Learning While Playing in Mean-Field Games: Convergence and Optimality

Qiaomin Xie [1] **Zhuoran Yang** [2] **Zhaoran Wang** [3] **Andreea Minca** [1]

Abstract

We study reinforcement learning in mean-field games. To achieve the Nash equilibrium, which consists of a policy and a mean-field state, existing algorithms require obtaining the optimal policy while fixing any mean-field state. In practice, however, the policy and the mean-field state evolve simultaneously, as each agent is learning while playing. To bridge such a gap, we propose a fictitious play algorithm, which alternatively updates the policy (learning) and the mean-field state (playing) by one step of policy optimization and gradient descent, respectively. Despite the nonstationarity induced by such an alternating scheme, we prove that the proposed algorithm converges to the Nash equilibrium with an explicit convergence rate. To the best of our knowledge, it is the first provably efficient algorithm that achieves learning while playing.

1. Introduction

Multi-agent reinforcement learning (MARL) (Shoham et al., 2007; Busoniu et al., 2008; Hernandez-Leal et al., 2017; 2018; Zhang et al., 2019) aims to tackle sequential decision-making problems in multi-agent systems by integrating the classical reinforcement learning framework with game-theoretical thinking (Başar & Olsder, 1998). Powered by deep learning, MARL recently has achieved striking empirical successes in games (Silver et al., 2016; 2017; Vinyals et al., 2019; Berner et al., 2019; Schrittwieser et al., 2019), robotics (Yang & Gu, 2004; Busoniu et al., 2006; Leottau et al., 2018), transportation (Kuyer et al., 2008; Mannion et al., 2016), and social science (Leibo et al., 2017; Jaques et al., 2019; Cao et al., 2018; McKee et al., 2020).

Despite the empirical successes, MARL is known to suffer from the scalability issue. Specifically, in a multi-agent system, each agent interacts with the environment *and* the other agents, with the goal of maximizing its own expected total return. Consequently, for each agent, its reward and state transition also depend on the states and actions of all the other agents. As the number of agents increases, the size of the joint state-action space grows exponentially, leading to severe computational and statistical challenges for reinforcement learning algorithms. Such challenges are sometimes referred to as the "curse of many agents" (Sonu et al., 2017).

To circumvent these challenges, a popular approach is through mean-field approximation, which imposes symmetry among the agents and specifies that, for each agent, the joint effect of all the other agents is summarized by a population quantity, which is oftentimes given by the empirical distribution $\mathcal{L}$ of the local states and actions of all the other agents, or by a functional of this empirical distribution. Under symmetry, the reward function and local state transition function are the same for each agent, and they are both functions of the local state-action and the population quantity. Under the mean-field approximation, a multi-agent system is modeled as a mean-field game (MFG) (Huang et al., 2003; Lasry & Lions, 2006a;b; 2007; Huang et al., 2007; Guéant et al., 2011; Carmona & Delarue, 2018), which is readily scalable to an arbitrary number of agents.

In this work, we aim to find the Nash equilibrium (Nash, 1950) of MFG with an infinite number of agents via reinforcement learning. An MFG consists of a population of symmetric agents, each of which has an infinitesimal effect over the whole population. By symmetry, it suffices to find a *symmetric* Nash equilibrium, where each agent adopts the same policy. Therefore, we can focus on a single agent, known as the representative agent, and view MFG as a game between the representative agent's local policy π and the mean-field state $\mathcal{L}$, where $\mathcal{L}$ aggregates the collective effect of the population. Specifically, the representative agent aims to find the optimal policy π when the mean-field state is fixed to $\mathcal{L}$, which reduces to solving a Markov decision process (MDP) induced by $\mathcal{L}$. Simultaneously, the mean-field state $\mathcal{L}$ evolves according to the transition kernel when all the agents adopt policy π. The Nash equilibrium of this two-player game, $(\pi^*, \mathcal{L}^*)$, corresponds to a symmetric Nash equilibrium π^* of the original MFG.

[1]School of Operations Research and Information Engineering, Cornell University [2]Department of Operations Research and Financial Engineering, Princeton University [3]Department of Industrial Engineering and Management Sciences, Northwestern University. Correspondence to: Qiaomin Xie <qiaomin.xie@cornell.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

In principle, the Nash equilibrium $(\pi^*, \mathcal{L}^*)$ can be obtained via fixed-point iteration, which generate a sequence $\{(\pi_t, \mathcal{L}_t)\}_{t\geq 0}$ as follows. In the t-th iteration, one solves for the optimal policy π_t of the MDP induced by $\mathcal{L}_t$. Then $\mathcal{L}_{t+1}$ is computed as the next mean-field state resulting from every agent following the policy π_t at the current mean-field state $\mathcal{L}_t$. Under appropriate assumptions, the mapping from $\mathcal{L}_t$ to $\mathcal{L}_{t+1}$ is a contraction (Saldi et al., 2018b; Anahtarci et al., 2020b; 2019), hence the above iterative algorithm converges to the unique fixed-point $\mathcal{L}^*$ of this mapping. Various reinforcement learning methods are proposed to approximately implement the fixed-point iteration to find the Nash equilibrium; see e.g. Guo et al. (2019; 2020); Anahtarci et al. (2019).

The above approach typically leads to an algorithm with a *double-loop*: each iteration requires (approximately) solving a standard reinforcement learning problem, namely learning the optimal policy of the MDP induced by the current mean-field state $\mathcal{L}_t$; this sub-problem itself is often solved by an iterative algorithm such as Q-learning (Watkins & Dayan, 1992; Mnih et al., 2015; Bellemare et al., 2017) or actor-critic methods (Konda & Tsitsiklis, 2000; Haarnoja et al., 2018; Schulman et al., 2015; 2017). A major challenge here is that the inner loop requires fixing the mean-field state at $\mathcal{L}_t$, which is difficult to implement since the mean-field state evolves simultaneously as agents play and updates their policy. Moreover, when the state space is enormous, function approximation tools such as deep neural networks are equipped to represent the value and policy functions in the reinforcement learning algorithm, making solving each inner subproblem computationally demanding.

To develop a practical and computationally efficient algorithm for MFG, we seek to answer the following question:

Can we design an online reinforcement learning algorithm for solving MFG which updates the policy and mean-field state simultaneously in each iteration?

In particular, such an algorithm does not require fixing the mean-field state across iteration. We provide an affirmative answer to this question by proposing a fictitious-play style policy optimization algorithm, where the policy π and mean-field state $\mathcal{L}$ are viewed as two players and updated simultaneously. Fictitious play (Brown, 1951) is a general algorithm framework for solving games, where each player first infers the opponent's strategy based on past behaviors and then improves its own policy accordingly. In the context of MFG, in each iteration, the policy player π first infers the mean-field state implicitly by performing policy evaluation of π under the MDP induced by the current $\mathcal{L}$. Then the policy π is updated via a *single* step of proximal policy optimization (PPO) (Schulman et al., 2017), with entropy regularization to ensure the uniqueness of the Nash equilibrium. Meanwhile, the update direction of the mean-field state $\mathcal{L}$ is given by how the state distribution of the agents evolves when the agents collectively execute policy π, and $\mathcal{L}$ is updated towards this direction with some stepsize. This algorithm is single-loop and online, as the mean-field state $\mathcal{L}$ is updated immediately when π is played and updated.

When the stepsizes for policy and mean-field state updates are properly chosen, we prove that our algorithm converges to the entropy-regularized Nash equilibrium at an $\widetilde{\mathcal{O}}(T^{-1/5})$ rate, where T is the total number of iterations and $\widetilde{\mathcal{O}}(\cdot)$ hides logarithmic terms. To our best knowledge, this is the first single-loop, online reinforcement learning algorithm for mean-field games with finite-time convergence guarantee.

We remark that the mean-field $\mathcal{L}$ is a distribution over the state space $\mathcal{S}$; when $\mathcal{S}$ is continuous, $\mathcal{L}$ is an infinite-dimensional object, making it computationally challenging to store and manipulate $\mathcal{L}$. To overcome this challenge, our algorithm can be optionally coupled with the kernel mean embedding approach (Smola et al., 2007; Gretton et al., 2006; Sriperumbudur et al., 2010), which provides a succinct representation of $\mathcal{L}$ by mapping it to an element in a reproducing kernel Hilbert space (RKHS). This approach allows for the flexibility of choosing the reproducing kernel appropriately for computationally efficient representation of the mean-field state.

