Dynamic Games and Applications
https://doi.org/10.1007/s13235-022-00450-2

Q-Learning in Regularized Mean-field Games

Berkay Anahtarci[1] · **Can Deha Kariksiz**[1] · **Naci Saldi**[2]

Accepted: 24 April 2022

Abstract
In this paper, we introduce a regularized mean-field game and study learning of this game under an infinite-horizon discounted reward function. Regularization is introduced by adding a strongly concave 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. The regularization term in general makes reinforcement learning algorithm more robust to the system components. 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.

Keywords Mean-field games · Q-learning · Regularized Markov decision processes · Discounted reward

1 Introduction

This paper deals with the learning of regularized mean-field games (MFGs) under an infinite-horizon discounted reward function. Regularization is introduced by adding a strongly concave regularization function to the one-stage reward function in the classical mean-field game model. In this model, a single agent interacts with a huge population of other agents and competes 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 taken to infinity. In the limiting case, a generic agent faces a single-agent stochastic control problem with a

This article is part of the topical collection "Multi-agent Dynamic Decision Making and Learning" edited by Konstantin Avrachenkov, Vivek S. Borkar and U. Jayakrishnan Nair.

✉ Naci Saldi
naci.saldi@bilkent.edu.tr

Berkay Anahtarci
berkay.anahtarci@ozyegin.edu.tr

Can Deha Kariksiz
deha.kariksiz@ozyegin.edu.tr

[1] Özyeğin University, Çekmeköy, Istanbul, Turkey

[2] Bilkent University, Çankaya, Ankara, Turkey

Published online: 23 May 2022

constraint on the state distribution at each time step. 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 [25], where the standard terminology of mean-field games was introduced, and independently as stochastic dynamic games by Huang, Malhamé and Caines [21], both considering 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 a coupled Hamilton–Jacobi–Bellman (HJB) equation and a Kolmogorov–Fokker–Planck (FPK) equation. We refer the reader to [6, 8, 9, 17, 19, 20, 29, 37] 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 the continuous-time framework, there are comparably fewer results available on discrete-time mean-field games in the literature. These works have mainly studied the settings where the state space is a discrete (finite or countable) set 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 [16], a mean-field game model with finite state is studied, and [1] 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 [12, 27, 28, 31]. References [7, 32, 39, 40] study discrete-time mean-field games subject to the average cost optimality criterion. In [34], authors consider a discrete-time risk-sensitive mean-field game with Polish state and action spaces. References [33, 35] consider a discrete-time mean-field game with Polish state and action spaces under the discounted cost optimality criterion for both the fully observed case and the partially observed case, respectively.

We note that the aforementioned papers, except linear models, mostly identify the existence of mean-field equilibrium and do not propose any algorithm with convergence guarantee to compute the mean-field equilibrium. In our recent work [3], this problem is explored for mean-field games with abstract state and action spaces under both discounted cost and average cost criteria, where we develop a value iteration algorithm and prove that this algorithm converges to the mean-field equilibrium. In [2], we generalize this value iteration algorithm to the model-free setting by using fitted Q-learning [4], which is preferred over a classical Q-learning algorithm since the action space is assumed to be a compact and convex subset of a finite dimensional Euclidean space. In order to establish the contractiveness of the optimality operator in this case, one needs to prove that the optimal policy is Lipschitz continuous with respect to the current mean-field term, since the optimal policy corresponding to the current mean-field term affects the next mean-field term in the value iteration algorithm. Although establishing the Lipschitz continuity of the optimal value function with respect to the mean-field term is straightforward, it is quite challenging to do the same for the optimal policy. To overcome this challenge, it was assumed in [2,Assumption-2.1(d)] that the function in the optimality equation is strongly convex and has Lipschitz continuous gradient, which restricts the applicability of the results. Moreover, as a result of these restrictive conditions, the proof of the contraction of the mean-field equilibrium operator is much more involved. Our novel approach in this paper is to introduce a strongly convex regularization function in the one-stage reward, which helps us to obtain Lipschitz continuity of the optimal policy

with respect to the mean-field term via duality between strong convexity and smoothness, and generalize the results in [2]. This allows us to significantly relax the assumptions on the system components and improve the theoretical analysis. In particular, as opposed to the unregularized case, we eliminate the need for strong convexity and smoothness assumptions on the system components when establishing the Lipschitz regularity of the optimal policy with respect to the mean-field term.

