Generative-Discriminative Mean Field Distribution Approximation in Multi-agent Reinforcement Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

Non-cooperative and cooperative games with a very large number of players remain generally intractable when the number of players increases. Introduced by Lasry and Lions (2007) and Huang et al. (2006), Mean Field Games (MFGs) rely on a mean-field approximation to allow the number of players to grow to infinity. In Mean-field reinforcement learning, When the state space is finite but very large, storing the population distribution in a tabular way for every state and computing the evolution of this distribution in an exact way is prohibitive in terms of memory and computational time. In continuous spaces, representing and updating the distribution is even more challenging, even if it is just for the purpose of implementing the RL environment and not to use it as an input to the policies. In this case, one needs to rely on approximations. This research aims to propose a model-based reinforcement learning algorithm, GD-MFRL that efficiently represents the distribution using function approximation in a two-part generative and discriminative setting; (i) one part learns to generate distributions by trial and error, and (ii) the other part tries to evaluate these distributions. The definition of such a framework requires answering several challenging research questions, including: How to evaluate the transfer quality in a Multiagent scenario?

1 Introduction

In Observing the mean field, we assume that the agent does not observe the distribution, or at least does not exploit this information to learn the equilibrium policy. Although this is the most common approach in the RL and MFGs literature, the question of learning population-dependent policies arises quite naturally since one could expect that agents learn how to react to the current distribution they observe. This is usual in MARL, see e.g. Yang et al. (2018) who consider Q-functions depending on the actions of all the other players. In MFGs, we can expect that, by learning a population-dependent policy, the agent will be able to generalize, i.e., to behave (approximately) optimally even for population configurations that have not been encountered during training. The concept of a value function depending on the population distribution is connected to the so-called Master equation in MFGs. Introduced by Lions (2012) in continuous MFGs (continuous time, continuous state, and action spaces), this partial differential equation (PDE) corresponds to the limit of systems of Hamilton-Jacobi-Bellman PDEs characterizing Nash equilibria in symmetric N-player games. We refer the interested reader to e.g. Bensoussan et al. (2015); and Cardaliaguet et al. (2019) for more details on this topic. With this approach, value functions and policies take as input a distribution, which is a high-dimensional object. As a consequence, they are much more challenging to approximate than population-independent policies. Perrin et al. (2022), introduced the concept of

Submitted to 38th Workshop on Aligning Reinforcement Learning Experimentalists and Theorists (ARLET 2024). Do not distribute.

master policies, which are population-dependent policies allowing to recover an equilibrium policy for any observed population distribution. They proposed to approximately compute master policies by a combination of Fictitious play, DRL, and suitable randomization of the initial distribution. Wu et al. (2024) extended the approach to non-stationary master policies and proposed an adaptation of the Munchausen OMD algorithm introduced by Lauriere et al., 2022 to compute policies taking the distribution as an input.

In distribution estimation, When the state space is finite but very large, storing the population distribution in a tabular way for every state and computing the evolution of this distribution in an exact way is prohibitive in terms of memory and computational time. In continuous spaces, representing and updating the distribution is even more challenging, even if it is just for the purpose of implementing the RL environment and not to use it as an input to the policies. In this case, one needs to rely on approximations. As already mentioned, a possible method consists of using an empirical distribution, whose evolution can be implemented by Monte Carlo samples of an interacting agent system. This amounts to using a finite population of agents to simulate the environment. For example, in linear-quadratic MFGs the interactions are only through the mean, which can be estimated even using a single agent, see Angiuli et al., 2022c,b in the stationary setting and Angiuli et al., 2021; uz Zaman et al., 2020; Miehling et al., 2022; uz Zaman et al., 2023a in the finite-horizon setting. However, it should be noted that even if a finite number of agents is used in the environment, this approach does not directly reduce the problem to a MARL problem because the goal is still to learn the equilibrium policy for the MFG instead of the finite-agent equilibrium policy.

Another approach consists of representing efficiently the distribution using function approximation. This raises the question of the choice of parameterization and the training method for the parameters. This approach can be implemented in a model-free way using Monte Carlo samples, which is particularly suitable for spaces that are too large to be explored in an exhaustive fashion.

2 Research Goals and Expected Contributions

This research aims to propose a model-based RL algorithm to allow distribution approximation in multi-agent reinforcement Learning, in a generative-discriminative setting. Specifying such a method requires the definition of (i) A model that learns the distributions and tries to consistently generate approximations; (ii) The second discriminative part that tries to understand and evaluate these approximations; and (iii) How to define knowledge interaction framework between the generator and discriminator. The agent extracts knowledge from trial and error and previously solved tasks to accelerate the learning of the distribution. The learning of this distribution can then be abstracted and added to the knowledge base.