Our Contributions. The contributions of this paper are two-fold. Algorithmically, we propose a single-loop fictitious-play style algorithm that updates both the policy and the mean-field state simultaneously in each iteration, where the policy is updated via entropy-regularized proximal policy optimization. Kernel mean embedding can be incorporated to represent the mean-field states, and the policy update subroutine can readily employ any function approximation schemes for efficient representation of the value and policy functions, which makes our method a general algorithmic framework for learning MFG with continuous state space. Theoretically, we establish rigorous guarantees that the policy and mean-field state sequence generated by the proposed algorithm converges to the Nash equilibrium of the MFG at an explicit $\widetilde{\mathcal{O}}(T^{-1/5})$ rate.

Related Works. Our work belongs to the literature on discrete-time MFG. A variety of works have focused on the existence of a Nash equilibrium and the behavior of Nash equilibrium as the number of agents goes to infinity under various settings of MFG. See, e.g., Gomes et al. (2010); Tembine & Huang (2011); Moon & Başar (2014); Biswas (2015); Saldi et al. (2018b;a; 2019); Więcek (2020) and the references therein. Our work is closely related to the line of research that aims to solve MFG via reinforcement learning methods. Most of the existing works propose to find the Nash equilibrium via fixed-point iterations in space of the mean-field states, which requires solving an MDP

induced by a mean-field state within each iteration (Guo et al., 2019; 2020; Anahtarci et al., 2019; 2020b; Fu et al., 2020; uz Zaman et al., 2020; Anahtarci et al., 2019). Among these works, Guo et al. (2019; 2020); Anahtarci et al. (2019; 2020b) propose to solve each MDP via Q-learning (Watkins & Dayan, 1992) or approximated value iteration (Munos & Szepesvári, 2008), whereas Fu et al. (2020); uz Zaman et al. (2020) solve each MDP using actor-critic (Konda & Tsitsiklis, 2000) under the linear-quadratic setting.

Most related to our work are Elie et al. (2020); Perrin et al. (2020), which study the convergence of a version of fictitious play for MFG. Similar to our algorithm, their fictitious play also regards the policy and the mean-field state as the two players. However, for policy update, they compute the best response policy to the current mean-field state by solving, to near optimality, the MDP induced by the mean-field state, and the obtained policy is added to the set of previous policy iterates to form a mixture policy. As a result, their algorithm is double-loop in essence due to solving an MDP in each iteration. In contrast, our algorithm is single-loop where each iteration involves policy evaluation rather than finding the optimaly policy; in particular, the policy is updated via a single PPO step in each iteration, and the mean-field state is updated before the policy converges to the optimum of the MDP associated with the current mean-field state. We remark that some recent works also consider single-loop algorithms. In particular, Subramanian & Mahajan (2019) propose a policy-gradient based approach to update policy; however, only asymptotic convergence guarantee is established via two-timescale stochastic approximation. The algorithm introduced by Angiuli et al. (2020) updates Q-function in each iteration like Q-learning, and can be applied to learn mean field games and control problems; however, no convergence guarantee is provided.

Notations. We use $\|\cdot\|_1$ to denote the vector ℓ_1-norm, and $\Delta(\mathcal{D})$ the probability simplex over $\mathcal{D}$. The KL divergence between $p_1, p_2 \in \Delta(\mathcal{A})$ is defined as $D_{\mathrm{KL}}(p_1\|p_2) := \sum_{a\in\mathcal{A}} p_1(a)\log\frac{p_1(a)}{p_2(a)}$. Let $\mathbf{1}_n \in \mathbb{R}^n$ denote the all-one vector. For two quantities x and y that may depend on problem parameters ($|\mathcal{A}|, \gamma$, etc.), if $x \geq Cy$ holds for a universals constant $C > 0$, we write $x \gtrsim y$, $x = \Omega(y)$ and $y = \mathcal{O}(x)$. We use $\widetilde{O}(\cdot)$ to denote $\mathcal{O}(\cdot)$ ignoring logarithmic factors.

2. Background and Preliminaries

In this section, we first review the standard setting of mean-field games (MFG), and then introduce a general MFG with mean embedding and entropy regularization.

2.1. Mean-Field Games

Consider a discrete-time Markov game involving an infinite number of identical and interchangeable agents. Let $\mathcal{S} \subseteq \mathbb{R}^d$ and $\mathcal{A} \subseteq \mathbb{R}^p$ be the state space and action space, respectively, that are common to all the agents. We assume that $\mathcal{S}$ is compact and $\mathcal{A}$ is finite. The reward and the state dynamic for each agent depend on the collective behavior of all agents through the mean-field state, i.e., the *distribution* of the states of all agents. As the agents are homogeneous and interchangeable, one can focus on a single representative agent of the population. Let $r : \mathcal{S}\times\mathcal{A}\times\Delta(\mathcal{S}) \to [0, R_{\max}]$ be the (bounded) reward function and P: $\mathcal{S}\times\mathcal{A}\times\Delta(\mathcal{S}) \to \Delta(\mathcal{S})$ be the state transition kernel. At each time t, the representative agent is in state $s_t \in \mathcal{S}$, and the probability distribution of s_t, denoted by $\mathcal{L}_t \in \Delta(\mathcal{S})$, corresponds to the mean-field state. Upon taking an action $a_t \in \mathcal{A}$, the agent receives a reward $r(s_t, a_t, \mathcal{L}_t)$ and transitions to a new state $s_{t+1} \sim \mathrm{P}(\cdot|s_t, a_t, \mathcal{L}_t)$. A Markovian policy for the agent is a function $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ that maps her own state to a distribution over actions,[1] i.e., $\pi(a|s)$ is the probability of taking action a in state s. Let Π be the set of all Markovian policies.

When an agent is operating under a policy $\pi \in \Pi$ and the population distribution flow is $\boldsymbol{\mathcal{L}} := (\mathcal{L}_t)_{t\geq 0}$, we define the expected cumulative discounted reward (or value function) of this agent as $V^\pi(s, \boldsymbol{\mathcal{L}}) := \mathbb{E}\big[\sum_{t=0}^\infty \gamma^t r(s_t, a_t, \mathcal{L}_t) \mid s_0 = s\big]$, where $a_t \sim \pi(\cdot|s_t)$, $s_{t+1} \sim \mathrm{P}(\cdot|s_t, a_t, \mathcal{L}_t)$, and $\gamma \in (0, 1)$ is the discount factor. The goal of this agent is to find a policy π that maximizes $V^\pi(s, \boldsymbol{\mathcal{L}})$ while interacting with the mean-field $\boldsymbol{\mathcal{L}}$.

We are interested in finding a *stationary (time-independent) Nash Equilibrium* (NE) of the game, which is a policy-population pair $(\pi^*, \mathcal{L}^*) \in \Pi \times \Delta(\mathcal{S})$ satisfying the following two properties:

- (Agent rationality) $V^{\pi^*}(s, \mathcal{L}^*) \geq V^\pi(s, \mathcal{L}^*), \forall \pi \in \Pi, s \in \mathcal{S}$.
- (Population consistency) $\mathcal{L}_t = \mathcal{L}^*, \forall t$ under policy π^* with initial mean-field state $\mathcal{L}_0 = \mathcal{L}^*$.

That is, π^* is the optimal policy under the mean-field $\mathcal{L}^*$, and $\mathcal{L}^*$ remains fixed under π^*. We formalize the notion of NE in Section 2.3 after introducing a more general setting of MFG.

2.2. Mean Embedding of Mean-Field States

Note that the mean-field state $\mathcal{L}$ is a distribution over $\mathcal{S}$, i.e., $\mathcal{L} \in \Delta(\mathcal{S})$, where $\Delta(\mathcal{S})$ is a continuous space. To learn the NE for $|\mathcal{S}| < \infty$, prior work (Guo et al., 2019) uses a

[1] In general, the policy may be a function of the mean-field state $\mathcal{L}_t$ as well. We have suppressed this dependency since our ultimate goal is to find a *stationary* equilibrium, under which the mean-field state remains fixed over time. See Guo et al. (2019); Anahtarci et al. (2020a) for a similar treatment.

discretization method that maintains an ϵ-net in $\Delta(\mathcal{S})$ and projects mean-field state updates to the ϵ-net. When the state space is continuous, the NE $(\pi^*, \mathcal{L}^*)$ is an infinite dimensional object, posing challenges for learning and tracking the NE. For instance, the discretization method seems computationally intractable. To overcome this challenge, one can optionally make use of a succinct representation of the mean-field via mean embedding, which embeds the mean-field states into a reproducing kernel Hilbert space (RKHS) (Smola et al., 2007; Gretton et al., 2006; Sriperumbudur et al., 2010).

