Q-Learning in Regularized Mean-field Games

Berkay Anahtarcı, Can Deha Karıksız, and Naci Saldi

Abstract—**In this paper, we introduce a regularized mean-field game and study learning of this game under an infinite-horizon discounted reward function. The game is defined by adding a regularization function to the one-stage reward function in the classical mean-field game model. We establish a value iteration based learning algorithm to this regularized mean-field game using fitted Q-learning. This regularization term in general makes reinforcement learning algorithm more robust with improved exploration. Moreover, it enables us to establish error analysis of the learning algorithm without imposing restrictive convexity assumptions on the system components, which are needed in the absence of a regularization term.**

I. Introduction

This paper deals with the learning of regularized mean-field games (MFGs) under infinite-horizon discounted reward function. In this game model, a single agent interacts with a huge population of other agents and compete with the collective behaviour of them through a mean-field term which converges to the distribution of a single generic agent, as the number of agents is being let go to infinity. In the limiting case, therefore, a generic agent faces a single-agent stochastic control problem with a constraint on the state distribution at each time. This condition specifies that the state distribution should be consistent with the behaviour of the total population. In other words, at each time step, the resulting distribution of the state of each agent is the same as the flow of the state distribution when the generic agent applies this policy. This stability condition between policy and state distribution flow is called the *mean-field equilibrium*.

The theory of MFGs has emerged in the work of Lasry and Lions [1] where the standard terminology of mean-field games was introduced, and independently as stochastic dynamic games by Huang, Malhamé and Caines [2]. They have both considered continuous time non-cooperative differential games with large but finite number of asymptotically negligible anonymous agents in interaction along with their infinite limits to establish approximate Nash equilibria. In continuous-time differential games, characterization of the mean-field equilibrium is given by coupled Hamilton-Jacobi-Bellman (HJB) equation and Kolmogorov-Fokker-Planck (FPK) equation. We refer the reader to [3]–[10] for studies of continuous-time mean-field games with different models and cost functions, such as games with major-minor players, risk-sensitive games, games with Markov jump parameters, and LQG games.

In comparison with continuous-time framework, there are relatively less results available on discrete-time mean-field games in the literature. These works have mainly studied the setting where the state space is discrete (finite or countable) and the agents are only coupled by their cost functions; that is, the mean-field term does not influence the evolution of the agents' states. In [11], a mean-field game model with finite state is studied, and [12] considers discrete-time mean-field games with an infinite-horizon discounted cost criterion over unbounded state spaces. Discrete-time mean-field games with linear state dynamics are studied in [13]–[16]. References [17]–[20] study discrete-time mean-field games subject to the average cost optimality criterion. In [21], authors consider a discrete-time risk-sensitive mean-field game with Polish state and action spaces. References [22], [23] consider a discrete-time mean-field game with Polish state and action spaces under the discounted cost optimality criterion for both fully-observed case and partially-observed case, respectively.

We note that the aforementioned papers, except linear models, mostly identify the existence of mean-field equilibrium and no algorithm with convergence guarantee has been proposed to compute this mean-field equilibrium. In our recent work [24], this problem is explored for mean-field games with abstract state and action spaces under both discounted cost and average cost criteria. We have developed a value iteration algorithm and proved that this algorithm converges to the mean-field equilibrium. In [25], we generalize this value iteration algorithm to the model-free setting using fitted Q-learning [26]. There, fitted Q-learning is used instead of classical Q-learning algorithm because the action space is assumed to be a compact and convex subset of a finite dimensional Euclidean space. In this work, we generalize this work to the regularized mean-field games.

In misspecified control models, greedy algorithms often results in policies that are far from optimal. Making use of regularization provides a way to overcome this problem. Most recent reinforcement algorithms also use regularization to increase exploration and robustness, and this regularization is generally established via entropy or relative entropy. We refer the reader to [27] for an exhaustive review of the literature on regularized Markov decision processes (MDPs) and [28] for a general framework on entropy-regularized MDPs. In this paper, we introduce regularized mean-field games, analogous to regularized MDPs. Our research seems to be the first one studying this problem. We propose a learning algorithm to compute an equilibrium solution for discrete-time regularized mean-field games under the discounted reward optimality criterion. A regularization term is added to the one-stage reward function in this game model. This is supposed to improve the algorithm's exploration and make algorithm more robust. Moreover, due to regularization term, an error analysis of the learning algorithm can be established under quite mild assumptions as opposed to the classical case. In the latter case, we need restrictive convexity conditions.

In the literature, the existence of mean-field equilibria has

Naci Saldi and Berkay Anahtarcı are with the Department of Natural and Mathematical Sciences, Özyeğin University, Cekmekoy, Istanbul, Turkey. Emails: {naci.saldi,berkay.anahtarci}@ozyegin.edu.tr, candeha@gmail.com.

been established for discrete-time mean-field games under the discounted optimality criterion in [22]. However, learning discrete-time mean-field games has not been studied much, even for the classical case, until recently. In [29], authors establish a Q-learning algorithm to compute approximate mean-field equilibria for finite state-action mean-field games. The analysis of convergence of the learning algorithm in this work is highly dependent upon the assumption that the operators in the algorithm are contractive. However, proving the contractive property of these operators is highly non-trivial and requires sophisticated convex analytic tools. Therefore, it is not judicious to state this as an assumption. In [30], authors develop a fictitious play iterative learning algorithm for mean-field games with compact state and action spaces, where the dynamics of the state and the one-stage cost function satisfy certain structure. They suggest an error analysis of the learning algorithm for the deterministic game model (no noise term in the state dynamics). Nevertheless, they do not identify the conditions on the system components for which the error bound between learned equilibrium and mean-field equilibrium converges to zero. In [31] authors study linear-quadratic mean-field games and establish the convergence of policy gradient algorithm. In [32], an actor-critic algorithm to learn mean-field equilibrium for linear-quadratic mean-field games is developed. In [33] a mean-field game in which agents can control their transition probabilities without any restriction is studied. In this case, the action space becomes the set of probability measures on the state space. With this specific model, they can transform a mean-field game into an equivalent deterministic Markov decision processes by extending the state and action spaces, and so, establish classical reinforcement learning algorithms to compute mean-field equilibrium.

In this paper, we consider a regularized discrete-time mean-field game with finite state and action spaces under infinite-horizon discounted reward. We introduce regularization as an additive term to the one-stage reward function. We establish a value iteration based learning algorithm to this regularized mean-field game using fitted Q-learning. This regularization term in general makes the learning algorithm more robust with improved exploration. Moreover, it enables us to establish error analysis of the learning algorithm without imposing restrictive convexity assumptions on the system components, which are needed in the absence of a regularization term.

The paper is set out as follows. In Section II, we introduce classical and regularized mean-field games as well as finite-agent game, and define the classical and regularized mean-field equilibria. In Section III, we define mean-field equilibrium operator and show that the mean-field equilibrium operator is contractive. In Section IV, we establish a Q-learning algorithm to compute approximate regularized-mean-field equilibrium and prove its convergence. Section V concludes the paper.

Notation. For a finite set E, we let $\mathcal{P}(\mathsf{E})$ denote the set of all probability distributions on E. In this paper, $\|\cdot\|_1$ and $\|\cdot\|_2$ denote l_1-norm and l_2-norm on $\mathcal{P}(\mathsf{E})$, respectively. Total variation norm on $\mathcal{P}(\mathsf{E})$ is denoted by $\|\cdot\|_{TV}$. For any $\mu, \nu \in \mathcal{P}(\mathsf{E})$, we have $\|\mu - \nu\|_{TV} = \inf\{\xi(1_{\{x \neq y\}}) : \xi(\cdot, \mathsf{E}) = \mu(\cdot) \text{ and } \xi(\mathsf{E}, \cdot) = \nu(\cdot)\}$ and the distribution ξ on $\mathsf{E} \times \mathsf{E}$ that achieves this infimum is called optimal coupling between μ and ν. It is known that $\|\cdot\|_1 = 2\|\cdot\|_{TV}$. In this paper, we will always endow $\mathcal{P}(\mathsf{E})$ with l_1-norm. For any $e \in \mathsf{E}$, δ_e is the Dirac delta distribution. We let $m(\cdot)$ denote the Lebesgue measure on appropriate finite dimensional Euclidean space $\mathbb{R}^d$. For any $a \in \mathbb{R}^d$ and $\rho > 0$, let $B(a, \rho) := \{b : \|a - b\|_1 \leq \rho\}$. For any $a, b \in \mathbb{R}^d$, $\langle a, b \rangle$ denotes the inner product. Let $Q : \mathsf{E}_1 \times \mathsf{E}_2 \to \mathbb{R}$, where E_1 and E_2 are two sets. Then, we define $Q_{\max}(e_1) := \sup_{e_2 \in \mathsf{E}_2} Q(e_1, e_2)$. For any function class $\mathcal{G}$, let $V_{\mathcal{G}}$ denote its pseudo-dimension. The notation $v \sim \nu$ means that the random element v has distribution ν.

II. Mean-Field Games

A discrete-time mean-field game is specified by

$$(\mathsf{X}, \mathsf{A}, p, r),$$

where X is the finite state space and A is the finite action space. The components $p : \mathsf{X} \times \mathsf{A} \times \mathcal{P}(\mathsf{X}) \to \mathcal{P}(\mathsf{X})$ and $r : \mathsf{X} \times \mathsf{A} \times \mathcal{P}(\mathsf{X}) \to [0, \infty)$ are the transition probability and the one-stage reward function, respectively. Therefore, given current state $x(t)$, action $a(t)$, and state-measure μ, the reward $r(x(t), a(t), \mu)$ is received immediately, and the next state $x(t+1)$ evolves to a new state probabilistically according to the following distribution:

$$x(t+1) \sim p(\cdot | x(t), a(t), \mu).$$

To complete the description of the model dynamics, we should also specify how the agent selects its action. To that end, a policy π is a conditional distribution on A given X; that is, $\pi : \mathsf{X} \to \mathcal{P}(\mathsf{A})$. Let Π denote the set of all policies.

