophic forgetting. The resulting policy can be applied to various initial distributions. Numerical experiments on four canonical examples demonstrate our algorithm has better convergence properties than SOTA algorithms, in particular a DRL version of Fictitious Play for population-dependent policies.

KEYWORDS

Mean Field Game; Multi-agent System; Reinforcement Learning

ACM Reference Format:
Zida Wu, Mathieu Laurière, Samuel Jia Cong Chua, Matthieu Geist, Olivier Pietquin, and Ankur Mehta. 2024. Population-aware Online Mirror Descent for Mean-Field Games by Deep Reinforcement Learning: Extended Abstract. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 3 pages.

1 INTRODUCTION

Multi-agent systems (MAS) [9] are prevalent in real-life scenarios involving a large number of players, such as flocking [7, 21], traffic flow [3], and swarm robotics [6], among others. The study of MAS has garnered significant attention throughout history. As the number of players increases in these multi-agent systems, scalability becomes a challenge [20, 25]. However, under symmetry and homogeneity assumptions, mean field approximations offer an effective approach for modeling population behaviors and learning decentralized policies that do not suffer from issues of the curse of dimensionality and non-stationarity.

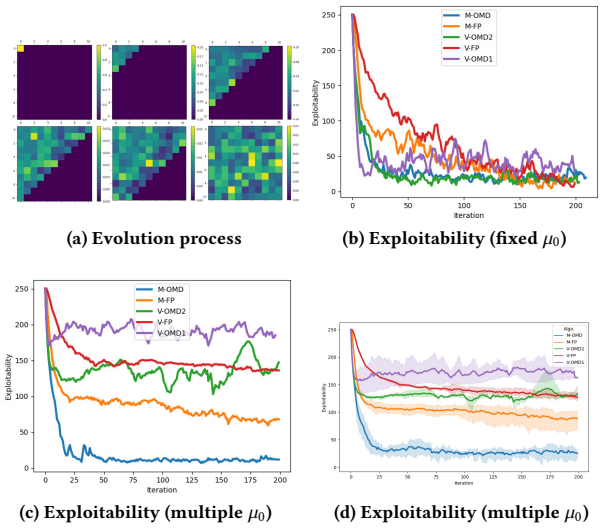
Mean field games (MFGs) [2, 5, 15, 16] provide a framework for large-population games where agents are identical in their behaviors (policy) and only interact through the distribution of all agents. This implies that, as the number of agents grows, the influence of an individual agent becomes negligible, reducing the interactions among agents to that between a representative individual and the population distribution. The main solution concept in MFGs corresponds to a Nash equilibrium, which represents the situation where no player has an incentive to deviate from its current policy unilaterally. Recently, several learning methods have been proposed to solve MFGs; see e.g. [17] for a survey. The most basic one relies on fixed point iterations, which amounts to iteratively updating the policy of a player and the mean field (MF). However, convergence of Banach-Picard fixed point iterations relies on a strict contraction condition [11, 19]. This condition necessitates Lipschitz continuity with sufficiently small Lipschitz constants, which often fails to hold [1, 8].

To address this limitation, several approaches have been proposed, usually based on some form of smoothing. Fictitious play (FP) [4, 14, 24] and Online Mirror Descent (OMD) [12, 13, 22, 23] are two effective strategies for learning equilibria in MFGs. However, FP requires storing all historical best responses and sampling from the best response pool during execution, while OMD requires averaging historical Q functions which is not feasible for neural networks. Moreover, the existing literature often assumes that agents always start from a fixed initial distribution.

2 ALGORITHM

In this paper, we propose a deep reinforcement learning (DRL) algorithm that achieves population-dependent Nash equilibrium without the need for averaging or sampling from history, inspired by Munchausen RL and OMD. Instead of keeping copies of history neural networks and summing the outputs in previous OMD-based algorithm [22], the regularized Q-function defined in this paper can mimic the summation $\sum_{i=0}^{k-1} Q^i$ by using implicit regularization thanks to a Kullback-Leibler (KL) divergence between the new policy and the previous one. We derive, in our MF context, the equivalence between regularized Q and cumulative Q values.

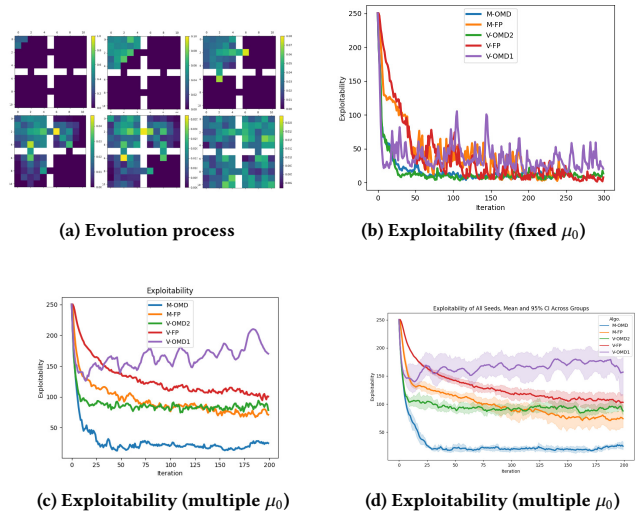
Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).