Specifically, given a positive definite kernel $k : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$, let $\mathcal{H}$ be the associated RKHS endowed with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and norm $\|\cdot\|_{\mathcal{H}}$. For each $\mathcal{L} \in \Delta(\mathcal{S})$, its mean embedding $\mu_{\mathcal{L}} \in \mathcal{H}$ is defined as

$$\mu_{\mathcal{L}}(s) := \mathbb{E}_{x \sim \mathcal{L}} \left[k(x, s) \right], \quad \forall s \in \mathcal{S}.$$

Let $\mathcal{M} := \{\mu_{\mathcal{L}} : \mathcal{L} \in \Delta(\mathcal{S})\} \subseteq \mathcal{H}$ be the set of all possible mean embeddings, which is convex. Note that when k is the identity kernel, we have $\mu_{\mathcal{L}} = \mathcal{L}$ and $\mathcal{M} = \Delta(\mathcal{S})$. Adopting mean-embedding of the state distribution allows one to track the mean-field via sample-based kernel regression (Szabó et al., 2015) when needed. On the other hand, when k is more structured (e.g., with a fast decaying eigen spectrum), $\mathcal{M}$ has significantly lower complexity than the set $\Delta(\mathcal{S})$ of raw mean-field states.

We assume that the MFG respects the mean embedding structure, in the sense that the reward $r : \mathcal{S} \times \mathcal{A} \times \mathcal{M} \to [0, R_{\max}]$ and transition kernel $\mathrm{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{M} \to \Delta(\mathcal{S})$ (with a slight abuse of notation) depend on the mean-field state $\mathcal{L}$ through its mean embedding representation $\mu_{\mathcal{L}}$. In particular, at each time t with state s_t and mean-field state $\mathcal{L}_t$, the representative agent takes action $a_t \sim \pi(\cdot|s_t)$, receives reward $r(s_t, a_t, \mu_{\mathcal{L}_t})$ and then transitions to a new state $s_{t+1} \sim \mathrm{P}(\cdot|s_t, a_t, \mu_{\mathcal{L}_t})$. The NE of the game is defined analogously. As mentioned, when k is the identity kernel, the above setting reduces to the standard setting in Section 2.1 with raw-mean field states.

We impose a standard regularity condition on the kernel k.

Assumption 1. *The MFG respects the mean embedding structure with a kernel $k : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$, which is bounded and universal, in the sense that $k(s, s) \leq 1, \forall s \in \mathcal{S}$ and the corresponding RKHS $\mathcal{H}$ is dense w.r.t. the L_∞ norm in the space of continuous functions on $\mathcal{S}$.*

The boundedness of the kernel in Assumption 1 is standard in the kernel learning literature (Caponnetto & De Vito, 2007; Muandet et al., 2012; Szabó et al., 2015; Lin et al., 2017). When the kernel is bounded, the embedding of each $\mathcal{L} \in \Delta(\mathcal{S})$ satisfies $\|\mu_{\mathcal{L}}\|_{\mathcal{H}} \leq \int_{x \sim \mathcal{L}} \|k(x, \cdot)\|_{\mathcal{H}} \, \mathrm{d}x \leq 1$. When one uses a universal kernel (e.g., Gaussian or Laplace kernel), the mean embedding mapping is injective and hence each embedding $\mu \in \mathcal{M}$ uniquely characterizes a distribution $\mathcal{L}$ in $\Delta(\mathcal{S})$ (Gretton et al., 2006; 2012).

2.3. Entropy Regularization

An entropy regularization approach, which augments the standard expected reward objective with an entropy term of the policy, has been used extensively in MDPs (Geist et al., 2019; Nachum et al., 2017). Recent work has shown that such regularization can accelerate the convergence of policy gradient algorithms (Cen et al., 2020; Shani et al., 2019). In particular, policy gradient algorithms can converge linearly when computing optimal value functions of the regularized MDP—a significant improvement over the non-regularized setting. For mean-field games, recent work by Anahtarci et al. (2020a) shows that with regularization, the NE is unique under quite mild assumptions as opposed to the unregularized case. To ensure the uniqueness of the NE and achieve fast algorithmic convergence, we thus consider entropy regularization. In particular, we define the entropy-regularized value function as

$$V_\mu^{\lambda,\pi}(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t [r(s_t, a_t, \mu) - \lambda \log \pi(a_t|s_t)] | s_0 = s \right],$$

where $a_t \sim \pi(\cdot|s_t), s_{t+1} \sim \mathrm{P}(\cdot|s_t, a_t, \mu)$, the parameter $\lambda > 0$ controls the regularization level and μ is the mean-embedding of some given mean-field state (fixed over time). Equivalently, one may view $V_\mu^{\lambda,\pi}$ as the usual value function of π with an entropy-regularized reward

$$r_\mu^{\lambda,\pi}(s, a) := r(s, a, \mu) - \lambda \log \pi(a|s), \ \forall s, a. \quad (1)$$

Also define the Q-function of a policy π as

$$Q_\mu^{\lambda,\pi}(s, a) = r(s, a, \mu) + \gamma \mathbb{E}\left[V_\mu^{\lambda,\pi}(s_1) \mid s_0 = s, a_0 = a \right],$$

which is related to the value function as

$$\begin{aligned} V_\mu^{\lambda,\pi}(s) &= \mathbb{E}_{a \sim \pi(\cdot|s)} \left[Q_\mu^{\lambda,\pi}(s, a) - \lambda \log \pi(a|s) \right] \\ &= \left\langle Q_\mu^{\lambda,\pi}(s, \cdot), \pi(\cdot|s) \right\rangle + \lambda \mathbb{H}\left(\pi(\cdot|s) \right), \end{aligned} \quad (2)$$

where $\mathbb{H}\left(\pi(\cdot|s)\right) := -\sum_a \pi(a|s) \log \pi(a|s)$ is the Shannon entropy of the distribution $\pi(\cdot|s)$. Since the reward function r is assumed to be $R_{\max}$-bounded, it is easy to show that the Q-function is also bounded as $\left\| Q_\mu^{\lambda,\pi} \right\|_\infty \leq Q_{\max} := (R_{\max} + \gamma\lambda \log |\mathcal{A}|)/(1 - \gamma)$; see Lemma 6.

Single-Agent MDP. When the mean-field state and its mean-embedding remain fixed over time, i.e., $\mathcal{L}_t = \mathcal{L}$ and $\mu_{\mathcal{L}_t} = \mu, \forall t$, a representative agent aims to solve the optimization problem

$$\max_{\pi : \mathcal{S} \to \Delta(\mathcal{A})} V_\mu^{\lambda,\pi}(s) \quad (3)$$

for each $s \in \mathcal{S}$. This problem corresponds to finding the (entropy-regularized) optimal policy for a single-agent discounted MDP, denoted by $\mathrm{MDP}_\mu := (\mathcal{S}, \mathcal{A}, \mathrm{P}(\cdot|\cdot,\cdot,\mu), r(\cdot,\cdot,\mu), \gamma)$, that is induced by $\mu \in \mathcal{M}$. Let $\pi_\mu^{\lambda,*}$ be the optimal solution to the problem (3), that is, the optimal regularized policy of MDP_μ. The optimal policy is unique whenever $\lambda > 0$. One can thus define a mapping $\Gamma_1^\lambda : \mathcal{M} \to \Pi$ via $\Gamma_1^\lambda(\mu) = \pi_\mu^{\lambda,*}$, which maps each embedded mean-field state μ to the optimal regularized policy $\pi_\mu^{\lambda,*}$ of MDP_μ. Let $Q_\mu^{\lambda,*}$ be the optimal regularized Q-function corresponding to the optimal policy $\pi_\mu^{\lambda,*}$.