In mean-field games, a state-measure $\mu \in \mathcal{P}(\mathsf{X})$ represents the collective behavior of the other agents; that is, μ can be considered as the infinite population limit of the empirical distribution of the states of other agents. Given any state-measure $\mu \in \mathcal{P}(\mathsf{X})$, a policy $\pi^* \in \Pi$ is optimal for μ if

$$J_\mu(\pi^*) = \sup_{\pi \in \Pi} J_\mu(\pi),$$

where

$$J_\mu(\pi) = E^\pi\left[\sum_{t=0}^{\infty} \beta^t r(x(t), a(t), \mu)\right]$$

is the discounted reward of policy π under the state-measure μ and $\beta \in (0,1)$ is the discount factor. Given μ, the states and actions are evolved as follows:

$$x(0) \sim \mu_0, \;\; x(t) \sim p(\,\cdot\,|x(t-1), a(t-1), \mu), \;\; t \geq 1,$$
$$a(t) \sim \pi(\,\cdot\,|x(t)), \;\; t \geq 0,$$

where μ_0 denotes the initial distribution of the state.

In this paper, we impose the following assumptions on the system components.

Assumption 1.

(a) The one-stage reward function r satisfies the following Lipschitz bound:

$$|r(x, a, \mu) - r(\hat{x}, \hat{a}, \hat{\mu})|$$

$$\leq L_1 \left(1_{\{x\neq\hat{x}\}} + 1_{\{a\neq\hat{a}\}} + \|\mu-\hat{\mu}\|_1\right), \forall x,\hat{x}, \forall a,\hat{a}, \forall \mu,\hat{\mu}.$$

(b) The stochastic kernel $p(\,\cdot\,|x,a,\mu)$ satisfies the following Lipschitz bound:

$$\begin{aligned}&\|p(\cdot|x,a,\mu)-p(\cdot|\hat{x},\hat{a},\hat{\mu})\|_1\\ &\leq K_1 \left(1_{\{x\neq\hat{x}\}} + 2\cdot 1_{\{a\neq\hat{a}\}} + \|\mu-\hat{\mu}\|_1\right), \forall x,\hat{x}, \forall a,\hat{a}, \forall \mu,\hat{\mu}.\end{aligned}$$

Note that we can equivalently describe the model above as follows. In this equivalent model, we take action space to be the set of probability measures $\mathsf{U} := \mathcal{P}(\mathsf{A})$ on the original action space A. Hence, the new action space U is an uncountable, convex, and compact subset of $\mathbb{R}^{\mathsf{A}}$ with dimension $|\mathsf{A}|-1$. With this new action space, the new transition probability $P : \mathsf{X}\times\mathsf{U}\times\mathcal{P}(\mathsf{X}) \to \mathcal{P}(\mathsf{X})$ and the new one-stage reward function $R : \mathsf{X}\times\mathsf{U}\times\mathcal{P}(\mathsf{X}) \to \mathbb{R}$ are defined as follows:

$$\begin{aligned}P(\,\cdot\,|x,u,\mu) &:= \sum_{a\in\mathsf{A}} p(\,\cdot\,|x,a,\mu)\,u(a),\\ R(x,u,\mu) &:= \sum_{a\in\mathsf{A}} r(x,a,\mu)\,u(a).\end{aligned}$$

In this equivalent model, a policy π is a deterministic function from state space X to the new action space U. Therefore, for a fixed μ, the reward function of any policy π is given by

$$J_\mu(\pi) = E^\pi\left[\sum_{t=0}^{\infty}\beta^t R(x(t),u(t),\mu)\right],$$

where

$$\begin{aligned}&x(0)\sim\mu_0,\;\; x(t)\sim P(\,\cdot\,|x(t-1),u(t-1),\mu),\;\; t\geq 1,\\ &u(t)=\pi(x(t)),\;\; t\geq 0.\end{aligned}$$

In the remainder of this paper, we replace the original mean-field game model with this equivalent one. We prove below the conditions satisfied by the new transition probability P and one-stage reward function R under Assumption 1.

Proposition 1. Under Assumption 1, P and R satisfy the following Lipschitz bounds:

$$\begin{aligned}&|R(x,u,\mu)-R(\hat{x},\hat{u},\hat{\mu})|\\ &\leq L_1 \left(1_{\{x\neq\hat{x}\}} + \|u-\hat{u}\|_1 + \|\mu-\hat{\mu}\|_1\right), \forall x,\hat{x}, \forall u,\hat{u}, \forall \mu,\hat{\mu}.\end{aligned}$$

and

$$\begin{aligned}&\|P(\cdot|x,u,\mu)-P(\cdot|\hat{x},\hat{u},\hat{\mu})\|_1\\ &\leq K_1 \left(1_{\{x\neq\hat{x}\}} + \|u-\hat{u}\|_1 + \|\mu-\hat{\mu}\|_1\right), \forall x,\hat{x}, \forall u,\hat{u}, \forall \mu,\hat{\mu}.\end{aligned}$$

Proof. Fix any $x,\hat{x},u,\hat{u},\mu,\hat{\mu}$. Then we have

$$\begin{aligned}&|R(x,u,\mu)-R(\hat{x},\hat{u},\hat{\mu})|\\ &= \left|\sum_{a\in\mathsf{A}} r(x,a,\mu)\,u(a) - \sum_{\hat{a}\in\mathsf{A}} r(\hat{x},a,\hat{\mu})\,\hat{u}(a)\right|\\ &\leq \left|\sum_{a\in\mathsf{A}} r(x,a,\mu)\,u(a) - \sum_{\hat{a}\in\mathsf{A}} r(x,a,\mu)\,\hat{u}(a)\right|\\ &\qquad + \left|\sum_{a\in\mathsf{A}} r(x,a,\mu)\,\hat{u}(a) - \sum_{\hat{a}\in\mathsf{A}} r(\hat{x},a,\hat{\mu})\,\hat{u}(a)\right|\\ &\leq L_1 \left(1_{\{x\neq\hat{x}\}} + \|u-\hat{u}\|_1 + \|\mu-\hat{\mu}\|_1\right).\end{aligned}$$

Similarly, we have

$$\begin{aligned}&\|P(\cdot|x,u,\mu)-P(\cdot|\hat{x},\hat{u},\hat{\mu})\|_1\\ &= \sum_{y\in\mathsf{X}} |P(y|x,u,\mu)-P(y|\hat{x},\hat{u},\hat{\mu})|\\ &= \sum_{y\in\mathsf{X}}\left|\sum_{a\in\mathsf{A}} p(y|x,a,\mu)\,u(a) - \sum_{a\in\mathsf{A}} p(y|\hat{x},a,\hat{\mu})\,\hat{u}(a)\right|\\ &\leq \sum_{y\in\mathsf{X}}\left|\sum_{a\in\mathsf{A}} p(y|x,a,\mu)\,u(a) - \sum_{a\in\mathsf{A}} p(y|x,a,\mu)\,\hat{u}(a)\right|\\ &\qquad + \sum_{y\in\mathsf{X}}\left|\sum_{a\in\mathsf{A}} p(y|x,a,\mu)\,\hat{u}(a) - \sum_{a\in\mathsf{A}} p(y|\hat{x},a,\hat{\mu})\,\hat{u}(a)\right|\\ &\overset{(1)}{\leq} K_1\|u-\hat{u}\|_1\\ &\qquad + \sum_{y\in\mathsf{X}}\left|\sum_{a\in\mathsf{A}} p(y|x,a,\mu)\,\hat{u}(a) - \sum_{a\in\mathsf{A}} p(y|\hat{x},a,\hat{\mu})\,\hat{u}(a)\right|\\ &\leq K_1 \left(1_{\{x\neq\hat{x}\}} + \|u-\hat{u}\|_1 + \|\mu-\hat{\mu}\|_1\right).\end{aligned}$$

To show that (1) follows from Assumption 1-(b), let us define the transition probability $M : \mathsf{A}\to\mathcal{P}(\mathsf{X})$ as

$$M(\cdot|a) := p(\cdot|x,a,\mu).$$

Let $\xi\in\mathcal{P}(\mathsf{A}\times\mathsf{A})$ be the optimal coupling of u and $\hat{u}$ that achieves total variation distance $\|u-\hat{u}\|_{TV}$. Similarly, for any $a,\hat{a}\in\mathsf{A}$, let $K(\cdot|a,\hat{a})\in\mathcal{P}(\mathsf{X}\times\mathsf{X})$ be the optimal coupling of $M(\cdot|a)$ and $M(\cdot|\hat{a})$ that achieves total variation distance $\|M(\cdot|a)-M(\cdot|\hat{a})\|_{TV}$. Note that

$$\begin{aligned}\sum_{y\in\mathsf{X}}\left|\sum_{a\in\mathsf{A}} p(y|x,a,\mu)\,u(a) - \sum_{a\in\mathsf{A}} p(y|x,a,\mu)\,\hat{u}(a)\right|&\\ = 2\|uM-\hat{u}M\|_{TV},&\end{aligned}$$

where

$$uM(\cdot) := \sum_{a\in\mathsf{A}} M(\cdot|a)\,u(a)$$

and

$$\hat{u}M(\cdot) := \sum_{a\in\mathsf{A}} M(\cdot|a)\,\hat{u}(a).$$

Let us define $\nu(\cdot) := \sum_{(a,\hat{a})\mathsf{A}\times\mathsf{A}} K(\cdot|a,\hat{a})\,\xi(a,\hat{a})$, and so, ν is a coupling of uM and $\hat{u}M$. Therefore, we have

$$\begin{aligned}2\,\|uM-\hat{u}M\|_{TV} &\leq 2\sum_{(x,y)\in\mathsf{X}\times\mathsf{X}} 1_{\{x\neq y\}}\,\nu(x,y)\\ &= 2\sum_{(a,\hat{a})\in\mathsf{A}\times\mathsf{A}}\;\sum_{(x,y)\in\mathsf{X}\times\mathsf{X}} 1_{\{x\neq y\}}\,K(x,y|a,\hat{a})\,\xi(a,\hat{a})\\ &= \sum_{(a,\hat{a})\in\mathsf{A}\times\mathsf{A}} \|M(\cdot|a)-M(\cdot|\hat{a})\|_1\,\xi(a,\hat{a})\\ &\leq 2\,K_1\sum_{(a,\hat{a})\in\mathsf{A}\times\mathsf{A}} 1_{\{a\neq\hat{a}\}}\,\xi(a,\hat{a})\\ &= K_1\,\|u-\hat{u}\|_1.\end{aligned}$$