In the literature, the existence of mean-field equilibria has been established for discrete-time mean-field games under the discounted optimality criterion in [33]. However, learning discrete-time mean-field games has not been studied much, even for the classical case, until recently. In [18], authors establish a Q-learning algorithm to compute approximate mean-field equilibria for finite state-action mean-field games, where the convergence of the learning algorithm is dependent upon the assumption that the operators in the algorithm are contractive. In [11], 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, and suggest an error analysis of the learning algorithm for the deterministic game model (no noise term in the state dynamics). In [10], authors study linear-quadratic mean-field games and establish the convergence of policy gradient algorithm. In [13], an actor-critic algorithm to learn mean-field equilibrium for linear-quadratic mean-field games is developed. In [41], 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, and the authors are able to transform a mean-field game into an equivalent deterministic Markov decision process by extending the state and action spaces, establishing classical reinforcement learning algorithms to compute mean-field equilibrium. In the continuous-time setup, the following early reference [42] develops a learning algorithm for mean-field oscillator game model to obtain approximate Nash equilibrium (see also Example in [26,Section IV-C] for learning algorithm developed for continuous-time LQG mean-field game problem).

In misspecified control models, greedy algorithms often result in policies that are far from optimal. Our approach of making use of regularization also provides a way to overcome this problem. Most recent reinforcement algorithms use regularization to increase exploration and robustness, and the regularization is generally established via entropy or relative entropy. We refer the reader to [14] for an exhaustive review of the literature on regularized Markov decision processes (MDPs) and [30] 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, making the algorithm more robust since one can establish the Lipschitz sensitivity of the optimal policy to the system components using duality between strong convexity and smoothness, which is a common necessity in robustness analysis [23,Remark 4.3],[22,Theorem 4.1]. As mentioned above, regularization additionally provides an error analysis of the learning algorithm that is established under quite milder assumptions compared to the unregularized case. Therefore, this work covers a wider range of systems in practice.

The paper is set out as follows. In Sect. 2, we introduce classical and regularized mean-field games as well as finite-agent game, and define the classical and regularized mean-field equilibria. In Sect. 3, we define mean-field equilibrium operator and show that the mean-field equilibrium operator is contractive. In Sect. 5, we establish a Q-learning algorithm to compute approximate regularized-mean-field equilibrium and prove its convergence. In Sect. 6, we

provide a numerical example to illustrate the effectiveness of the learning algorithm. Section 7 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, $\|\cdot\|_1$ denotes l_1-norm on $\mathcal{P}(\mathsf{E})$. Total variation norm on $\mathcal{P}(\mathsf{E})$ is denoted by $\|\cdot\|_{TV}$. For any probability measures $\mu, \nu \in \mathcal{P}(\mathsf{E})$, we have $\|\mu - \nu\|_{TV} = \inf\left\{E^{\xi}[1_{\{x\neq y\}}] : \xi(\cdot, \mathsf{E}) = \mu(\cdot) \text{ and } \xi(\mathsf{E}, \cdot) = \nu(\cdot)\right\}$ 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 $\|\cdot\|_1 = 2\|\cdot\|_{TV}$ [15,p. 141]. 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 [38]. The notation $v \sim \nu$ means that the random element v has distribution ν.

2 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 behaviour of the other agents;[1] that is, μ can be considered as the infinite population limit of the empirical distribution of the states of other agents.