Throughout the paper, we fix a state distribution $\nu_0 \in \Delta(\mathcal{S})$, which will serve as the initial state of our policy optimization algorithm. For each $\mu \in \mathcal{M}$ and a policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, define

$$J_\mu^\lambda(\pi) := \mathbb{E}_{s\sim\nu_0}\left[V_\mu^{\lambda,\pi}(s)\right] \tag{4}$$

as the expectation of the value function $V_\mu^{\lambda,\pi}(s)$ of policy π on the regularized MDP_μ. We define the discounted state visitation distribution ρ_μ^π induced by a policy π on MDP_μ as:

$$\rho_\mu^\pi(s) := (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{P}(s_t = s), \tag{5}$$

where $\mathbb{P}(s_t = s)$ is the state distribution when $s_0 \sim \nu_0$ and the actions are chosen according to π.

Mean-field Dynamics. When all agents follow the same policy π, we can define another mapping $\Gamma_2 : \Pi \times \mathcal{M} \to \mathcal{M}$ that describes the dynamic of the embedded mean-field state. In particular, given the current embedding μ corresponding to some mean-field state $\mathcal{L}$, the next embedded mean-field state $\mu^+ = \Gamma_2(\pi, \mu)$ is given by

$$\begin{aligned}\mu^+ &= \mu_{\mathcal{L}^+},\\ \mathcal{L}^+(s') &= \int_{\mathcal{S}}\sum_{a\in\mathcal{A}}\mathcal{L}(s)\pi(a|s)\mathrm{P}(s'|s,a,\mu)\mathrm{d}s.\end{aligned}$$

Note that the evolution of the mean-field depends on the agents' policy in a deterministic manner.

Entropy-regularized Mean-field Nash Equilibrium (NE). With the above notations, we can formally define our notion of equilibrium.

Definition 1. *A stationary (time-independent) entropy-regularized Nash equilibrium for the MFG is a policy-population pair $(\pi^*, \mu^*) \in \Pi \times \mathcal{M}$ that satisfies*

$$\begin{aligned}&\textit{(agent rationality)} \quad \pi^* = \Gamma_1^\lambda(\mu^*),\\ &\textit{(population consistency)} \quad \mu^* = \Gamma_2(\pi^*, \mu^*).\end{aligned}$$

When $\lambda = 0$, the above definition reduces to that of the (unregularized) NE discussed in Section 2.1, which requires π^* to the unregularized optimal policy of MDP_{μ^*}. For general values of λ, the regularized NE (π^*, μ^*) approximates the unregularized NE (Geist et al., 2019), in the sense that π^* is an approximate optimal policy of MDP_{μ^*} satisfying

$$\max_{\pi\in\Pi}\{J_{\mu^*}^0(\pi)\} - J_{\mu^*}^\lambda(\pi^*) \le \frac{\lambda\log|\mathcal{A}|}{1-\gamma}. \tag{6}$$

That is, the approximation bias induced by entropy regularization is small with a sufficiently small regularization parameter λ.

One may further define the composite mapping $\Lambda^\lambda : \mathcal{M} \to \mathcal{M}$ as $\Lambda^\lambda(\mu) = \Gamma_2\left(\Gamma_1^\lambda(\mu), \mu\right)$. When Λ^λ is a contraction, the regularized NE exists and is unique (Anahtarci et al., 2020a). Moreover, the iterates $\{(\pi_t, \mu_t)\}_{t\ge0}$ given by the two-step update

$$\pi_t = \Gamma_1^\lambda(\mu_t), \qquad \mu_{t+1} = \Gamma_2(\pi_t, \mu_t)$$

converge to the regularized NE at a linear rate. Note that the first step above requires an oracle for computing the exact optimal policy $\pi_{\mu_t}^{\lambda,*}$. In most cases, such an exact oracle is not available; various single-agent reinforcement learning algorithms have been considered for computing an approximate optimal policy, including Q-learning (Guo et al., 2019) and policy gradient methods (Guo et al., 2020; Subramanian & Mahajan, 2019). The recent work by Elie et al. (2020) considers fictitious play iterative learning scheme. We remark that their convergence guarantee requires being able to compute the approximate optimal policy to an arbitrary precision with high probability.

3. Fictitious Play Algorithm for MFG: Learning While Playing

In this section, we present a fictitious play algorithm, which *learns* the optimal policy π^* of the NE while the mean-field state evolves simultaneously with agents *playing*. As given in Algorithm 1 for a representative agent, each iteration of the algorithm involves three steps: *playing* (line 4), *learning* through policy evaluation (line 6) and policy improvement (line 7). Below we explain each step in more details.

Playing and Mean-field State Evolution. In the t-th iteration, to control the evolution of (embedded) mean-field state towards the Nash equilibrium, the representative agent adopts the following protocol: with probability β_t, she takes an action according to the current policy π_t, observes the reward and the next state accordingly; with probability $1-\beta_t$, she does nothing and remains in the current state. With all agents following this protocol, the underlying embedded mean-field state evolves as follows:

$$\mu_{t+1} = (1-\beta_t)\mu_t + \beta_t \cdot \Gamma_2(\pi_t, \mu_t).$$

Algorithm 1 Mean-Embedded Fictitious Play

1: **Input**: initial policy-state pair (π_0, μ_0), step size sequence $\{\alpha_t, \beta_t\}_{t\geq 0}$, mixing parameter η, entropy regularization parameter λ
2: Sample $s_0 \sim \mu_0$
3: **for** Iteration $t = 0, 1, 2, \ldots, T$ **do**
4: (**Playing**): with probability β_t play action $a_t \sim \pi_t(\cdot|s_t)$, observe reward $r = (s_t, a_t, \mu_t)$ and next state $s_{t+1} \sim \mathrm{P}(\cdot|s_t, a_t, \mu_t)$; and with probability $1-\beta_t$ do nothing, $s_{t+1} = s_t$
5: Environment transitions:

$$\mu_{t+1} = (1-\beta_t)\mu_t + \beta_t \cdot \Gamma_2(\pi_t, \mu_t). \quad (7)$$

6: (**Policy evaluation**): Compute an approximate version $\widehat{Q}_t^\lambda$ of the Q-function $Q_{\mu_t}^{\lambda,\pi_t}$ of policy π_t w.r.t. the regularized MDP_{μ_t}
7: (**Policy improvement**): Update the policy by

$$\widehat{\pi}_{t+1}(\cdot|s_t) \propto (\pi_t(\cdot|s_t))^{1-\alpha_t\lambda} \exp\left(\alpha_t \widehat{Q}_t^\lambda(s_t, \cdot)\right) \quad (8)$$

$$\pi_{t+1}(\cdot|s_t) = (1-\eta)\widehat{\pi}_{t+1}(\cdot|s_t) + \eta \cdot \mathbf{1}_{|\mathcal{A}|}(\cdot)/|\mathcal{A}| \quad (9)$$

8: **end for**

That is, the next (embedded) mean-field state μ_{t+1} is a weighted average of the current μ_t and the mean-field state $\Gamma_2(\pi_t, \mu_t)$ induced by the current policy π_t. This evolution can be viewed as a single step of the (soft) fixed point iteration for the equation $\mu = \Gamma_2(\pi_t, \mu)$, with step size β_t. We remark that unlike prior work that explicitly updates and computes the (artificial) mean-field state, the mean-field state in our algorithm corresponds to the *real* system state, which evolves according to (7) as all agents play according to the described protocol. Therefore, the mean-field state can be estimated accurately from sampled states of a sufficiently large number K of randomly selected agents. In particular, the estimation error for the mean-embedding of the mean-field state decays as $1/\sqrt{K}$ (Muandet et al., 2017).

Policy Evaluation. We next evaluate the current policy π_t with respect to the regularized single-agent MDP_{μ_t} induced by the current embedded mean-field state μ_t. In particular, we compute an approximation $\widehat{Q}_t^\lambda : \mathcal{S} \times \mathcal{A} \to [0, Q_{\max}]$ of the true Q-function $Q_t^\lambda := Q_{\mu_t}^{\lambda,\pi_t}$. Our theorem characterizes how the convergence depends on the policy evaluation error in this step. As the state space is continuous, we can use either function approximation (e.g., linear, RKHS or neural networks (Farahmand et al., 2016; Cai et al., 2019)) or non-parametric method such as k-nearest neighbor (Shah & Xie, 2018).