3 Background and Related Work

3.1 Mean Field Games

An MFG describes a game for a continuum of identical agents and is fully characterized by the dynamics and the payoff function of a representative agent. More precisely, denoting by μ_t the state distribution of the population, and by $\xi_t \in \mathbb{R}^l$ and $\alpha_t \in \mathbb{R}^k$ the state and the control of an infinitesimal agent, the dynamics of the infinitesimal agent is given by

$$\xi_{t+1} = \xi_t + b(\xi_t, \mu_t, \alpha_t) + \sigma\epsilon_{t+1}(1) \tag{1}$$

where $b : \mathbf{R}^l \times \mathbf{R}^k \times \rho\mathbf{R}^l\mathbf{R}^l$ is a drift (or transition) function, σ is a $l \times l$ matrix and ϵ_{t+1} is a noise term taking values in $\mathbf{R}^l$. We assume that the sequence of noises $\epsilon_{t(}t \geq 0)$ is i.i.d. (e.g. Gaussian). The objective of each infinitesimal agent is to maximize its total expected payoff, given a flow of distributions $\mu = \mu_t(t \geq 0)$ and a strategy α (i.e., a stochastic process adapted to the filtration generated by $\epsilon_{t(}t \geq 0)$ as: $J_\mu(\alpha) = \mathbf{E}_{\xi t \alpha t}[\Sigma_t \geq 0\gamma^t\phi(\xi_t, \mu_t, \alpha_t)]$, where $\gamma \in (0, 1)$ is a discount factor and $\phi\mathbf{R}^l \times \mathbf{R}^k \times \rho\mathbf{R}^l\mathbf{R}^l$ is an instantaneous payoff function. Since this payoff depends on the population's state distribution, and since the other agents would also aim to maximize their payoff, a natural approach is to generalize the notion of Nash equilibrium to this framework. A mean-field (Nash) equilibrium is defined as a pair $(\hat{\mu}, \hat{\alpha} = (\hat{\mu_t}, \alpha_t)_{(}\hat{t} \geq 0)$ of a flow of distributions

and strategies such that the following two conditions are satisfied: $\hat{\alpha}$ is the best response against $\hat{\mu}$ (optimality) and $\hat{\mu}$ is the distribution generated by $\hat{\alpha}$ (consistency), i.e.,

1. $\hat{\alpha}$ maximizes $\alpha J_{\hat{\mu}(\alpha)}$; 2. for every $t \geq 0$, $\hat{\mu_t}$ is the distribution of ξ_t when it follows the dynamics (1) with (α_t, μ_t) replaced by $\alpha_t\hat{)}, \hat{\mu_t}$.

Finding a mean field equilibrium thus amounts to finding a fixed point in the space of (flows of) probability distributions. The existence of equilibria can be proven through classical fixed point theorems (Carmona and Delarue, 2018). In most mean field games considered in the literature, the equilibrium is unique, which can be proved using either a strict contraction argument or the so-called Lasry-Lions monotonicity condition (Lasry and Lions, 2007). Computing solutions to MFGs is a challenging task, even when the state is in a small dimension, due to the coupling between the optimality and the consistency conditions. This coupling typically implies that one needs to solve a forward-backward system where the forward equation describes the evolution of the distribution and the backward equation characterizes the optimal control. One can not be solved prior to the other one, which leads to numerical difficulties.

3.2 Reinforcement Learning

The Reinforcement Learning (RL) paradigm is the machine learning answer to the optimal control problem. It aims at learning an optimal policy for an agent that interacts in an environment composed of states, by performing actions. Formally, the problem is framed under the Markov Decision Processes (MDP) framework. An MDP is a tuple (S, A, p, r, γ) where S is a state space, A is an action space, $p : S \times AP(S)$ is a transition kernel, $r : S \times AR$ is a reward function and γ is a discount factor (see Eq. (2)). Using action a when the current state is s leads to a new state distributed according to P(s, a) and produces a reward R(s, a). A policy $\pi : SP(A), s\pi(|s)$ provides a distribution over actions for each state RL aims at learning a policy π^* which maximizes the total return defined as the expected (Discounted) the sum of future rewards:

$$R(\pi) = E_a t,_s \, t_{+1}[\Sigma_(t \geq 0)\gamma^t(s_t, a_t)](2) \tag{2}$$

with $a_t \sim \pi(|s_t)$ and $s_{t+1}p(|s_t, a_t)$. Note that if the dynamics (p and r) are known to the agent, the problem can be solved using e.g. dynamic programming. Most of the time, these quantities are unknown and RL is required. A plethora of algorithms exist to address the RL problem. Yet, we need to focus on methods that allow continuous action spaces as we want to control accelerations. One category of such algorithms is based on the Policy Gradient (PG) theorem (Sutton et al., 1999) and makes use of the gradient ascent principle: $\pi\pi + \alpha(\pi)/\partial\pi$, where α is a learning rate. Yet, PG methods are known to be high-variance because they use Monte Carlo rollouts to estimate the gradient. A vast literature thus addresses the variance reduction problem. Most of the time, it involves a hybrid architecture, namely Actor-Critic, which relies on both a representation of the policy and of the so-called state-action value function $(s, a)Q^\pi(s, a)$. $Q^\pi(s, a)$ is the total return conditioned on starting in state s and using action a before using policy π for subsequent time steps. It can be estimated by bootstrapping, using the Markov property, through the Bellman equations. Most recent implementations rely on deep neural networks to approximate π and Q (e.g. (Haarnoja et al., 2018)).