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:

$$\begin{aligned}&|r(x, a, \mu) - r(\hat{x}, \hat{a}, \hat{\mu})|\\ &\quad\leq L_1\left(1_{\{x\neq\hat{x}\}} + 2 \cdot 1_{\{a\neq\hat{a}\}} + \|\mu - \hat{\mu}\|_1\right), \forall x, \hat{x}, \forall a, \hat{a}, \forall \mu, \hat{\mu}.\end{aligned}$$

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

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

[1] In classical mean-field game literature, the exogenous behaviour of the other agents is in general modelled by a state-measure flow $\{\mu_t\}$, $\mu_t \in \mathcal{P}(\mathsf{X})$ for all t, which means that total population behaviour is non-stationary. In this paper, we only consider the stationary case; that is, $\mu_t = \mu$ for all t. Establishing a learning algorithm for the non-stationary case is more challenging.

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:

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

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 and π, the states and actions are evolved as follows:

$$x(t) \sim P(\cdot\,|x(t-1), u(t-1), \mu),\ t \geq 1,$$
$$u(t) = \pi(x(t)),\ t \geq 0.$$

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:*

$$|R(x, u, \mu) - R(\hat{x}, \hat{u}, \hat{\mu})| \leq L_1 \left(1_{\{x \neq \hat{x}\}} + \|u - \hat{u}\|_1 + \|\mu - \hat{\mu}\|_1\right),$$
$$\|P(\cdot|x, u, \mu) - P(\cdot|\hat{x}, \hat{u}, \hat{\mu})\|_1 \leq K_1 \left(1_{\{x \neq \hat{x}\}} + \|u - \hat{u}\|_1 + \|\mu - \hat{\mu}\|_1\right),$$

$\forall x, \hat{x}, \forall u, \hat{u}, \forall \mu, \hat{\mu}$.

Proof The proof is in "Appendix 8.2". □

In the following section, we introduce regularized mean-field games and the adapted optimality notion.

2.1 Regularized Mean-Field Games

A theory of regularized Markov decision processes (MDPs) has been introduced in [14]. 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 [14] 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 differentiable ρ-strongly convex function with respect to the l_1-norm $\|\cdot\|_1$ (see "Appendix 8.1" for definition). 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

$$R^{\text{reg}}(x, u, \mu) := R(x, u, \mu) - \Omega(u).$$

A typical example for Ω is the negative entropy $\Omega(u) = \sum_{a \in \mathsf{A}} \ln(u(a))\, u(a)$. Another similar 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 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. This is indeed the main motivation here for introducing the regularization term.

Now, it is time to define the optimality notion that is adapted in this paper. To this end, we first define the regularized discounted cost of any policy given any state measure.

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

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

where $\beta \in (0, 1)$ is the discount factor and x is the initial state. For this model, we define the set-valued mapping $\Psi^{\mathrm{reg}} : \mathcal{P}(\mathsf{X}) \to 2^{\Pi}$ as follows (here, 2^{Π} is the collection of all subsets of Π):

$$\Psi^{\mathrm{reg}}(\mu) = \{\hat{\pi} \in \Pi : J_\mu^{\mathrm{reg}}(\hat{\pi}, x) = \sup_{\pi} J_\mu^{\mathrm{reg}}(\pi, x) \ \text{ for all } \ x \in \mathsf{X}\}.$$

The set $\Psi^{\mathrm{reg}}(\mu)$ is the set of optimal policies for μ. 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 an 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, $\Lambda^{\mathrm{reg}}(\pi)$ is always non-empty. This can be established via Kakutani's fixed point theorem (see [2,Lemma 3]). Then, the notion of equilibrium for this regularized game model is defined as follows.

Definition 1 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; that is the transition probability P and the one-stage reward function R are not available to the decision maker. To that end, we define the following.

Definition 2 Let $(\pi_*, \mu_*) \in \Pi \times \mathcal{P}(\mathsf{X})$ be a *regularized mean-field equilibrium*. A policy $\pi_\varepsilon \in \Pi$ is an ε-regularized-mean-field equilibrium policy if

$$\sup_{x\in\mathsf{X}} \|\pi_\varepsilon(x) - \pi_*(x)\|_1 \le \varepsilon.$$

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 policy with high confidence. This learned approximate regularized-mean-field equilibrium policy can then be used in finite-agent game model as an approximate Nash equilibrium.

3 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 policy. To that end, in addition to Assumption 1, we assume the following. This assumption ensures that the MFE operator is contractive.