We remark that the policy evaluation step does not require fixing the underlying mean-field state $\mathcal{L}_t$ to get extra samples from the MDP induced by $\mathcal{L}_t$ (equivalently, MDP_{μ_t}). Specifically, in the t-th iteration, we equip each agent with the same policy π_t. We can uniformly randomly select N agents that actually play, and let s_t^i denote the local state of the i-th agent. Note that $s_t^i \sim \mathcal{L}_t$, and the i-th agent takes action $a_t^i \sim \pi_t(\cdot \mid s_t^i)$, observes a reward $r_t^i = r(s_t^i, a_t^i, \mu_t)$ and the next state $s_{t+1}^i \sim \mathrm{P}(\cdot|s_t^i, a_t^i, \mu_t), \forall i \in [N]$. We thus have N transition data $\{(s_t^i, a_t^i, r_t^i, s_{t+1}^i), i \in [N]\}$ under the policy π_t for MDP_{μ_t}, which can be used for any off-the-shelf policy evaluation solver (e.g., TD(0) or LSTD) to estimate the desired Q-function. For instance, the ℓ_2^2 error of the Q-function under a variant of TD-learning decays at rate $1/\sqrt{N}$ (Cai et al., 2019). Therefore, the policy evaluation step in our algorithm can utilize the feedback information of all agents through a *single* transition of the game under policy π_t. On the other hand, since we only have samples under the current policy π_t, it is difficult (if not infeasible) to solve for the optimal policy of the MDP induced by $\mathcal{L}_t$. This is exactly the advantage of our algorithm, which only requires policy evaluation rather than finding the optimal policy.

Policy Improvement. To update the policy estimate π_t, we first compute an intermediate policy $\widehat{\pi}_{t+1}$ by a *single* policy improvement step: for each $s \in \mathcal{S}$,

$$\widehat{\pi}_{t+1}(\cdot|s) = \underset{\pi(\cdot|s)\in\Delta(\mathcal{A})}{\operatorname{argmax}} \Big\{ \alpha_t \big\langle \widehat{Q}_t^\lambda(s, \cdot) - \lambda \log \pi_t(\cdot|s), \pi(\cdot|s) - \pi_t(\cdot|s) \big\rangle - D_{\mathrm{KL}}\left(\pi(\cdot|s) \| \pi_t(\cdot|s)\right) \Big\}, \quad (10)$$

where $\alpha_t > 0$ is the stepsize. This step corresponds to one iteration of Proximal Policy Optimization (PPO) (Schulman et al., 2017). It can also be viewed as one mirror descent iteration, where the shifted Q-function $\widehat{Q}_t^\lambda(s, \cdot) - \lambda \log \pi_t(\cdot|s)$ plays the role of the gradient. The maximizer $\widehat{\pi}_{t+1}$ in equation (10) can be computed in closed form as done in equation (8) in Algorithm 1. We then compute the new policy π_{t+1} by mixing $\widehat{\pi}_{t+1}$ with a small amount of uniform distribution, as done in equation (9). "Mixing in" a uniform distribution is a standard technique to prevent the policy from approaching the boundary of the probability simplex and becoming degenerate. Doing so allows us to upper bound a quantity of the form $D_{\mathrm{KL}}\left(p \,\|\, \pi_{t+1}(\cdot|s)\right)$ (cf. Lemma 2), which otherwise may be infinite. It also ensures that the KL divergence satisfies a Lipschitz condition (cf. Lemma 3).

We remark that our algorithm is similar to the classical fictitious play approach for finding NEs, where each agent plays a response to the empirical average of its opponent's past behaviors. In our algorithm, the representative agent views the population of all agents collectively as an opponent. Expanding the recursion (8) and ignoring the difference

between $\widehat{\pi}_{t+1}$ and π_{t+1}, we can write the policy π_{t+1} as

$$\pi_{t+1}(\cdot|s) \propto \exp\left(\sum_{\tau=0}^{t} w_\tau \widehat{Q}_\tau^\lambda(s,\cdot)\right)$$

for some positive weights $\{w_\tau\}$. Therefore, the representative agent is playing a policy that responds to the (weighted) average of all previous Q functions, which reflects the representative agent's belief on the aggregate population policy. Such a soft-max policy gives a good response, while ensuring the policy does not change too quickly. The evolution of the mean-field is controlled in a similar manner.

Our algorithm is *online* in the sense that it alternatively updates the mean-field state (*playing*) by one step of gradient descent, and improves the policy (*learning*) by one step of policy optimization using the feedback information from playing the game. Note that our algorithm only performs a single policy improvement step to compute the updated policy π_{t+1}. It is unnecessary to compute the exact optimal policy $\pi_t^* = \Gamma_1^\lambda(\mu_t)$ under μ_t (which would require an inner loop for solving MDP_{μ_t} by fixing the mean-field state μ_t), as μ_t is only an approximate anyway of the true NE mean-field μ^*.

4. Main Results

In this section, we establish the theoretical guarantees on learning the regularized NE (π^*, μ^*) of the MFG for our fictitious play algorithm. To state our theorem, we first discuss several regularity assumptions on the MFG model. Recall the definition (5) of the discounted state visitation distribution and let $\rho^* := \rho_{\mu^*}^{\pi^*} \in \Delta(\mathcal{S})$ be the visitation distribution induced by the NE (π^*, μ^*). We make use of the following distance metric between two policies $\pi, \pi' \in \Pi$:

$$D(\pi, \pi') := \mathbb{E}_{s\sim\rho^*}\left[\|\pi(\cdot|s) - \pi'(\cdot|s)\|_1\right]. \qquad (11)$$

To ensure the convergence of learning algorithms, we consider mean-field games where the NE exists. For discrete-time unregularized MFG, some prior works investigate sufficient conditions on the system parameters (i.e., transition kernel and reward function) for the NE operator (i.e., $\Gamma : \mathcal{M} \to \mathcal{M}$) being contractive (Adlakha et al., 2015; Anahtarci et al., 2020b; 2019; Saldi et al., 2018b; 2019), which guarantees the existence of NE. For the regularized MFG considered in this paper, recent work by Anahtarci et al. (2020a) shows that the NE exists under quite mild assumptions as opposed to the unregularized case. In particular, when both the reward function and the Markov transition kernel are Lipschitz continuous, one can show that for any mean-field state, the corresponding optimal policy is Lipschitz continuous w.r.t. the mean-field term. Moreover, the NE operator is contractive under certain condition on the Lipschitz constants of the reward and transitions kernel.

We are thus motivated to assume Lipschitzness for the two mappings $\Gamma_1^\lambda : \mathcal{M} \to \Pi$ and $\Gamma_2 : \Pi \times \mathcal{M} \to \mathcal{M}$ defined in Section 2.3.

Assumption 2 states that $\Gamma_1^\lambda(\mu)$ is Lipschitz in the mean-embedded mean-field state μ with respect to the RKHS norm; Assumption 3 states that $\Gamma_2(\pi, \mu)$ is Lipschitz in each of its arguments when the other argument is fixed.

Assumption 2. *There exists a constant $d_1 > 0$, such that for any $\mu, \mu' \in \mathcal{M}$, it holds that*

$$D\left(\Gamma_1^\lambda(\mu), \Gamma_1^\lambda(\mu')\right) \le d_1 \|\mu - \mu'\|_{\mathcal{H}}.$$

Assumption 3. *There exist constants $d_2 > 0, d_3 > 0$ such that for any policies $\pi, \pi' \in \Pi$ and embedded mean-field states $\mu, \mu' \in \mathcal{M}$, it holds that*

$$\|\Gamma_2(\pi, \mu) - \Gamma_2(\pi', \mu)\|_{\mathcal{H}} \le d_2 D(\pi, \pi'),$$
$$\|\Gamma_2(\pi, \mu) - \Gamma_2(\pi, \mu')\|_{\mathcal{H}} \le d_3 \|\mu - \mu'\|_{\mathcal{H}}.$$

Assumptions 2 and 3 immediately imply Lipschitzness of the composite mapping $\Lambda^\lambda : \mathcal{M} \to \mathcal{M}$, which we recall is defined as $\Lambda^\lambda(\mu) = \Gamma_2\left(\Gamma_1^\lambda(\mu), \mu\right)$. The proof is provided in Section A. We remark that the operator Λ is contractive if $d_1 d_2 + d_3 < 1$, which holds if the transition kernel and reward function are Lipschitz and the corresponding Lipschitz constants are small enough.