Hence, (1) follows. This completes the proof. □

Now, we can define the optimality criteria of the model. To this end, we need to define two set-valued mappings. The first set-valued mapping $\Psi : \mathcal{P}(\mathsf{X}) \to 2^{\Pi}$ is defined as follows:

$$\Psi(\mu) = \{\hat{\pi} \in \Pi : J_{\mu}(\hat{\pi}) = \sup_{\pi} J_{\mu}(\pi) \ \ \text{and} \ \ \mu_0 = \mu\}.$$

The set $\Psi(\mu)$ is the set of optimal policies for μ when the initial distribution is μ as well. The second set-valued mapping $\Lambda : \Pi \to 2^{\mathcal{P}(\mathsf{X})}$ is defined as follows: for any $\pi \in \Pi$, the state-measure $\mu_{\pi} \in \Lambda(\pi)$ is the invariant distribution of the transition probability $P(\,\cdot\,|x, \pi(x), \mu_{\pi})$; that is,

$$\mu_{\pi}(\,\cdot\,) = \sum_{x \in \mathsf{X}} P(\,\cdot\,|x, \pi(x), \mu_{\pi})\, \mu_{\pi}(x).$$

Under Assumption 1 and Proposition 1, $\Lambda(\pi)$ is always non-empty.

We may now define the notion of equilibrium (called mean-field equilibrium) for mean-field games using these mappings Ψ, Λ as follows.

Definition 1. A pair $(\pi_*, \mu_*) \in \Pi \times \mathcal{P}(\mathsf{X})$ is a *mean-field equilibrium* if $\pi_* \in \Psi(\mu_*)$ and $\mu_* \in \Lambda(\pi_*)$.

A. Regularized Mean-Field Games

A theory of regularized Markov decision processes (MDPs) has been introduced in [27]. In this work, regularization is introduced via subtracting a strongly convex function from the one-stage reward function. This type of modifications is in general applied to reinforcement learning algorithms to ensure robust learners with improved exploration. We refer the reader to [27] for comprehensive review on a variety of regularized MDPs used in the literature.

Analogous to regularized MDPs, in this section, we introduce regularized mean-field games. To that end, let $\Omega : \mathsf{U} \to \mathbb{R}$ be a ρ-strongly convex function. Let L_{reg} be the Lipschitz constant of Ω on U, whose existence is guaranteed by strong convexity of Ω. The only difference between classical MFGs and regularized ones is the regularization term in the one-stage reward function. In regularized MFGs, the reward function is given by

$$R^{\mathrm{reg}}(x, u, \mu) \coloneqq R(x, u, \mu) - \Omega(u).$$

A typical example for Ω is the negative entropy $\Omega(u) = \sum_{a \in \mathsf{A}} \ln(u(a))\, u(a)$. Another example is the relative entropy between u and uniform distribution; that is, $\Omega(u) = \sum_{a \in \mathsf{A}} \ln(u(a))\, u(a) + \ln(|\mathsf{A}|)$. In both of these examples, as a result of entropy regularization, agent visits optimal and as well as almost optimal actions more often and randomly. This improves the exploration of the algorithm. Moreover, due to strong convexity of Ω, Lipschitz sensitivity of the optimal action on state, state-measure, and other uncertain parameters can be established via Legendre-Fenchel duality. This makes the learning algorithm more robust.

In regularized MFGs, for a fixed μ, the reward function of any policy π is given by

$$J_{\mu}^{\mathrm{reg}}(\pi) = E^{\pi}\left[\sum_{t=0}^{\infty} \beta^t R^{\mathrm{reg}}(x(t), u(t), \mu)\right].$$

For this model, we define the set-valued mapping $\Psi^{\mathrm{reg}} : \mathcal{P}(\mathsf{X}) \to 2^{\Pi}$ as follows:

$$\Psi^{\mathrm{reg}}(\mu) = \{\hat{\pi} \in \Pi : J_{\mu}^{\mathrm{reg}}(\hat{\pi}) = \sup_{\pi} J_{\mu}^{\mathrm{reg}}(\pi) \ \ \text{and} \ \ \mu_0 = \mu\}.$$

Similarly, we define the set-valued mapping $\Lambda^{\mathrm{reg}} : \Pi \to 2^{\mathcal{P}(\mathsf{X})}$ as follows: for any $\pi \in \Pi$, the state-measure $\mu_{\pi} \in \Lambda^{\mathrm{reg}}(\pi)$ is the invariant distribution of the transition probability $P(\,\cdot\,|x, \pi(x), \mu_{\pi})$; that is,

$$\mu_{\pi}(\,\cdot\,) = \sum_{x \in \mathsf{X}} P(\,\cdot\,|x, \pi(x), \mu_{\pi})\, \mu_{\pi}(x).$$

Then, the notion of equilibrium for this regularized game model is defined as follows, which is similar to the definition of mean-field equilibrium.

Definition 2. A pair $(\pi_*, \mu_*) \in \Pi \times \mathcal{P}(\mathsf{X})$ is a *regularized mean-field equilibrium* if $\pi_* \in \Psi^{\mathrm{reg}}(\mu_*)$ and $\mu_* \in \Lambda^{\mathrm{reg}}(\pi_*)$.

In this paper, our goal is to develop a Q-learning algorithm for computing an approximate regularized mean-field equilibrium when the model is unknown. To that end, we define the following.

Definition 3. Let $(\pi_*, \mu_*) \in \Pi \times \mathcal{P}(\mathsf{X})$ be a *regularized mean-field equilibrium*. A pair $(\pi_{\varepsilon}, \mu_*) \in \Pi \times \mathcal{P}(\mathsf{X})$ is an ε-regularized-mean-field equilibrium if

$$J_{\mu_*}^{\mathrm{reg}}(\pi_{\varepsilon}) \geq \sup_{\pi \in \Pi} J_{\mu_*}^{\mathrm{reg}}(\pi) - \varepsilon = J_{\mu_*}^{\mathrm{reg}}(\pi_*) - \varepsilon;$$

that is, instead of optimality, we require that π_{ε} is ε-optimal.

With this definition, our goal now is to learn an ε-regularized-mean-field equilibrium using Q-learning algorithm.

B. Finite Agent Game

The regularized mean-field game model is indeed the infinite-population limit of the regularized finite-agent game model that will be described below. In finite-agent game model, we have N-agents and, for each agent $i \in \{1, 2, \ldots, N\}$, $x_i^N(t) \in \mathsf{X}$ and $u_i^N(t) \in \mathsf{U}$ denote the state and the action of Agent i at time t, respectively. The empirical distribution of the states of agents at time t is defined as follows:

$$e_t^{(N)}(\,\cdot\,) \coloneqq \frac{1}{N} \sum_{i=1}^{N} \delta_{x_i^N(t)}(\,\cdot\,) \in \mathcal{P}(\mathsf{X}).$$

This empirical distribution affects both the system dynamics and one-stage reward function. Therefore, for each $t \geq 0$, next states $(x_1^N(t+1), \ldots, x_N^N(t+1))$ of agents have the following conditional distribution given current states $(x_1^N(t), \ldots, x_N^N(t))$ and actions $(u_1^N(t), \ldots, u_N^N(t))$:

$$\prod_{i=1}^{N} P\big(dx_i^N(t+1) \big| x_i^N(t), u_i^N(t), e_t^{(N)}\big).$$

A *policy* π for a generic agent is a deterministic function from X to U. The set of all policies for Agent i is denoted by Π_i. The initial states $x_i^N(0)$ are independent and identically distributed according to μ_0.

Let $\boldsymbol{\pi}^{(N)} := (\pi^1, \ldots, \pi^N)$, $\pi^i \in \Pi_i$, denote an N-tuple of policies. Under such an N-tuple of policies, the regularized discounted reward of Agent i is defined as

$$J_i^{(N)}(\boldsymbol{\pi}^{(N)}) = E^{\boldsymbol{\pi}^{(N)}}\left[\sum_{t=0}^{\infty} \beta^t R^{\text{reg}}(x_i^N(t), u_i^N(t), e_t^{(N)})\right].$$

Then, the goal of the agents is to achieve a Nash equilibrium, which is defined as follows.

Definition 4. An N-tuple of policies $\boldsymbol{\pi}^{(N*)} = (\pi^{1*}, \ldots, \pi^{N*})$ is a *Nash equilibrium* if

$$J_i^{(N)}(\boldsymbol{\pi}^{(N*)}) = \sup_{\pi^i \in \Pi_i} J_i^{(N)}(\boldsymbol{\pi}_{-i}^{(N*)}, \pi^i)$$

for each $i = 1, \ldots, N$, where $\boldsymbol{\pi}_{-i}^{(N*)} := (\pi^{j*})_{j \neq i}$.