Assumption 2 We assume that

$$\frac{3\,K_1}{2}\left(1+\frac{1}{\rho}\,\frac{K_{\text{Lip}}}{1-\beta}\right)<1,$$

where

$$K_{\text{Lip}}:=\frac{L_1}{1-\beta\,K_1/2}>0.$$

Recall that 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

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

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

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

It is also a well-known fact that $Q_{\mu,\max}^{\text{reg},*}$ satisfies the following Bellman optimality equation:

$$\begin{aligned}Q_{\mu,\max}^{\text{reg},*}(x)&=\sup_u\left[R^{\text{reg}}(x,u,\mu)+\beta\sum_{y\in\mathsf{X}}Q_{\mu,\max}^{\text{reg},*}(y)\,P(y|x,u,\mu)\right]\\&=:\,T_\mu\,Q_{\mu,\max}^{\text{reg},*}(x).\end{aligned}$$

Here, L_μ and T_μ are $\|\cdot\|_\infty$-contractions with contraction factor β, and the unique fixed point of L_μ is $Q_\mu^{\text{reg},*}$ and the unique fixed point of T_μ is $Q_{\mu,\max}^{\text{reg},*}$.

Let $\mathcal{C}$ denote the set of all Q-functions satisfying the following properties: any $Q\in\mathcal{C}$ is uniformly $\left(K_{\text{Lip}}+L_{\text{reg}}\right)$- Lipschitz continuous and ρ-strongly concave with respect to u. We endow $\mathcal{C}$ with the sup-norm $\|\cdot\|_\infty$ throughout the paper.

Lemma 1 *For any μ, $Q^{\text{reg},*}_{\mu,\max}$ is K_{Lip}-Lipschitz continuous; that is,*

$$|Q^{\text{reg},*}_{\mu,\max}(x) - Q^{\text{reg},*}_{\mu,\max}(y)| \leq K_{\text{Lip}}\, 1_{\{x \neq y\}}.$$

Proof The proof is in "Appendix 8.3". □

Now, we define the MFE operator. To that end, we define $H_1 : \mathcal{P}(\mathsf{X}) \to \mathcal{C}$ as $H_1(\mu) = Q^{\text{reg},*}_{\mu}$ (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:

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

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

Lemma 2 *The mapping H_1 is a Lipschitz continuous with the Lipschitz constant* $K_{H_1} := \dfrac{K_{\text{Lip}}}{1-\beta}$.

Proof The proof is in "Appendix 8.4" □

Before we prove that H is contractive, we establish that for any mean-field term, the optimal policy is Lipschitz continuous with respect to the mean-field term.

Lemma 3 *For any $\mu, \hat{\mu}$, let f_μ and $f_{\hat{\mu}}$ denote the corresponding optimal policies. Then, it follows that*

$$\|f_\mu(x) - f_{\hat{\mu}}(\hat{x})\|_1 \leq \frac{1}{\rho} K_{H_1}(1_{\{x \neq \hat{x}\}} + \|\mu - \hat{\mu}\|_1),$$

for all $x, \hat{x}$.

Proof The proof is in "Appendix 8.5" □

Remark 1 In the absence of the regularization term, one can establish the Lipschitz continuity of the optimal policy with respect to the mean-field term if it is assumed that the following function

$$F : (x, v, \mu, u) \mapsto R(x, u, \mu) + \beta \sum_{y} v(y)\, P(y|x, u, \mu)$$

is strongly concave with respect to u and has a Lipschitz continuous gradient in u with respect to x, v, μ, which are in general restrictive conditions. Indeed, these conditions were imposed in our previous work [2,Assumption 2.1(d)] on learning unregularized mean-field games. As a result of these restrictive conditions, the analysis of the convergence of the algorithm is much more involved. Therefore, introducing a regularization term into the one-stage reward function significantly relaxes these conditions on the system components and simplifies the analysis. Moreover, because of the Lipschitz sensitivity of the optimal policy, the algorithm is supposed to be more robust to the uncertainties in the environment.

Note that in classical algorithms developed for MDPs, such as Q-learning, value iteration, and policy iteration, it is not required to establish the Lipschitz continuity of the optimal policy.