Lemma 1. *Suppose Assumptions 2 and 3 hold. Then for each $\mu, \mu' \in \mathcal{M}$, it holds that*

$$\left\|\Lambda^\lambda(\mu) - \Lambda^\lambda(\mu')\right\|_{\mathcal{H}} \le (d_1 d_2 + d_3) \|\mu - \mu'\|_{\mathcal{H}}.$$

We next impose an assumption on the boundedness of certain concentrability coefficients. This type of assumption, standard in analysis of policy optimization algorithms (Kakade & Langford, 2002; Shani et al., 2019; Bhandari & Russo, 2019; Agarwal et al., 2020), allows one to define the policy optimization error in an average-case sense with respect to appropriate distributions over the states.

Assumption 4 (Finite Concentrability Coefficients)**.** *There exist two constants $C_\rho, \overline{C}_\rho > 0$ such that for each $\mu \in \mathcal{M}$,*

$$\left\|\frac{\rho_\mu^{\pi_\mu^{\lambda,*}}}{\rho^*}\right\|_\infty := \sup_s \left[\frac{\rho_\mu^{\pi_\mu^{\lambda,*}}(s)}{\rho^*(s)}\right] \le C_\rho,$$
$$\left\{\mathbb{E}_{s\sim\rho_\mu^{\pi_\mu^{\lambda,*}}}\left[\left|\frac{\rho^*(s)}{\rho_\mu^{\pi_\mu^{\lambda,*}}(s)}\right|^2\right]\right\}^{1/2} \le \overline{C}_\rho.$$

Finally, our last assumption stipulates that the state visitation distributions are smooth with respect to the (embedded) mean-field states of the MFG. This assumption is analogous to those in the literature on MDP and two-player games (Fei et al., 2020; Radanovic et al., 2019), which requires the visitation distributions to be smooth w.r.t. the policy.

Assumption 5. *There exists a constant $d_0 > 0$, such that for any $\mu, \mu' \in \mathcal{M}$, it holds that the discounted state visitation distributions induced by the corresponding optimal policy $\pi_\mu^{\lambda,*}$ for regularized MDP_μ and $\pi_{\mu'}^{\lambda,*}$ for regularized $\mathrm{MDP}_{\mu'}$ satisfy*

$$\left\| \rho_\mu^{\pi_\mu^{\lambda,*}} - \rho_{\mu'}^{\pi_{\mu'}^{\lambda,*}} \right\|_1 \leq d_0 \left\| \mu - \mu' \right\|_{\mathcal{H}}.$$

We now state our theoretical guarantees on the convergences of the policy-population sequence $\{\pi_t, \mu_t\}$ in Algorithm 1 to the NE $\{\pi^*, \mu^*\}$. For the embedded mean-field states, it is natural to consider the distance $\|\mu_t - \mu^*\|_{\mathcal{H}}$ in RKHS norm. For convergence to NE policy μ^*, recall that μ^* is the optimal policy to MDP_{μ^*}, and each iteration of our algorithm involves a single policy improvement step to update π_t rather than solving MDP_{μ_t} to its optimal policy $\pi_t^* := \Gamma_1^\lambda(\mu_t)$. As such, we analyze the difference between these two policies in terms of $D(\pi_t, \pi_t^*)$, where the metric D is defined in equation (11). Also let $\rho_t^* := \rho_{\mu_t}^{\pi_t^*}$ denote the discounted visitation distribution induced by the optimal policy π_t^* of MDP_{μ_t}.[2] With the above considerations in mind, we have the following theorem, which is proved in Section C.

Theorem 1. *Suppose that Assumptions 1–5 hold with $\overline{d} := d_1 d_2 + d_3 < 1$ and that the error in the policy evaluation step in Algorithm 1 satisfies*

$$\mathbb{E}_{s\sim\rho_t^*}\left[\left\| Q_t^\lambda(s,\cdot) - \widehat{Q}_t^\lambda(s,\cdot) \right\|_\infty^2 \right] \leq \varepsilon_Q^2, \qquad \forall t \in [T].$$

With the choice of

$$\eta = c_\eta T^{-1},\ \alpha_t \equiv \alpha = c_\alpha T^{-2/5},\ \beta_t \equiv \beta = c_\beta T^{-4/5},$$

for some universal constants $c_\eta > 0$, $c_\alpha > 0$ and $c_\beta > 0$ in Algorithm 1, the resulting policy and embedded mean-field state sequence $\{(\pi_t, \mu_t)\}_{t=0}^T$ satisfy

$$D\Big(\frac{1}{T}\sum_{t=0}^{T-1}\pi_t, \frac{1}{T}\sum_{t=0}^{T-1}\pi_t^*\Big) \leq \frac{1}{T}\sum_{t=0}^{T-1} D(\pi_t, \pi_t^*)$$
$$\lesssim \frac{1}{\sqrt{\lambda}} \cdot \Big(\sqrt{\log T} \cdot T^{-1/5} + \sqrt{\varepsilon_Q} \Big), \tag{12}$$
$$\Big\| \frac{1}{T}\sum_{t=0}^{T-1}\mu_t - \mu^* \Big\|_{\mathcal{H}} \leq \frac{1}{T}\sum_{t=0}^{T-1} \|\mu_t - \mu^*\|_{\mathcal{H}}$$
$$\lesssim \frac{1}{\sqrt{\lambda}} \cdot \Big(\sqrt{\log T} \cdot T^{-1/5} + \sqrt{\varepsilon_Q} \Big). \tag{13}$$

Proof outline: To analyze the convergence behavior of the policy-population sequence $\{(\pi_t, \mu_t)\}$, we keep track of two error terms: $\sigma_\pi^t := \mathbb{E}_{s\sim\rho_t^*}\left[D_{\mathrm{KL}}\left(\pi_t^*(\cdot|s) \| \pi_t(\cdot|s)\right)\right]$ (the gap between the current policy and the optimal policy under the current mean field) and $\sigma_\mu^t := \|\mu_t - \mu^*\|_{\mathcal{H}}$ (the gap between the current mean field and equilibrium mean field). The key step is showing that the policy gap satisfies the bound $\sigma_\pi^{t+1} \leq (1 - \lambda\alpha)\sigma_\pi^t + \mathcal{O}\left(\|\mu_{t+1} - \mu_t\|_{\mathcal{H}} + \varepsilon_Q \alpha + \alpha^2 \right)$. Per our update rule for the mean field, it is ensured that $\|\mu_{t+1} - \mu_t\|_{\mathcal{H}} \leq 2\beta$. Combining with the previous bound establishes the recursion

$$\sigma_\pi^t \leq \frac{1}{\lambda\alpha}\left(\sigma_\pi^t - \sigma_\pi^{t+1}\right) + \mathcal{O}\Big(\frac{\beta}{\lambda\alpha} + \frac{\varepsilon_Q}{\lambda} + \frac{\alpha}{\lambda}\Big),$$

which implies convergence of σ_π^t under our choice of step sizes (α, β). We remark that the convergence relies on the distinction of time-scales between α and β. This bound for σ_π^t can be further translated to an error bound in the distance metric $D(\cdot,\cdot)$, as stated in the theorem, under the concentrability coefficient assumption. On the other hand, we show that the mean field gap satisfies the bound

$$\sigma_\mu^t \leq \mathcal{O}\Big(\frac{1}{\beta}\left(\sigma_\mu^t - \sigma_\mu^{t+1}\right) + \sqrt{\sigma_\pi^t}\Big),$$

which together with the bound for σ_π^t implies convergence of σ_μ^t. □

Theorem 1 bounds the distance between π_t and the optimal policy π_t^* of $\mathrm{MDP}_{\mu_t^*}$. By directly measuring the distance between π_t and the NE policy π^*, we can define the notion of an θ-approximate NE of the game.