It is known that establishing the existence of Nash equilibria and computing it are in general prohibitive for finite-agent game model as a result of the decentralized nature of the problem (see [22, pp. 4259]). Therefore, it is of interest to obtain an approximate Nash equilibrium, whose definition is given below.

Definition 5. An N-tuple of policies $\boldsymbol{\pi}^{(N*)} = (\pi^{1*}, \ldots, \pi^{N*})$ constitutes an *δ-Nash equilibrium* if

$$J_i^{(N)}(\boldsymbol{\pi}^{(N*)}) \geq \sup_{\pi^i \in \Pi_i} J_i^{(N)}(\boldsymbol{\pi}_{-i}^{(N*)}, \pi^i) - \delta$$

for each $i = 1, \ldots, N$, where $\boldsymbol{\pi}_{-i}^{(N*)} := (\pi^{j*})_{j \neq i}$.

Due to symmetry in mean-field game model, if the number of agents is large enough, one can obtain approximate Nash equilibrium by studying the infinite population limit $N \to \infty$ of the game (i.e., mean-field game). Indeed, one can prove that if each agent in the finite-agent game model adopts the ε-regularized-mean-field equilibrium policy, the resulting policy will be an approximate Nash equilibrium for all sufficiently large N-agent game models. Indeed, this is the statement of the below theorem.

Theorem 1. Let (π_ε, μ_*) be an ε-regularized-mean-field equilibrium. Then, for any $\epsilon > 0$, there exists a positive integer $N(\epsilon)$, such that, for each $N \geq N(\epsilon)$, the N-tuple of policies $\boldsymbol{\pi}^{(N)} = \{\pi_\varepsilon, \pi_\varepsilon, \ldots, \pi_\varepsilon\}$ is an $(\varepsilon + \epsilon)$-Nash equilibrium for the game with N agents.

Proof. Note that we must prove that

$$J_i^{(N)}(\boldsymbol{\pi}^{(N)}) \geq \sup_{\pi^i \in \Pi_i} J_i^{(N)}(\boldsymbol{\pi}_{-i}^{(N)}, \pi^i) - \varepsilon - \epsilon \tag{1}$$

for each $i = 1, \ldots, N$, when N is sufficiently large. As the transition probabilities and the one-stage reward functions are the same for all agents, it is sufficient to prove (1) for Agent 1 only. Given $\epsilon > 0$, for each $N \geq 1$, let $\tilde{\pi}^{(N)} \in \Pi_1$ be such that

$$J_1^{(N)}(\tilde{\pi}^{(N)}, \pi_\varepsilon, \ldots, \pi_\varepsilon) > \sup_{\pi' \in \Pi_1} J_1^{(N)}(\pi', \pi_\varepsilon, \ldots, \pi_\varepsilon) - \frac{\epsilon}{3}.$$

Then, by [22, Corollary 4.11], we have

$$\begin{aligned}
\lim_{N \to \infty} J_1^{(N)}(\tilde{\pi}^{(N)}, \pi_\varepsilon, \ldots, \pi_\varepsilon) &= \lim_{N \to \infty} J_{\boldsymbol{\mu}_*}^{\text{reg}}(\tilde{\pi}^{(N)}) \\
&\leq \sup_{\pi'} J_{\boldsymbol{\mu}_*}^{\text{reg}}(\pi') \\
&\leq J_{\boldsymbol{\mu}_*}^{\text{reg}}(\pi_\varepsilon) + \varepsilon \\
&= \lim_{N \to \infty} J_1^{(N)}(\pi_\varepsilon, \pi_\varepsilon, \ldots, \pi_\varepsilon) + \varepsilon.
\end{aligned}$$

Therefore, there exists $N(\epsilon)$ such that

$$\begin{aligned}
&\sup_{\pi' \in \Pi_1} J_1^{(N)}(\pi', \pi_\varepsilon, \ldots, \pi_\varepsilon) - \epsilon - \varepsilon \\
&\leq J_1^{(N)}(\tilde{\pi}^{(N)}, \pi_\varepsilon, \ldots, \pi_\varepsilon) - \frac{2\epsilon}{3} - \varepsilon \\
&\leq J_{\boldsymbol{\mu}_*}^{\text{reg}}(\pi_\varepsilon) - \frac{\epsilon}{3} \\
&\leq J_1^{(N)}(\pi_\varepsilon, \pi_\varepsilon, \ldots, \pi_\varepsilon).
\end{aligned}$$

for all $N \geq N(\epsilon)$. □

Theorem 1 states that if one can learn ε-regularized-mean-field equilibrium, then the learned policy will be an approximate Nash equilibrium for the finite-agent game problem, where computing or learning the exact Nash equilibrium is in general prohibitive.

In the next section, we will first introduce a mean-field equilibrium (MFE) operator, which can be used to compute mean-field equilibrium when the model is known, and prove that this operator is contractive. Then, under model-free setting, we approximate this MFE operator with a random one and establish a learning algorithm. Using this random operator, we obtain ε-regularized-mean-field equilibrium with high confidence. This learned approximate regularized-mean-field equilibrium can then be used in finite-agent game model as an approximate Nash equilibrium.

III. Mean-Field Equilibrium Operator

In this section, we introduce a mean-field equilibrium (MFE) operator, whose fixed point is a mean-field equilibrium. We prove that this operator is contractive. Using this result, we then establish a Q-learning algorithm to obtain approximate regularized mean-field equilibrium. To that end, in addition to Assumption 1, we assume the following. This assumption ensures that the MFE operator is contractive. To state the assumption, we need to define the constants below:

$$\begin{aligned}
r_{\max} &:= \sup_{(x,u,\mu) \in \mathsf{X} \times \mathsf{U} \times \mathcal{P}(\mathsf{X})} |R^{\text{reg}}(x, u, \mu)|, \\
Q_{\max} &:= \frac{r_{\max}}{1 - \beta}, \; Q_{\text{Lip}} := L_1 + L_{\text{reg}} + \beta \, Q_{\max} \, K_1.
\end{aligned}$$

Assumption 2. We assume that

$$\frac{3\,K_1}{2}\left(1 + \frac{\sqrt{|\mathsf{A}|}}{\rho}\,\frac{L_1 + \beta\,Q_{\max}\,K_1}{1 - \beta}\right) < 1.$$

Given any state-measure μ, the regularized value function J_μ^{reg} of policy π with initial state x is defined as

$$J_\mu^{\text{reg}}(\pi, x) := E^\pi\left[\sum_{t=0}^{\infty} \beta^t R^{\text{reg}}(x(t), u(t), \mu) \,\middle|\, x(0) = x\right].$$

Then, the optimal regularized value function is given by

$$J_\mu^{\text{reg},*}(x) := \sup_{\pi \in \Pi} J_\mu^{\text{reg}}(\pi, x).$$

Similarly, we define the optimal regularized Q-function as

$$Q_\mu^{\mathrm{reg},*}(x,u) = R^{\mathrm{reg}}(x,u,\mu) + \beta \sum_{y\in\mathsf{X}} J_\mu^{\mathrm{reg},*}(y)\, P(y|x,u,\mu).$$

Note that $Q^{\mathrm{reg},*}_{\mu,\max}(x) := \sup_{u\in\mathsf{U}} Q_\mu^{\mathrm{reg},*}(x,u) = J^*_\mu(x)$ for all $x\in\mathsf{X}$. Therefore, we have the following optimality equation:

$$\begin{aligned} Q_\mu^{\mathrm{reg},*}(x,u) &= R^{\mathrm{reg}}(x,u,\mu) + \beta \sum_{y\in\mathsf{X}} Q_{\mu,\max}^{\mathrm{reg},*}(y)\, P(y|x,u,\mu) \\ &=: H_\mu Q_\mu^{\mathrm{reg},*}(x,u). \end{aligned}$$

Here, it is known that H_μ is a $\|\cdot\|_\infty$-contraction with modulus β and the unique fixed point of H_μ is $Q_\mu^{\mathrm{reg},*}$.

Let $\mathcal{C}$ denote the set of all Q-functions. We assume that any $Q\in\mathcal{C}$ is uniformly Q_{Lip}- Lipschitz continuous and ρ-strongly concave with respect to u. For any $Q\in\mathcal{C}$, the sup-norm is defined as $\|Q\|_\infty := \sup_{(x,u)\in\mathsf{X}\times\mathsf{U}} |Q(x,u)|$. In this paper, we will always endow $\mathcal{C}$ with the sup-norm $\|\cdot\|_\infty$.

Now, we define the MFE operator. To that end, we define $H_1 : \mathcal{P}(\mathsf{X}) \to \mathcal{C}$ as $H_1(\mu) = Q_\mu^{\mathrm{reg},*}$ (optimal regularized Q-function) and $H_2 : \mathcal{P}(\mathsf{X})\times\mathcal{C} \to \mathcal{P}(\mathsf{X})$ as

$$H_2(\mu, Q)(\cdot) := \sum_{x\in\mathsf{X}} P(\cdot|x, f_Q(x), \mu)\, \mu(x),$$

where $f_Q(x) = \arg\max_{u\in\mathsf{U}} Q(x,u)$ for all $x\in\mathsf{X}$. With these definitions, we can give the definition of the optimality operator as follows:

$$H : \mathcal{P}(\mathsf{X}) \ni \mu \mapsto H_2\left(\mu, H_1(\mu)\right) \in \mathcal{P}(\mathsf{X}).$$

Our goal is to prove that H is contractive. In the following lemma, we prove that H_1 is contractive, which will be used to prove that H operator is also contractive.

Lemma 1. The mapping H_1 is a contraction with contraction constant $K_{H_1} := \dfrac{L_1 + \beta\, Q_{\max}\, K_1}{1-\beta}$.