(a) Evolution process (b) Exploitability (fixed μ_0)

(c) Exploitability (multiple μ_0) (d) Exploitability (multiple μ_0)

Figure 1: Example 1: Exploration in one room. (a): density evolution using the policy learnt by M-OMD, starting from the μ_0 used for (b). (b): exploitability vs training iteration for a single μ_0. (c): average exploitability when training over 5 different μ_0 (single run of each algo.). (d): averaged curve over 5 runs and std dev.



(a) Evolution process (b) Exploitability (fixed μ_0)

(c) Exploitability (multiple μ_0) (d) Exploitability (multiple μ_0)

Figure 2: Example 2: Exploration in four connected rooms. (a): density evolution using the policy learnt by M-OMD, starting from the μ_0 used for (b). (b): exploitability vs training iteration for a single μ_0. (c): average exploitability when training over 5 different μ_0 (single run of each algo.). (d): average over 5 runs & std dev.

In addition, through the design of an additional inner-loop replay buffer, the agents can effectively learn to achieve Nash equilibrium from any distribution, mitigating catastrophic forgetting.

3 EXPERIMENT

Environments. Exploration is a canonical problem in MFG [10], in which a large group of agents tries to uniformly distribute into empty areas but in a decentralized way. In this section, we introduce two variants, with different geometries of domain. The first is in a big empty room, and the second is in four connected rooms, which makes the problem much more challenging. In each experiment, we explore two different scenarios respectively. The first scenario, referred to as **fixed** μ_0 in the sequel, where the population always starts from a fixed initial distribution. The second scenario, referred to as **multiple** μ_0 aims to examine the effectiveness of the master policy. In this scenario, we set different initial distributions simultaneously used for training. Instead of training multiple Nash equilibria with different networks, the master policy aims to use one single network to learn the equilibrium policies for different initial distributions. Intuitively, population-independent policies cannot perform well in this scenario (unless the equilibrium policy does not vary when the initial distribution changes, which amounts to saying that there are no interactions).

Baselines. We compare our algorithm with 4 baselines, including several SOTA algorithms in the domain of Deep RL for MFGs. In the figures, **vanilla FP (V-FP)** refers to an adaptation of (tabular) FP from [24] to deep neural networks. V-FP uses classic fictitious play to iteratively learn the Nash equilibrium, implicitly assuming agents always start from a fixed distribution. **Master FP (M-FP)** is the population-dependent FP from [23], which aimed to handle any initial distribution via FP. **Vanilla OMD1 (V-OMD1)** is the Deep OMD introduced in [18] based on Munchausen trick. **Vanilla OMD2 (V-OMD2)** is our algorithm *without* the input of MF state, while our full algorithm is called **Master OMD (M-OMD)**. With this terminology, M-FP and M-OMD learn population-dependent policies, while V-FP, V-OMD1 and V-OMD2 do not. V-OMD2 can be viewed as an ablation study of our main algorithm (M-OMD), where we remove the distribution dependence to see the performance.

4 CONLUSION

This paper presents an algorithm called Master OMD (M-OMD) for computing population-dependent Nash equilibria in MFGs, which is more efficient than the SOTA algorithm (M-FP). By extending the Munchausen OMD algorithm to population-aware functions, we propose an effective Q-updating rule that enables the realization of this algorithm. In contrast to stationary MFGs and finite horizon MFGs assuming a fixed initial distribution, our work focuses on models where the initial population is a priori unknown and evolves. Extensive numerical experiments demonstrate clearly the advantages of our proposed M-OMD algorithm over baselines. We leave for future work the theoretical analysis, such as a proof of convergence, perhaps under monotonicity conditions. Furthermore, it would be interesting to extend the algorithm to other settings, such as multi-population MFGs.

ACKNOWLEDGMENTS

M.L. is affiliated with the Shanghai Frontiers Science Center of Artificial Intelligence and Deep Learning.

REFERENCES

[1] Berkay Anahtarci, Can Deha Kariksiz, and Naci Saldi. 2023. Q-learning in regularized mean-field games. *Dynamic Games and Applications* 13, 1 (2023), 89–117.

[2] Alain Bensoussan, Jens Frehse, and Phillip Yam. 2013. *Mean field games and mean field type control theory*. Vol. 101. Springer.

[3] Birgit Burmeister, Afsaneh Haddadi, and Guido Matylis. 1997. Application of multi-agent systems in traffic and transportation. *IEE Proceedings-Software* 144, 1 (1997), 51–60.