Definition 2. *For each $\theta > 0$, a policy-population pair (π, μ) is called an θ-approximate (entropy-regularized) NE of the MFG if*

$$D(\pi, \pi^*) \leq \theta \quad \textit{and} \quad \|\mu - \mu^*\|_{\mathcal{H}} \leq \theta.$$

The following corollary of Theorem 1 states that after T iterations of our algorithm, the average policy-population pair $(\frac{1}{T}\sum_{t=0}^{T-1}\pi_t, \frac{1}{T}\sum_{t=0}^{T-1}\mu_t)$ is an $\widetilde{\mathcal{O}}\left(T^{-1/5}\right)$-approximate NE.

Corollary 1. *Under the assumptions of Theorem 1, we have*

$$D\Big(\frac{1}{T}\sum_{t=0}^{T-1}\pi_t, \pi^*\Big) + \Big\|\frac{1}{T}\sum_{t=0}^{T-1}\mu_t - \mu^*\Big\|_{\mathcal{H}}$$
$$\lesssim \frac{1}{\sqrt{\lambda}} \cdot \Big(\sqrt{\log T} \cdot T^{-1/5} + \sqrt{\varepsilon_Q} \Big).$$

We prove this corollary in Section D.

The above results require an ℓ_2-error of ε_Q for policy evaluation. A variety of algorithms have been shown to achieve such guarantees, including TD(0) and LSTD (Bhandari et al., 2018). It is also worth emphasizing that the convergence rate to the regularized NE scales inverse proportionally with $\sqrt{\lambda}$, implying that convergence can be accelerated with a higher

[2] The subscript in ρ_t^* emphasizes that ρ_t^* only depends on the mean-field state μ_t at time t through $\pi_t^* = \Gamma_1^\lambda(\mu_t)$.

level of entropy regularization. On other hand, the approximation error of the regularized NE for the original unregularized NE scales proportionally with λ (cf. (6)). Therefore, it is desirable to choose the regularization parameter λ that balances the target accuracy level and convergence rate.

We also remark that the ℓ_∞ condition on concentrability coefficient in Assumption 4 can be relaxed to an ℓ_2 condition of the form $\left\{\mathbb{E}\left[\left|\rho_\mu^{\pi_\mu^{\lambda,*}}(s)/\rho^*(s)\right|^2\right]\right\}^{1/2} \leq C_\rho$, under which we can establish an $\widetilde{O}(T^{-1/9})$ convergence rate; see Theorem 2 and Corollary 2 in Section E in the Supplementary Material for the details.

5. Conclusion

In this paper, we develop a provably efficient fictitious play algorithm for stationary mean-field games. In comparison to the existing work that requires solving an MDP induced by a mean-field state within each iteration, our algorithm updates both the policy and the mean-field state simultaneously in each iteration. We prove that the policy and mean-field state sequence under the proposed algorithm converges to the Nash equilibrium of the MFG at an explicit rate.

A number of directions are of interest for future research. An immediate step is to investigate whether the convergence rate can be improved. The $\widetilde{O}(T^{-1/5})$ convergence rate we obtain here relies on constant step-sizes. It would be interesting to see if using time-varying step-sizes can attain a faster convergence rate. Another research direction worth pursuing is generalizing our approach for developing decentralized/distributed learning schemes.

Acknowledgements

Qiaomin Xie is partially supported by the National Science Foundation grant CNS-1955997 and Google 2020 System Research Award. Zhuoran Yang acknowledges the support of Simons Institute (Theory of Reinforcement Learning). Zhaoran Wang acknowledges National Science Foundation (Awards 2048075, 2008827, 2015568, 1934931), Simons Institute (Theory of Reinforcement Learning), Amazon, J.P. Morgan, and Two Sigma for their supports. Andreea Minca is partially supported by the National Science Foundation grant 1653354.

References

Adlakha, S., Johari, R., and Weintraub, G. Y. Equilibria of dynamic games with many players: Existence, approximation, and market structure. *Journal of Economic Theory*, 156:269–316, 2015.

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. Optimality and approximation with policy gradient methods in markov decision processes. In *Conference on Learning Theory, COLT 2020*, volume 125 of *Proceedings of Machine Learning Research*, pp. 64–66. PMLR, 2020.

Anahtarci, B., Kariksiz, C. D., and Saldi, N. Fitted q-learning in mean-field games. *arXiv preprint arXiv:1912.13309*, 2019.

Anahtarci, B., Kariksiz, C. D., and Saldi, N. Q-learning in regularized mean-field games. *arXiv preprint arXiv:2003.12151*, 2020a.

Anahtarci, B., Kariksiz, C. D., and Saldi, N. Value iteration algorithm for mean-field games. *Systems & Control Letters*, 143:104744, 2020b.

Angiuli, A., Fouque, J.-P., and Laurière, M. Unified reinforcement q-learning for mean field game and control problems. *arXiv preprint arXiv:2006.13912*, 2020.

Başar, T. and Olsder, G. J. *Dynamic noncooperative game theory*. SIAM, 1998.

Bellemare, M. G., Dabney, W., and Munos, R. A distributional perspective on reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 449–458. PMLR, 2017.

Berner, C., Brockman, G., Chan, B., Cheung, V., Dębiak, P., Dennison, C., Farhi, D., Fischer, Q., Hashme, S., Hesse, C., et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019.

Bhandari, J. and Russo, D. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*, 2019.

Bhandari, J., Russo, D., and Singal, R. A finite time analysis of temporal difference learning with linear function approximation. In *Conference On Learning Theory, COLT 2018*, volume 75 of *Proceedings of Machine Learning Research*, pp. 1691–1692. PMLR, 2018.

Biswas, A. Mean field games with ergodic cost for discrete time markov processes. *arXiv preprint arXiv:1510.08968*, 2015.

Brown, G. W. Iterative solution of games by fictitious play. *Activity analysis of production and allocation*, 13(1):374–376, 1951.

Busoniu, L., De Schutter, B., and Babuska, R. Decentralized reinforcement learning control of a robotic manipulator. In *2006 9th International Conference on Control, Automation, Robotics and Vision*, pp. 1–6. IEEE, 2006.

Busoniu, L., Babuska, R., and De Schutter, B. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(2):156–172, 2008.

Cai, Q., Yang, Z., Lee, J. D., and Wang, Z. Neural temporal-difference and q-learning provably converge to global optima. *arXiv preprint arXiv:1905.10027*, 2019.

Cao, K., Lazaridou, A., Lanctot, M., Leibo, J. Z., Tuyls, K., and Clark, S. Emergent communication through negotiation. In *6th International Conference on Learning Representations, ICLR 2018*. OpenReview.net, 2018.

Caponnetto, A. and De Vito, E. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Carmona, R. and Delarue, F. *Probabilistic Theory of Mean Field Games with Applications I-II*. Springer, 2018.

Cen, S., Cheng, C., Chen, Y., Wei, Y., and Chi, Y. Fast global convergence of natural policy gradient methods with entropy regularization. *arXiv preprint arXiv:2007.06558*, 2020.

Elie, R., Pérolat, J., Laurière, M., Geist, M., and Pietquin, O. On the convergence of model free learning in mean field games. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*, pp. 7143–7150. AAAI Press, 2020.

Farahmand, A.-m., Ghavamzadeh, M., Szepesvári, C., and Mannor, S. Regularized policy iteration with nonparametric function spaces. *The Journal of Machine Learning Research*, 17(1):4809–4874, 2016.

Fei, Y., Yang, Z., Wang, Z., and Xie, Q. Dynamic regret of policy optimization in non-stationary environments. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, 2020.

Fu, Z., Yang, Z., Chen, Y., and Wang, Z. Actor-critic provably finds nash equilibria of linear-quadratic mean-field games. In *8th International Conference on Learning Representations, ICLR 2020*. OpenReview.net, 2020.

Geist, M., Scherrer, B., and Pietquin, O. A theory of regularized markov decision processes. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2160–2169. PMLR, 2019.

Gomes, D. A., Mohr, J., and Souza, R. R. Discrete time, finite state space mean field games. *Journal de mathématiques pures et appliquées*, 93(3):308–328, 2010.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. J. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems 2006*, pp. 513–520. MIT Press, 2006.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Guéant, O., Lasry, J.-M., and Lions, P.-L. Mean field games and applications. In *Paris-Princeton lectures on mathematical finance 2010*, pp. 205–266. Springer, 2011.

Guo, X., Hu, A., Xu, R., and Zhang, J. Learning mean-field games. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*, pp. 4967–4977, 2019.