Proof. Under Assumption 1, it is straightforward to prove that H_1 maps $\mathcal{P}(\mathsf{X})$ into $\mathcal{C}$.

For any $\mu, \hat{\mu} \in \mathcal{P}(\mathsf{X})$, we have

$$\begin{aligned} &\|H_1(\mu) - H_1(\hat{\mu})\|_\infty = \|Q_\mu^{\mathrm{reg},*} - Q_{\hat{\mu}}^{\mathrm{reg},*}\|_\infty \\ &= \sup_{x,u} \Bigg| R(x,u,\mu) + \beta \sum_y Q_{\mu,\max}^{\mathrm{reg},*}(y) P(y|x,u,\mu) \\ &\qquad\qquad - R(x,u,\hat{\mu}) - \beta \sum_y Q_{\hat{\mu},\max}^{\mathrm{reg},*}(y) P(y|x,u,\hat{\mu}) \Bigg| \\ &\le L_1\, \|\mu - \hat{\mu}\|_1 \\ &+ \beta \left| \sum_y Q_{\mu,\max}^{\mathrm{reg},*}(y) P(y|x,u,\mu) - \sum_y Q_{\mu,\max}^{\mathrm{reg},*}(y) P(y|x,u,\hat{\mu}) \right| \\ &+ \beta \left| \sum_y Q_{\mu,\max}^{\mathrm{reg},*}(y) P(y|x,u,\hat{\mu}) - \sum_y Q_{\hat{\mu},\max}^{\mathrm{reg},*}(y) P(y|x,u,\hat{\mu}) \right| \\ &\le L_1\, \|\mu - \hat{\mu}\|_1 + \beta\, Q_{\max}\, K_1\, \|\mu - \hat{\mu}\|_1 + \beta\, \|Q_\mu^{\mathrm{reg},*} - Q_{\hat{\mu}}^{\mathrm{reg},*}\|_\infty. \end{aligned}$$

This completes the proof. □

Now, using Lemma 1, we can prove that H is contractive.

Proposition 2. The mapping H is a contraction with contraction constant K_H, where

$$K_H := \frac{3\, K_1}{2} \left(1 + \frac{\sqrt{|\mathsf{A}|}}{\rho}\, K_{H_1} \right).$$

Proof. For any $\mu \in \mathcal{P}(\mathsf{X})$, we have

$$\begin{aligned} Q_\mu^{\mathrm{reg},*}(x,u) &= H_\mu Q_\mu^{\mathrm{reg},*}(x,u) \\ &= R(x,u,\mu) + \beta \sum_{y\in\mathsf{X}} Q_{\mu,\max}^{\mathrm{reg},*}(y)\, P(y|x,u,\mu) - \Omega(u) \\ &= \langle q_x^\mu, u \rangle - \Omega(u), \end{aligned}$$

where

$$q_x^\mu(\cdot) := r(x,\cdot,\mu) + \beta \sum_{y\in\mathsf{X}} Q_{\mu,\max}^{\mathrm{reg},*}(y)\, p(y|x,\cdot,\mu).$$

As a result of this representation of $Q_\mu^{\mathrm{reg},*}$ and ρ-strong convexity of Ω, by [27, Proposition 1], $Q_\mu^{\mathrm{reg},*}(x,\cdot)$ has a unique minimizer $f_{Q_\mu^{\mathrm{reg},*}}(x) \in \mathsf{U}$ for any $x\in\mathsf{X}$, and, for any $\mu, \hat{\mu} \in \mathcal{P}(\mathsf{X})$ and $x, \hat{x} \in \mathsf{X}$, we have

$$\|f_{Q_\mu^{\mathrm{reg},*}}(x) - f_{Q_{\hat{\mu}}^{\mathrm{reg},*}}(\hat{x})\|_2 \le \frac{1}{\rho}\, \|q_x^\mu - q_{\hat{x}}^{\hat{\mu}}\|_2 \le \frac{1}{\rho}\, \|q_x^\mu - q_{\hat{x}}^{\hat{\mu}}\|_1.$$

Note that we have

$$\begin{aligned} &\|q_x^\mu - q_{\hat{x}}^{\hat{\mu}}\|_1 \\ &= \sum_{a\in\mathsf{A}} \Bigg| r(x,a,\mu) + \beta \sum_{y\in\mathsf{X}} Q_{\mu,\max}^{\mathrm{reg},*}(y)\, p(y|x,a,\mu) \\ &\qquad\qquad - r(\hat{x},a,\hat{\mu}) - \beta \sum_{y\in\mathsf{X}} Q_{\hat{\mu},\max}^{\mathrm{reg},*}(y)\, p(y|\hat{x},a,\hat{\mu}) \Bigg| \\ &\le L_1(1_{\{x\neq\hat{x}\}} + \|\mu - \hat{\mu}\|_1) \\ &+ \beta \sum_{a\in\mathsf{A}} \Bigg| \sum_y Q_{\mu,\max}^{\mathrm{reg},*}(y) p(y|x,a,\mu) \\ &\qquad\qquad - \sum_y Q_{\hat{\mu},\max}^{\mathrm{reg},*}(y) p(y|x,a,\mu) \Bigg| \\ &+ \beta \sum_{a\in\mathsf{A}} \Bigg| \sum_y Q_{\hat{\mu},\max}^{\mathrm{reg},*}(y) p(y|x,a,\mu) \\ &\qquad\qquad - \sum_y Q_{\hat{\mu},\max}^{\mathrm{reg},*}(y) p(y|\hat{x},a,\hat{\mu}) \Bigg| \\ &\le L_1(1_{\{x\neq\hat{x}\}} + \|\mu - \hat{\mu}\|_1) + \beta \|Q_\mu^{\mathrm{reg},*} - Q_{\hat{\mu}}^{\mathrm{reg},*}\|_\infty \\ &\qquad\qquad + \beta Q_{\max} K_1 (1_{\{x\neq\hat{x}\}} + \|\mu - \hat{\mu}\|_1) \\ &\le (L_1 + \beta Q_{\max} K_1)(1_{\{x\neq\hat{x}\}} + \|\mu - \hat{\mu}\|_1) + \beta K_{H_1} \|\mu - \hat{\mu}\|_1 \\ &\le K_{H_1}(1_{\{x\neq\hat{x}\}} + \|\mu - \hat{\mu}\|_1). \end{aligned}$$

Therefore we obtain

$$\|f_{Q_\mu^{\mathrm{reg},*}}(x) - f_{Q_{\hat{\mu}}^{\mathrm{reg},*}}(\hat{x})\|_2 \le \frac{1}{\rho}\, K_{H_1}(1_{\{x\neq\hat{x}\}} + \|\mu - \hat{\mu}\|_1).$$

Since $\|v\|_1 \le \sqrt{|\mathsf{A}|}\, \|v\|_2$, we have

$$\begin{aligned} &\|f_{Q_\mu^{\mathrm{reg},*}}(x) - f_{Q_{\hat{\mu}}^{\mathrm{reg},*}}(\hat{x})\|_1 \\ &\qquad\qquad \le \frac{\sqrt{|\mathsf{A}|}}{\rho}\, K_{H_1}(1_{\{x\neq\hat{x}\}} + \|\mu - \hat{\mu}\|_1). \end{aligned} \tag{2}$$

Now, fix any $\mu, \hat{\mu} \in \mathcal{P}(\mathsf{X})$. Using (2), we have

$$
\begin{aligned}
&\|H_2(\mu, H_1(\mu)) - H_2(\hat{\mu}, H_1(\hat{\mu}))\|_1 \\
&= \sum_y \left| \sum_x P(y|x, f_{Q_\mu^{\mathrm{reg},*}}(x), \mu)\, \mu(x) \right. \\
&\qquad\qquad \left. - \sum_x P(y|x, f_{Q_{\hat{\mu}}^{\mathrm{reg},*}}(x), \hat{\mu})\, \hat{\mu}(x) \right| \\
&\leq \sum_y \left| \sum_x P(y|x, f_{Q_\mu^{\mathrm{reg},*}}(x), \mu)\, \mu(x) \right. \\
&\qquad\qquad \left. - \sum_x P(y|x, f_{Q_{\hat{\mu}}^{\mathrm{reg},*}}(x), \hat{\mu})\, \mu(x) \right| \\
&+ \sum_y \left| \sum_x P(y|x, f_{Q_{\hat{\mu}}^{\mathrm{reg},*}}(x), \hat{\mu})\, \mu(x) \right. \\
&\qquad\qquad \left. - \sum_x P(y|x, f_{Q_{\hat{\mu}}^{\mathrm{reg},*}}(x), \hat{\mu})\, \hat{\mu}(x) \right| \\
&\overset{(1)}{\leq} \sum_x \left\| P(\cdot|x, f_{Q_\mu^{\mathrm{reg},*}}(x), \mu) - P(\cdot|x, f_{Q_{\hat{\mu}}^{\mathrm{reg},*}}(x), \hat{\mu}) \right\|_1 \mu(x) \\
&\qquad\qquad + \frac{K_1}{2} \left(1 + \frac{\sqrt{|\mathsf{A}|}}{\rho} K_{H_1} \right) \|\mu - \hat{\mu}\|_1 \\
&\leq K_1 \left(\|f_{Q_\mu^{\mathrm{reg},*}}(x) - f_{Q_{\hat{\mu}}^{\mathrm{reg},*}}(x)\| + \|\mu - \hat{\mu}\|_1 \right) \\
&\qquad\qquad + \frac{K_1}{2} \left(1 + \frac{\sqrt{|\mathsf{A}|}}{\rho} K_{H_1} \right) \|\mu - \hat{\mu}\|_1 \\
&\leq \frac{3\,K_1}{2} \left(1 + \frac{\sqrt{|\mathsf{A}|}}{\rho} K_{H_1} \right) \|\mu - \hat{\mu}\|_1. \qquad (3)
\end{aligned}
$$