[4] Pierre Cardaliaguet and Saeed Hadikhanloo. 2017. Learning in mean field games: the fictitious play. *ESAIM: Control, Optimisation and Calculus of Variations* 23, 2 (2017), 569–591.

[5] René Carmona and François Delarue. 2018. *Probabilistic theory of mean field games with applications I-II*. Springer.

[6] Soon-Jo Chung, Aditya Avinash Paranjape, Philip Dames, Shaojie Shen, and Vijay Kumar. 2018. A survey on aerial swarm robotics. *IEEE Transactions on Robotics* 34, 4 (2018), 837–855.

[7] Felipe Cucker and Steve Smale. 2007. Emergent behavior in flocks. *IEEE Transactions on automatic control* 52, 5 (2007), 852–862.

[8] Kai Cui and Heinz Koeppl. 2021. Approximately solving mean field games via entropy-regularized deep reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 1909–1917.

[9] Ali Dorri, Salil S Kanhere, and Raja Jurdak. 2018. Multi-agent systems: A survey. *Ieee Access* 6 (2018), 28573–28593.

[10] Matthieu Geist, Julien Pérolat, Mathieu Laurière, Romuald Elie, Sarah Perrin, Olivier Bachem, Rémi Munos, and Olivier Pietquin. 2021. Concave utility reinforcement learning: the mean-field game viewpoint. *arXiv preprint arXiv:2106.03787* (2021).

[11] Xin Guo, Anran Hu, Renyuan Xu, and Junzi Zhang. 2019. Learning mean-field games. *Advances in Neural Information Processing Systems* 32 (2019).

[12] Saeed Hadikhanloo. 2017. Learning in anonymous nonatomic games with applications to first-order mean field games. *arXiv preprint arXiv:1704.00378* (2017).

[13] Saeed Hadikhanloo. 2018. *Learning in mean field games*. Ph.D. Dissertation. Université Paris sciences et lettres.

[14] Saeed Hadikhanloo and Francisco J Silva. 2019. Finite mean field games: fictitious play and convergence to a first order continuous mean field game. *Journal de Mathématiques Pures et Appliquées* 132 (2019), 369–397.

[15] Minyi Huang, Peter E Caines, and Roland P Malhamé. 2007. Large-population cost-coupled LQG problems with nonuniform agents: individual-mass behavior and decentralized ε-Nash equilibria. *IEEE transactions on automatic control* 52, 9 (2007), 1560–1571.

[16] Jean-Michel Lasry and Pierre-Louis Lions. 2007. Mean field games. *Japanese journal of mathematics* 2, 1 (2007), 229–260.

[17] Mathieu Laurière, Sarah Perrin, Matthieu Geist, and Olivier Pietquin. 2022. Learning mean field games: A survey. *arXiv preprint arXiv:2205.12944* (2022).

[18] Mathieu Laurière, Sarah Perrin, Sertan Girgin, Paul Muller, Ayush Jain, Theophile Cabannes, Georgios Piliouras, Julien Pérolat, Romuald Élie, and Olivier Pietquin. 2022. Scalable deep reinforcement learning algorithms for mean field games. In *International Conference on Machine Learning*. PMLR, 12078–12095.

[19] Minne Li, Zhiwei Qin, Yan Jiao, Yaodong Yang, Jun Wang, Chenxi Wang, Guobin Wu, and Jieping Ye. 2019. Efficient ridesharing order dispatching with mean field multi-agent reinforcement learning. In *The world wide web conference*. 983–994.

[20] Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems* 30 (2017).

[21] Reza Olfati-Saber. 2006. Flocking for multi-agent dynamic systems: Algorithms and theory. *IEEE Transactions on automatic control* 51, 3 (2006), 401–420.

[22] Julien Perolat, Sarah Perrin, Romuald Elie, Mathieu Laurière, Georgios Piliouras, Matthieu Geist, Karl Tuyls, and Olivier Pietquin. 2022. Scaling Mean Field Games by Online Mirror Descent. *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems* (2022), 1028–1037.

[23] Sarah Perrin, Mathieu Laurière, Julien Pérolat, Romuald Élie, Matthieu Geist, and Olivier Pietquin. 2022. Generalization in mean field games by learning master policies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 9413–9421.

[24] Sarah Perrin, Julien Pérolat, Mathieu Laurière, Matthieu Geist, Romuald Elie, and Olivier Pietquin. 2020. Fictitious play for mean field games: Continuous time analysis and applications. *Advances in Neural Information Processing Systems* 33 (2020), 13199–13213.

[25] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder De Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. 2020. Monotonic value function factorisation for deep multi-agent reinforcement learning. *The Journal of Machine Learning Research* 21, 1 (2020), 7234–7284.