Guo, X., Hu, A., Xu, R., and Zhang, J. A general framework for learning mean-field games. *arXiv preprint arXiv:2003.06069*, 2020.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1856–1865. PMLR, 2018.

Hernandez-Leal, P., Kaisers, M., Baarslag, T., and de Cote, E. M. A survey of learning in multiagent environments: Dealing with non-stationarity. *arXiv preprint arXiv:1707.09183*, 2017.

Hernandez-Leal, P., Kartal, B., and Taylor, M. E. Is multiagent deep reinforcement learning the answer or the question? a brief survey. *learning*, 21:22, 2018.

Huang, M., Caines, P. E., and Malhamé, R. P. Individual and mass behaviour in large population stochastic wireless power control problems: centralized and nash equilibrium solutions. In *42nd IEEE International Conference on Decision and Control (IEEE Cat. No. 03CH37475)*, volume 1, pp. 98–103. IEEE, 2003.

Huang, M., Caines, P. E., and Malhamé, R. P. Large-population cost-coupled LQG problems with nonuniform agents: individual-mass behavior and decentralized ε-Nash equilibria. *IEEE Transactions on Automatic Control*, 52(9):1560–1571, 2007.

Jaques, N., Lazaridou, A., Hughes, E., Gülçehre, Ç., Ortega, P. A., Strouse, D., Leibo, J. Z., and de Freitas, N. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *Proceedings of the 36th*

International Conference on Machine Learning, ICML 2019, volume 97 of *Proceedings of Machine Learning Research*, pp. 3040–3049. PMLR, 2019.

Kakade, S. M. and Langford, J. Approximately optimal approximate reinforcement learning. In *Machine Learning, Proceedings of the Nineteenth International Conference (ICML 2002)*, pp. 267–274. Morgan Kaufmann, 2002.

Konda, V. R. and Tsitsiklis, J. N. Actor-critic algorithms. In *Advances in neural information processing systems*, pp. 1008–1014, 2000.

Kuyer, L., Whiteson, S., Bakker, B., and Vlassis, N. Multiagent reinforcement learning for urban traffic control using coordination graphs. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 656–671. Springer, 2008.

Lasry, J.-M. and Lions, P.-L. Jeux à champ moyen. i–le cas stationnaire. *Comptes Rendus Mathématique*, 343(9): 619–625, 2006a.

Lasry, J.-M. and Lions, P.-L. Jeux à champ moyen. ii–horizon fini et contrôle optimal. *Comptes Rendus Mathématique*, 343(10):679–684, 2006b.

Lasry, J.-M. and Lions, P.-L. Mean field games. *Japanese journal of mathematics*, 2(1):229–260, 2007.

Leibo, J. Z., Zambaldi, V., Lanctot, M., Marecki, J., and Graepel, T. Multi-agent reinforcement learning in sequential social dilemmas. *arXiv preprint arXiv:1702.03037*, 2017.

Leottau, D. L., Ruiz-del Solar, J., and Babuška, R. Decentralized reinforcement learning of robot behaviors. *Artificial Intelligence*, 256:130–159, 2018.

Lin, S.-B., Guo, X., and Zhou, D.-X. Distributed learning with regularized least squares. *The Journal of Machine Learning Research*, 18(1):3202–3232, 2017.

Mannion, P., Duggan, J., and Howley, E. An experimental review of reinforcement learning algorithms for adaptive traffic signal control. In *Autonomic road transport support systems*, pp. 47–66. Springer, 2016.

McKee, K. R., Gemp, I., McWilliams, B., Duéñez-Guzmán, E. A., Hughes, E., and Leibo, J. Z. Social diversity and social preferences in mixed-motive reinforcement learning. *arXiv preprint arXiv:2002.02325*, 2020.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *nature*, 518(7540): 529–533, 2015.

Moon, J. and Başar, T. Discrete-time lqg mean field games with unreliable communication. In *53rd IEEE Conference on Decision and Control*, pp. 2697–2702. IEEE, 2014.

Muandet, K., Fukumizu, K., Dinuzzo, F., and Schölkopf, B. Learning from distributions via support measure machines. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012.*, pp. 10–18, 2012.

Muandet, K., Fukumizu, K., Sriperumbudur, B., Schölkopf, B., et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

Munos, R. and Szepesvári, C. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9 (May):815–857, 2008.

Nachum, O., Norouzi, M., Xu, K., and Schuurmans, D. Bridging the gap between value and policy based reinforcement learning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pp. 2775–2785, 2017.

Nash, J. F. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences of the United States of America*, 36(1):48–49, 1950.

Perrin, S., Pérolat, J., Laurière, M., Geist, M., Elie, R., and Pietquin, O. Fictitious play for mean field games: Continuous time analysis and applications. *arXiv preprint arXiv:2007.03458*, 2020.

Radanovic, G., Devidze, R., Parkes, D. C., and Singla, A. Learning to collaborate in markov decision processes. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5261–5270. PMLR, 2019.

Salakhutdinov, R. Deep learning. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2014*, pp. 1973. ACM, 2014.

Saldi, N., Basar, T., and Raginsky, M. Discrete-time risk-sensitive mean-field games. *arXiv preprint arXiv:1808.03929*, 2018a.

Saldi, N., Basar, T., and Raginsky, M. Markov–nash equilibria in mean-field games with discounted cost. *SIAM Journal on Control and Optimization*, 56(6):4256–4287, 2018b.

Saldi, N., Başar, T., and Raginsky, M. Approximate nash equilibria in partially observed stochastic games with

mean-field interactions. *Mathematics of Operations Research*, 44(3):1006–1033, 2019.

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., et al. Mastering atari, go, chess and shogi by planning with a learned model. *arXiv preprint arXiv:1911.08265*, 2019.

Schulman, J., Levine, S., Abbeel, P., Jordan, M. I., and Moritz, P. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pp. 1889–1897. JMLR.org, 2015.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Shah, D. and Xie, Q. Q-learning with nearest neighbors. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018*, pp. 3115–3125, 2018.

Shani, L., Efroni, Y., and Mannor, S. Adaptive trust region policy optimization: Global convergence and faster rates for regularized mdps. *arXiv preprint arXiv:1909.02769*, 2019.

Shoham, Y., Powers, R., and Grenager, T. If multi-agent learning is the answer, what is the question? *Artificial intelligence*, 171(7):365–377, 2007.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., and Lanctot, M. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484, 2016.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., and Bolton, A. Mastering the game of Go without human knowledge. *Nature*, 550(7676):354–359, 2017.

Smola, A., Gretton, A., Song, L., and Schölkopf, B. A hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory*, pp. 13–31. Springer, 2007.

Sonu, E., Chen, Y., and Doshi, P. Decision-theoretic planning under anonymity in agent populations. *Journal of Artificial Intelligence Research*, 59:725–770, 2017.

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. R. Hilbert space embeddings and metrics on probability measures. *The Journal of Machine Learning Research*, 11:1517–1561, 2010.

Subramanian, J. and Mahajan, A. Reinforcement learning in stationary mean-field games. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 251–259, 2019.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Szabó, Z., Gretton, A., Póczos, B., and Sriperumbudur, B. K. Two-stage sampled learning theory on distributions. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2015*, volume 38 of *JMLR Workshop and Conference Proceedings*. JMLR.org, 2015.

Tembine, H. and Huang, M. Mean field difference games: Mckean-vlasov dynamics. In *2011 50th IEEE Conference on Decision and Control and European Control Conference*, pp. 1006–1011. IEEE, 2011.

uz Zaman, M. A., Zhang, K., Miehling, E., and Başar, T. Approximate equilibrium computation for discrete-time linear-quadratic mean-field games. In *2020 American Control Conference (ACC)*, pp. 333–339. IEEE, 2020.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575 (7782):350–354, 2019.

Watkins, C. J. and Dayan, P. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

Więcek, P. Discrete-time ergodic mean-field games with average reward on compact spaces. *Dynamic Games and Applications*, 10(1):222–256, 2020.

Wooldridge, M. *An introduction to multiagent systems*. John Wiley & Sons, 2009.

Yang, E. and Gu, D. Multiagent reinforcement learning for multi-robot systems: A survey. *Technical Report CSM 04*, 2004.

Zhang, K., Yang, Z., and Başar, T. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *arXiv preprint arXiv:1911.10635*, 2019.