Note that (2) and Proposition 1 lead to

$$
\begin{aligned}
&\|P(\cdot|x, f_{Q_{\hat{\mu}}^{\mathrm{reg},*}}(x), \hat{\mu}) - P(\cdot|y, f_{Q_{\hat{\mu}}^{\mathrm{reg},*}}(y), \hat{\mu})\|_1 \\
&\qquad\qquad \leq K_1 \left(1 + \frac{\sqrt{|\mathsf{A}|}}{\rho} K_{H_1} \right). \qquad (4)
\end{aligned}
$$

Hence, (1) follows from [34, Lemma A2]. This completes the proof. □

Under Assumption 1 and Assumption 2, H is a contraction mapping. Therefore, by Banach Fixed Point Theorem, H has an unique fixed point. Let μ_* be this unique fixed point and $Q_{\mu_*}^{\mathrm{reg},*} = H_1(\mu_*)$. Let $\pi_*(x) = f_{Q_{\mu_*}^{\mathrm{reg},*}}(x)$. Then, one can prove that the pair (π_*, μ_*) is a regularized mean-field equilibrium. Hence, we can compute this regularized mean-field equilibrium via applying H recursively starting from arbitrary μ_0. This indeed leads to a value iteration algorithm for computing mean-field equilibrium. However, if the model is unknown, we replace H with a random operator and establish a learning algorithm via this random operator. To prove the convergence of this learning algorithm, the contraction property of H is crucial.

IV. Q-Learning Algorithm

In this section, we establish a learning algorithm for obtaining approximate regularized mean-field equilibrium. In this learning algorithm, we replace operators H_1 and H_2 with random operators $\hat{H}_1$ and $\hat{H}_2$, respectively. Therefore, we have two stages in each iteration of the learning algorithm. In the first stage, the optimal regularized Q-function $Q_\mu^{\mathrm{reg},*}$ for a given μ is learned via fitted Q-learning algorithm, which has been introduced in [26] to learn optimal Q-functions of Markov decision processes. This stage replaces the operator H_1 with a random operator $\hat{H}_1$. In this fitted Q-learning algorithm, Q-functions are picked from a fixed function class $\mathcal{F} \subset \mathcal{C}$. This function class $\mathcal{F}$ can be chosen as the set of neural networks with some fixed architecture or linear span of some finite number of basis functions or the set $\mathcal{C}$ itself. Depending on $\mathcal{F}$, an additional representation error in the learning algorithm will be present. Let $\mathcal{F}_{\max} := \{Q_{\max} : Q \in \mathcal{F}\}$.

In the second stage of each iteration, the state-measure is updated via simulating corresponding transition probability. This stage replaces the operator H_2 with a random operator $\hat{H}_2$. Below, we give the overall description of the algorithm first, and then, the descriptions of $\hat{H}_1$ and $\hat{H}_2$ are given along with their error analysis, respectively.

Algorithm 1 Algorithm H

```
Inputs $\left(K, \{[N_k, L_k]\}_{k=0}^{K}, \{M_k\}_{k=0}^{K-1}, \mu_0\right)$
Start with $\mu_0$
for $k = 0, \ldots, K-1$ do
    $\mu_{k+1} = \hat{H}\left([N_k, L_k], M_k\right)(\mu_k)$
        $:= \hat{H}_2[M_k]\left(\mu_k, \hat{H}_1[N_k, L_k](\mu_k)\right)$
end for
return $\mu_K$ and $Q_K = \hat{H}_1([N_K, L_K])(\mu_K)$
```

We proceed by giving the description of $\hat{H}_1$ first. Let ν be a probability measure on X. We fix some function $\pi_b : \mathsf{X} \to \mathcal{P}(\mathsf{U})$ such that, for any $x \in \mathsf{X}$, the distribution $\pi_b(x)(\cdot)$ on U has a density with respect to Lebesgue measure m. We denote this density with $\pi_b(x,u)$. We assume that $\pi_0 := \inf_{(x,u) \in \mathsf{X} \times \mathsf{U}} \pi_b(x,u) > 0$. Now, we can give the definition of the random operator $\hat{H}_1$.

Algorithm 2 Algorithm $\hat{H}_1$

```
Inputs $([N, L], \mu)$
Start with $Q_0 = 0$
for $l = 0, \ldots, L-1$ do
    generate i.i.d. samples $\{(x_t, u_t, r_t, y_{t+1})_{t=1}^{N}\}$ using
        $x_t \sim \nu,\ u_t \sim \pi_b(x_t)(\cdot),\ r_t = R^{\mathrm{reg}}(x_t, u_t, \mu)$,
        $y_{t+1} \sim P(\cdot|x_t, u_t, \mu)$
    and set
        $Q_{l+1} = \arg\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{t=1}^{N} \frac{1}{m(\mathsf{U})\, \pi_b(x_t, u_t)} \Big| f(x_t, u_t)$
            $- \Big[ r_t + \beta \max_{u' \in \mathsf{U}} Q_l(y_{t+1}, u') \Big] \Big|^2$
end for
return $Q_L$
```

Before we describe $\hat{H}_2$, the error analysis of algorithm $\hat{H}_1$ is given. Note that there exists $\alpha > 0$ such that for any $u \in \mathsf{U}$ and $\xi > 0$, we have $m\left(B(a,\xi) \cap \mathsf{U}\right) \geq \min\left\{\alpha\, m(B(a,\xi)), m(\mathsf{A})\right\}$, where m is the Lebesgue measure on U (when considered as a subset of $\mathbb{R}^{|\mathsf{A}|-1}$) (see [35]). To this end, we need to define the following constants:

$$E(\mathcal{F}) := \sup_{\mu \in \mathcal{P}(\mathsf{X})} \sup_{Q \in \mathcal{F}} \inf_{Q' \in \mathcal{F}} \left[\sum_x \int_{\mathsf{U}} |Q'(x,u) - H_\mu Q(x,u)|^2 \frac{m(du)}{m(\mathsf{U})} v(x) \right]^{1/2}$$

$$L_{\max} := (1+\beta) Q_{\max} + r_{\max}, \;\; C := \frac{L_{\max}^2}{m(\mathsf{U})\, \pi_0}$$

$$\Upsilon = 8\, e^2 \left(V_{\mathcal{F}} + 1\right) \left(V_{\mathcal{F}_{\max}} + 1\right) \times \left(\frac{64 e Q_{\max} L_{\max} (1+\beta)}{m(\mathsf{U}) \pi_0} \right)^{V_{\mathcal{F}} + V_{\mathcal{F}_{\max}}}$$

$$V = V_{\mathcal{F}} + V_{\mathcal{F}_{\max}}, \;\; \gamma = 512 C^2$$

$$\Delta := \frac{1}{1-\beta} \left[\frac{m(\mathsf{U}) |\mathsf{A}|!}{\alpha (2/Q_{\mathrm{Lip}})^{|\mathsf{A}|-1}}\, E(\mathcal{F}) \right]^{\frac{1}{\dim_{\mathsf{A}} + 1}}$$

$$\Lambda := \frac{1}{1-\beta} \left[\frac{m(\mathsf{U}) |\mathsf{A}|!}{\alpha (2/Q_{\mathrm{Lip}})^{|\mathsf{A}|-1}} \right]^{\frac{1}{|\mathsf{A}|}}.$$

The following theorem gives the error analysis of the algorithm $\hat{H}_1$.

Theorem 2. ([25, Theorem 4.1]) For any $(\varepsilon, \delta) \in (0,1)^2$, with probability at least $1-\delta$, we have

$$\left\| \hat{H}_1[N,L](\mu) - H_1(\mu) \right\|_\infty \leq \varepsilon + \Delta$$

if $\frac{\beta^L}{1-\beta} Q_{\max} < \frac{\varepsilon}{2}$ and $N \geq m_1(\epsilon, \delta, L)$, where

$$m_1(\varepsilon, \delta, L) := \frac{\gamma (2\Lambda)^{4|\mathsf{A}|}}{\varepsilon^{4|\mathsf{A}|}} \ln\left(\frac{\Upsilon (2\Lambda)^{2V|\mathsf{A}|} L}{\delta \varepsilon^{2V|\mathsf{A}|}} \right).$$

Here, the constant error Δ is as a result of the representation error $E(\mathcal{F})$ in the algorithm.

Next, we describe the random operator $\hat{H}_2$, and then, give the error analysis.

Algorithm 3 Algorithm $\hat{H}_2$

```
Inputs $(M, \mu, Q)$
for $x \in \mathsf{X}$ do
    generate i.i.d. samples $\{y_t^x\}_{t=1}^M$ using
        $y_t^x \sim P(\cdot|x, f_Q(x), \mu)$
    and define
        $P_M(\cdot|x, f_Q(x), \mu) = \frac{1}{M} \sum_{t=1}^M \delta_{y_t^x}(\cdot).$
end for
return $\sum_{x \in \mathsf{X}} P_M(\cdot|x, f_Q(x), \mu)\, \mu(x)$
```

Theorem 3. ([25, Theorem 4.2]) For any $(\varepsilon, \delta) \in (0,1)^2$, with probability at least $1-\delta$

$$\left\| \hat{H}_2[M](\mu, Q) - H_2(\mu, Q) \right\|_1 \leq \varepsilon$$

if $M \geq m_2(\epsilon, \delta)$, where

$$m_2(\epsilon, \delta) := \frac{|\mathsf{X}|^2}{\varepsilon^2} \ln\left(\frac{2\,|\mathsf{X}|^2}{\delta} \right).$$

Using above error analyses of the algorithms $\hat{H}_1$ and $\hat{H}_2$, we can now obtain the following error analysis for the algorithm $\hat{H}$. Then, the main result of this paper can be stated as a corollary of this result.