4 Partial Results

In order to define a representation that allows distribution approximation, we propose a GD-MFRL extension to MAS, called Generative Discriminative Mean-Field Reinforcement Learning. GD-MFRL is inspired by the insight that approximation can be seen and modeled as a game in the MAS; hence, the environment is described by a set of a generator and a discriminator, in which the former can generate approximations and the latter enhances these approximations. While GD-MFRL enables distribution approximation by trial and error, eliminating the need for recalibration to have neural networks satisfying that the model is calibrated w.r.t f.

5 Next Steps

GD-MFRL is a promising model that allows approximation of the distribution. Now, the next step in our research is to define how the method will iteratively adjust the interactions between the generative and discriminative components until they reach a consensus on a value that accurately reflects reality and aligns with their initial beliefs. Abstract policies have been successfully used, thus we now plan to build abstract policies based on GD-MFRL. We still need to specify a mapping method to find correspondences between generators and discriminators in different domains, and how the transfer of knowledge among agents may be executed with abstract policies.

References

[1] Lasry, J.-M. and Lions, P.-L. (2007). Mean field games. Japanese Journal of Mathematics, 2(1).

[2] Huang, M., Malhamé, R. P., and Caines, P. E. (2006). Large population stochastic dynamic games: closed-loop McKean-Vlasov systems and the Nash certainty equivalence principle. Communications in Information and Systems, 6(3):221–252.

[3] Yang, Y., Luo, R., Li, M., Zhou, M., Zhang, W., and Wang, J. (2018b). Mean field multi-agent reinforcement learning. In International conference on machine learning, pages 5571–5580. PMLR.

[4] Lions, P.-L. (2006-2012). Lecture at the Collège de France.

[5] Bensoussan, A., Frehse, J., and Yam, S. C. P. (2015). The master equation in mean field theory. Journal de Mathématiques Pures et Appliquées, 103(6):1441–1474.

[6] Cardaliaguet, P., Delarue, F., Lasry, J.-M., and Lions, P. L. (2019). The master equation and the convergence problem in mean field games. Princeton University Press.

[7] Perrin, S., Laurière, M., Pérolat, J., Élie, R., Geist, M., and Pietquin, O. (2022). Generalization in mean field games by learning master policies. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, pages 9413–9421.

[8] Wu, Z., Laurière, M., Chua, S. J. C., Geist, M., Pietquin, O., and Mehta, A. (2024). Population-aware online mirror descent for mean-field games by deep reinforcement learning.

[9] Lauriere, M., Perrin, S., Girgin, S., Muller, P., Jain, A., Cabannes, T., Piliouras, G., Pérolat, J., Elie, R., Pietquin, O., et al. (2022). Scalable deep reinforcement learning algorithms for mean field games. pages 12078–12095.

[10] Angiuli, A., Fouque, J.-P., and Laurière, M. (2022c). Unified reinforcement Q-learning for mean field game and control problems. Mathematics of Control, Signals, and Systems, pages 1–55.

[11] Angiuli, A., Fouque, J.-P., and Lauriere, M. (2021). Reinforcement learning for mean field games, with applications to economics. To appear in Machine Learning And Data Sciences For Financial Markets (arXiv preprint arXiv:2106.13755).

[12] uz Zaman, M. A., Zhang, K., Miehling, E., and Bas, ar, T. (2020). Reinforcement learning in non-stationary discretetime linear-quadratic mean-field games. In 2020 59th IEEE Conference on Decision and Control (CDC), pages 2278–2284. IEEE.

[13] Miehling, E., Ba¸sar, T., et al. (2022). Reinforcement learning for non-stationary discrete-time linear–quadratic mean-field games in multiple populations. Dynamic Games and Applications, pages 1–47.

[14] uz Zaman, M. A., Koppel, A., Bhatt, S., and Basar, T. (2023a). Oracle-free reinforcement learning in mean-field games along a single sample path. In International Conference on Artificial Intelligence and Statistics, pages 10178–10206. PMLR.

[15] Rene Carmona and Franc¸ois ´ Delarue. Probabilistic theory of mean field games with applications. I. 2018.

[16] Jean-Michel Lasry and PierreLouis Lions. Mean field games. Jpn. J. Math., 2007.

[17] R. S. Sutton, D. Mcallester, S. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In proc. of NeurIPS. MIT Press, 1999.

[18] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. CoRR, abs/1801.01290, 2018.