Theorem 4. Fix any $(\varepsilon, \delta) \in (0,1)^2$. Define

$$\varepsilon_1 := \frac{(1-K_H)^2\, \varepsilon^2}{16\, \theta\, (K_1)^2}, \;\; \varepsilon_2 := \frac{(1-K_H)\, \varepsilon}{4},$$

where $\theta := \frac{4\sqrt{|\mathsf{A}|}}{\rho}$. Let K, L be such that

$$\frac{(K_H)^K}{1-K_H} \leq \frac{\varepsilon}{2}, \;\; \frac{\beta^L}{1-\beta} Q_{\max} \leq \frac{\varepsilon_1}{2}.$$

Then, pick N, M such that

$$N \geq m_1\left(\varepsilon_1, \frac{\delta}{2K}, L\right), \;\; M \geq m_2\left(\varepsilon_2, \frac{\delta}{2K}\right). \tag{5}$$

Let (μ_K, Q_K) be the output of the learning algorithm established by random operator $\hat{H}$ with inputs

$$\left(K, \{[N,L]\}_{k=0}^K, \{M\}_{k=0}^{K-1}, \mu_0\right).$$

Then, with probability at least $1-\delta$

$$\|\mu_K - \mu_*\|_1 \leq \frac{K_1 \sqrt{\theta\, \Delta}}{(1-K_H)} + \varepsilon,$$

where μ_* is the unique fixed point of H in regularized mean-field equilibrium. Moreover, with probability at least $1 - \frac{\delta}{2K}$

$$\|Q_K - H_1(\mu_K)\|_\infty \leq \varepsilon_1 + \Delta.$$

Proof. First of all, the last statement follows from Theorem 2.

To prove the first statement, note that for any $\mu \in \mathcal{P}(\mathsf{X})$ and $Q, \hat{Q} \in \mathcal{C}$, we have

$$\begin{aligned} &\|H_2(\mu, Q) - H_2(\mu, \hat{Q})\|_1 \\ &= \sum_{y \in \mathsf{X}} \left| \sum_{x \in \mathsf{X}} P(y|x, f_Q(x), \mu)\, \mu(x) - \sum_{x \in \mathsf{X}} P(y|x, f_{\hat{Q}}(x), \mu)\, \mu(x) \right| \\ &\leq \sum_{x \in \mathsf{X}} \|P(\cdot|x, f_Q(x), \mu) - P(\cdot|x, f_{\hat{Q}}(x), \mu)\|_1\, \mu(x) \\ &\leq \sum_{x \in \mathsf{X}} K_1\, \|f_Q(x) - f_{\hat{Q}}(x)\|_1\, \mu(x). \end{aligned} \tag{6}$$

Suppose that Q is of the following form:

$$\begin{aligned} Q(x,u) &= R^{\mathrm{reg}}(x,u,\mu) + \beta \sum_{y \in \mathsf{X}} v(y)\, P(y|x,u,\mu) \\ &= \langle q_x^{\mu,v}, u \rangle - \Omega(u), \end{aligned}$$

where $v : \mathsf{X} \to \mathbb{R}$ and

$$q_x^{\mu,v}(\cdot) := r(x, \cdot, \mu) + \beta \sum_{y \in \mathsf{X}} v(y)\, p(y|x, \cdot, \mu).$$

Note that the mapping $f_Q(x)$ is the unique minimizer of $Q(x,\cdot)$ and $f_{\hat{Q}}(x)$ is the unique minimizer of $\hat{Q}(x,\cdot)$. Let us set $f_Q(x) = u$ and $f_{\hat{Q}}(x) = u'$. Then we have

$$Q(x,u) - Q(x,u')$$

$$
\begin{aligned}
&= \langle q_x^{\mu,v}, u \rangle - \Omega(u) - \langle q_x^{\mu,v}, u' \rangle + \Omega(u') \\
&= \langle q_x^{\mu,v}, u - u' \rangle + \Omega(u') - \Omega(u) \\
&\overset{(1)}{\geq} \langle q_x^{\mu,v}, u - u' \rangle + \langle \nabla\Omega(u), u' - u \rangle + \frac{\rho}{2} \|u - u'\|_2^2 \\
&= \langle \nabla Q(x,u), u - u' \rangle + \frac{\rho}{2} \|u - u'\|_2^2 \\
&= \frac{\rho}{2} \|u - u'\|_2^2,
\end{aligned}
$$

where the last statement follows from the fact that $u = f_Q(x)$ is the unique minimizer of differentiable concave function $Q(x, \cdot)$ (i.e., $\nabla Q(x,u) = 0$) and (1) follows from strong convexity of Ω [36, Definition 3.1]. Since $\|v\|_1 \leq \sqrt{|\mathsf{A}|}\|v\|_2$, we have

$$
\begin{aligned}
&\|f_Q(x) - f_{\hat{Q}}(x)\|_1^2 \leq \frac{2\sqrt{|\mathsf{A}|}}{\rho} \left(Q(x, f_Q(x)) - Q(x, f_{\hat{Q}}(x)) \right) \\
&= \frac{2\sqrt{|\mathsf{A}|}}{\rho} \bigg(Q(x, f_Q(x)) - \hat{Q}(x, f_{\hat{Q}}(x)) \\
&\qquad\qquad + \hat{Q}(x, f_{\hat{Q}}(x)) - Q(x, f_{\hat{Q}}(x)) \bigg) \\
&= \frac{2\sqrt{|\mathsf{A}|}}{\rho} \bigg(\max_{u \in \mathsf{U}} Q(x,u) - \max_{u \in \mathsf{U}} \hat{Q}(x,u) \\
&\qquad\qquad + \hat{Q}(x, f_{\hat{Q}}(x)) - Q(x, f_{\hat{Q}}(x)) \bigg) \\
&\leq \frac{4\sqrt{|\mathsf{A}|}}{\rho} \|Q - \hat{Q}\|_\infty =: \theta \, \|Q - \hat{Q}\|_\infty. \qquad (7)
\end{aligned}
$$

Hence, combining (6) and (7) yields

$$
\|H_2(\mu, Q) - H_2(\mu, \hat{Q})\|_1 \leq \sqrt{\theta}\, K_1 \sqrt{\|Q - \hat{Q}\|_\infty}. \qquad (8)
$$

Using (8), for any $k = 0, \ldots, K-1$, we have

$$
\begin{aligned}
&\|H(\mu_k) - \hat{H}([N,L],M)(\mu_k)\|_1 \\
&\leq \|H_2(\mu_k, H_1(\mu_k)) - H_2(\mu_k, \hat{H}_1[N,L](\mu_k))\|_1 \\
&+ \|H_2(\mu_k, \hat{H}_1[N,L](\mu_k)) - \hat{H}_2[M](\mu_k, \hat{H}_1[N,L](\mu_k))\|_1 \\
&\leq \sqrt{\theta}\, K_1 \sqrt{\|H_1(\mu_k) - \hat{H}_1[N,L](\mu_k)\|_\infty} \\
&+ \|H_2(\mu_k, \hat{H}_1[N,L](\mu_k)) - \hat{H}_2[M](\mu_k, \hat{H}_1[N,L](\mu_k))\|_1.
\end{aligned}
$$

The last term is bounded from above by

$$
K_1 \sqrt{\theta(\varepsilon_1 + \Delta)} + \varepsilon_2
$$

with probability at least $1 - \frac{\delta}{K}$ by Theorem 2 and Theorem 3. Therefore, with probability at least $1 - \delta$

$$
\begin{aligned}
&\|\mu_K - \mu_*\|_1 \\
&\leq \sum_{k=0}^{K-1} K_H^{K-(k+1)} \|\hat{H}([N,L],M)(\mu_k) - H(\mu_k)\|_1 \\
&\qquad\qquad + \|H^K(\mu_0) - \mu_*\|_1 \\
&\leq \sum_{k=0}^{K-1} K_H^{K-(k+1)} \left(K_1 \sqrt{\theta(\varepsilon_1 + \Delta)} + \varepsilon_2 \right) + \frac{(K_H)^K}{1 - K_H} \\
&\leq \frac{K_1 \sqrt{\theta\, \Delta}}{(1 - K_H)} + \varepsilon.
\end{aligned}
$$

This completes the proof. □

Now, we give the main result of this paper as a corollary of Theorem 4. It states that, by using learning algorithm $\hat{H}$, one can obtain approximate regularized-mean-field equilibrium with high confidence. Since approximate regularized mean-field equilibrium constitutes an approximate Nash equilibrium for the finite-agent game model with sufficiently many agents, this learning algorithm also provides approximate Nash equilibrium.

Corollary 1. Fix any $(\varepsilon, \delta) \in (0,1)^2$. Suppose that K, L, N, M satisfy the conditions in Theorem 4. Let (μ_K, Q_K) be the output of the learning algorithm established by random operator $\hat{H}$ with inputs

$$
\left(K, \{[N,L]\}_{k=0}^{K}, \{M\}_{k=0}^{K-1}, \mu_0 \right).
$$

Define $\pi_K(x) := \arg\max_{u \in \mathsf{U}} Q_K(x,u)$. Then, with probability at least $1 - \delta(1 + \frac{1}{2K})$, the pair (π_K, μ_*) is a $\kappa(\varepsilon, \Delta)$-regularized mean-field equilibrium, where

$$
\begin{aligned}
&\kappa(\varepsilon, \Delta) \\
&= 2 \frac{1}{1-\beta} \left(\frac{(1-K_H)^2 \varepsilon^2}{16\, \theta\, (K_1)^2} + \Delta + K_{H_1} \left(\frac{K_1 \sqrt{\theta\, \Delta}}{(1-K_H)} + \varepsilon \right) \right).
\end{aligned}
$$

Therefore, by Theorem 1, an N-tuple of policies $\pi^{(N)} = \{\pi_K, \pi_K, \ldots, \pi_K\}$ is an $\kappa(\varepsilon, \Delta) + \epsilon$-Nash equilibrium for the regularized game with $N \geq N(\epsilon)$ agents.

Proof. By Theorem 4, with probability at least $1 - \delta(1 + \frac{1}{2K})$, we have

$$
\begin{aligned}
&\|Q_K - H_1(\mu_*)\|_\infty \\
&\quad \leq \|Q_K - H_1(\mu_K)\|_\infty + \|H_1(\mu_K) - H_1(\mu_*)\|_\infty \\
&\quad \leq \varepsilon_1 + \Delta + K_{H_1} \|\mu_K - \mu_*\|_1 \\
&\quad \leq \varepsilon_1 + \Delta + K_{H_1} \left(\frac{K_1 \sqrt{\theta\, \Delta}}{(1-K_H)} + \varepsilon \right) \\
&\quad = \frac{(1-K_H)^2 \varepsilon^2}{16\, \theta\, (K_1)^2} + \Delta + K_{H_1} \left(\frac{K_1 \sqrt{\theta\, \Delta}}{(1-K_H)} + \varepsilon \right).
\end{aligned}
$$

Let $\pi_K(x) := \arg\max_{u \in \mathsf{U}} Q_K(x,u)$. By [25, Proposition 3.2], with probability at least $1 - \delta(1 + \frac{1}{2K})$, we have

$$
\begin{aligned}
&\|J_{\mu_*}^{\mathrm{reg}}(\pi_*, \cdot) - J_{\mu_*}^{\mathrm{reg}}(\pi_K, \cdot)\|_\infty \\
&\leq 2 \frac{1}{1-\beta} \left(\frac{(1-K_H)^2 \varepsilon^2}{16\, \theta\, (K_1)^2} + \Delta + K_{H_1} \left(\frac{K_1 \sqrt{\theta\, \Delta}}{(1-K_H)} + \varepsilon \right) \right) \\
&= \kappa(\varepsilon, \Delta).
\end{aligned}
$$

Hence, (π_K, μ_*) is $\kappa(\varepsilon, \Delta)$-mean-field equilibrium with probability at least $1 - \delta(1 + \frac{1}{2K})$. □

Remark 1. In Corollary 1, there is a constant error Δ, which is a function of representation error $E(\mathcal{F})$. If we choose the class of Q-functions $\mathcal{F}$ as $\mathcal{C}$, then there will be no representation error, i.e, $E(\mathcal{F}) = 0$, and so, $\Delta = 0$. Hence, in this case, we have the following error bound:

$$
\kappa(\varepsilon, 0) := 2 \frac{1}{1-\beta} \left(\frac{(1-K_H)^2 \varepsilon^2}{16\, \theta\, (K_1)^2} + K_{H_1} \varepsilon \right),
$$

which goes to zero as $\varepsilon \to 0$.

V. Conclusion

In this paper, we have established a learning algorithm for discrete time regularized mean-field games subject to discounted reward criterion via fitted Q-learning. It is known, at least heuristically and experimentally, that adding regularization term to the one-stage reward function makes the learning algorithm more robust and improves exploration. In addition to these advantages, with regularization term, the error analysis of the learning algorithm has been established under milder assumptions compared to the classical version of the game model.

References

[1] J. Lasry and P.Lions, “Mean field games,” *Japan. J. Math.*, vol. 2, pp. 229–260, 2007.

[2] M. Huang, R. Malhamé, and P. Caines, “Large population stochastic dynamic games: Closed loop McKean-Vlasov systems and the Nash certainty equivalence principle,” *Communications in Information Systems*, vol. 6, pp. 221–252, 2006.

[3] M. Huang, P. Caines, and R. Malhamé, “Large-population cost coupled LQG problems with nonuniform agents: Individual-mass behavior and decentralized ϵ-Nash equilibria,” *IEEE. Trans. Autom. Control*, vol. 52, no. 9, pp. 1560–1571, 2007.

[4] H. Tembine, Q. Zhu, and T. Başar, “Risk-sensitive mean field games,” *IEEE. Trans. Autom. Control*, vol. 59, no. 4, pp. 835–850, 2014.

[5] M. Huang, “Large-population LQG games involving major player: The Nash certainty equivalence principle,” *SIAM J. Control Optim.*, vol. 48, no. 5, pp. 3318–3353, 2010.

[6] A. Bensoussan, J. Frehse, and P. Yam, *Mean Field Games and Mean Field Type Control Theory*. Springer, New York, 2013.

[7] P. Cardaliaguet, *Notes on Mean-field Games*, 2011.

[8] R. Carmona and F. Delarue, “Probabilistic analysis of mean-field games,” *SIAM J. Control Optim.*, vol. 51, no. 4, pp. 2705–2734, 2013.

[9] D. Gomes and J. Saúde, “Mean field games models - a brief survey,” *Dyn. Games Appl.*, vol. 4, no. 2, pp. 110–154, 2014.

[10] J. Moon and T. Başar, “Robust mean field games for coupled Markov jump linear systems,” *International Journal of Control*, vol. 89, no. 7, pp. 1367–1381, 2016.

[11] D. Gomes, J. Mohr, and R. Souza, “Discrete time, finite state space mean field games,” *J. Math. Pures Appl.*, vol. 93, pp. 308–328, 2010.

[12] S. Adlakha, R. Johari, and G. Weintraub, “Equilibria of dynamic games with many players: Existence, approximation, and market structure,” *Journal of Economic Theory*, vol. 156, pp. 269–316, 2015.

[13] R. Elliot, X. Li, and Y. Ni, “Discrete time mean-field stochastic linear-quadratic optimal control problems,” *Automatica*, vol. 49, pp. 3222–3233, 2013.

[14] J. Moon and T. Başar, “Discrete-time decentralized control using the risk-sensitive performance criterion in the large population regime: a mean field approach,” in *ACC 2015*, Chicago, Jul. 2015.

[15] M. Nourian and G. Nair, “Linear-quadratic-Gaussian mean field games under high rate quantization,” in *CDC 2013*, Florence, Dec. 2013.

[16] J. Moon and T. Başar, “Discrete-time mean field Stackelberg games with a large number of followers,” in *CDC 2016*, Las Vegas, Dec. 2016.

[17] A. Biswas, “Mean field games with ergodic cost for discrete time Markov processes,” arXiv:1510.08968, 2015.

[18] P. Wiecek, “Discrete-time ergodic mean-field games with average reward on compact spaces,” *Dynamic Games and Applications*, pp. 1–35, 2019.

[19] P. Wiecek and E. Altman, “Stationary anonymous sequential games with undiscounted rewards,” *Journal of Optimization Theory and Applications*, vol. 166, no. 2, pp. 686–710, 2015.

[20] N. Saldi, “Discrete-time average-cost mean-field games on Polish spaces,” arXiv:1908.08793 (accepted to Turkish Journal of Mathematics), 2019.

[21] N. Saldi, T. Başar, and M. Raginsky, “Approximate Markov-Nash equilibria for discrete-time risk-sensitive mean-field games,” to appear in Mathematics of Operations Research, 2019.

[22] ——, “Markov–Nash equilibria in mean-field games with discounted cost,” *SIAM Journal on Control and Optimization*, vol. 56, no. 6, pp. 4256–4287, 2018.

[23] ——, “Approximate Nash equilibria in partially observed stochastic games with mean-field interactions,” *Mathematics of Operations Research*, vol. 44, no. 3, pp. 1006–1033, 2019.

[24] B. Anahtarci, C. Kariksiz, and N. Saldi, “Value iteration algorithm for mean field games,” arXiv:1909.01758, 2019.

[25] ——, “Fitted Q-learning in mean-field games,” arXiv:1912.13309, 2019.

[26] A. Antos, R. Munos, and C. Szepesvári, “Fitted Q-iteration in continuous action-space MDPs,” in *Proceedings of the 20th International Conference on Neural Information Processing Systems*, 2007, pp. 9–16.

[27] M. Geist, B. Scherrer, and O. Pietquin, “A theory of regularized Markov decision processes,” arXiv:1901.11275, 2019.

[28] G. Neu, A. Jonsson, and V. Gomez, “A unified view of entropy-regularized Markov decision processes,” arXiv:1705.07798, 2017.

[29] X. Guo, A. Hu, R. Xu, and J. Zhang, “Learning mean-field games,” arXiv:1901.09585, 2019.

[30] R. Elie, J. Perolat, M. Lauriere, M. Geist, and O. Pietquin, “Approximate fictitious play for mean-field games,” arXiv:1907.02633, 2019.

[31] R. Carmona, M. Lauriere, and Z. Tan, “Linear-quadratic mean-field reinforcement learning: convergence of policy gradient methods,” arXiv:1910.04295, 2019.

[32] Z. Fu, Z. Yang, Y. Chen, and Z. Wang, “Actor-critic provably finds Nash equilibria of linear-quadratic mean-field games,” arXiv:1910.07498, 2019.

[33] J. Yang, X. Ye, R. Trivedi, X. Hu, and H.Zha, “Learning deep mean field games for modelling large population behaviour,” arXiv:1711.03156, 2018.

[34] L. Kontorovich and K. Ramanan, “Concentration inequalities for dependent random variables via the martingale method,” *The Annals of Probability*, vol. 36, no. 6, pp. 2126–2158, 2008.

[35] A. Antos, R. Munos, and C. Szepesvári, “Fitted Q-iteration in continuous action-space MDPs,” Tech. Rep., 2007, pp.22. inria-00185311v1.

[36] B. Hajek and M. Raginsky, “Statistical learning theory,” *Lecture Notes*, 2019.