However, in mean-field games, since the optimal policy f_μ directly affects the behaviour of the next mean-field term through

$$H_2(\mu, Q_\mu^{\text{reg},*})(\cdot) = \sum_x P(\cdot|x, f_\mu(x), \mu)\, \mu(x),$$

one must also establish the Lipschitz continuity of the optimal policy f_μ in mean-field games. This is indeed the most challenging part in the analysis compared to the analysis of the algorithms developed for MDPs.

Now, we can prove using Lemmas 2 and 3, that H is contractive.

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

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

Proof Now, fix any $\mu, \hat{\mu} \in \mathcal{P}(\mathsf{X})$. Using Lemma 3, we have

$$\begin{aligned}
&\|H_2(\mu, H_1(\mu)) - H_2(\hat{\mu}, H_1(\hat{\mu}))\|_1 \\
&= \sum_y \left| \sum_x P(y|x, f_\mu(x), \mu)\, \mu(x) - \sum_x P(y|x, f_{\hat{\mu}}(x), \hat{\mu})\, \hat{\mu}(x) \right| \\
&\le \sum_y \left| \sum_x P(y|x, f_\mu(x), \mu)\, \mu(x) - \sum_x P(y|x, f_{\hat{\mu}}(x), \hat{\mu})\, \mu(x) \right| \\
&\quad + \sum_y \left| \sum_x P(y|x, f_{\hat{\mu}}(x), \hat{\mu})\, \mu(x) - \sum_x P(y|x, f_{\hat{\mu}}(x), \hat{\mu})\, \hat{\mu}(x) \right| \\
&\overset{(I)}{\le} \sum_x \left\| P(\cdot|x, f_\mu(x), \mu) - P(\cdot|x, f_{\hat{\mu}}(x), \hat{\mu}) \right\|_1 \mu(x) + \frac{K_1}{2}\left(1 + \frac{K_{H_1}}{\rho}\right) \|\mu - \hat{\mu}\|_1 \\
&\le K_1 \left(\sup_x \|f_\mu(x) - f_{\hat{\mu}}(x)\|_1 + \|\mu - \hat{\mu}\|_1 \right) + \frac{K_1}{2}\left(1 + \frac{K_{H_1}}{\rho}\right) \|\mu - \hat{\mu}\|_1 \\
&\le \frac{3\,K_1}{2}\left(1 + \frac{K_{H_1}}{\rho}\right) \|\mu - \hat{\mu}\|_1. \qquad (1)
\end{aligned}$$

Note that Lemma 3 and Proposition 1 lead to

$$\|P(\cdot|x, f_{\hat{\mu}}(x), \hat{\mu}) - P(\cdot|y, f_{\hat{\mu}}(y), \hat{\mu})\|_1 \le K_1 \left(1 + \frac{K_{H_1}}{\rho}\right) 1_{\{x \ne y\}}.$$

Hence, (I) follows from [24,Lemma A2]. This completes the proof. □

Under Assumptions 1 and 2, H is a contraction mapping. Therefore, by Banach Fixed Point Theorem, H has a unique fixed point. Let μ_* be this unique fixed point and $Q_{\mu_*}^{\text{reg},*} = H_1(\mu_*)$. Let $\pi_*(x) = f_{\mu_*}(x)$. 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; that is the transition probability P and the one-stage reward function R are not available to the decision maker, 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.

4 Finite Agent Game

The regularized mean-field game model in Sect. 2 is indeed the infinite-population limit of the regularized finite-agent game model that will be described below. In a 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:

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

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 $(u_1^N(t), \ldots, u_N^N(t))$:

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

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

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

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

Definition 3 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 [33,pp. 4259]). Therefore, it is of interest to obtain an approximate Nash equilibrium, whose definition is given below.

Definition 4 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 model in Sect. 2). Indeed, one can prove that if each agent in the finite-agent game model adopts the ε-regularized-mean-field equilibrium policy in Definition 2 of the infinite population limit, 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.

Before we state the theorem, let us define the following constants:

$$C_1 := \left(\frac{3\,K_1}{2} + \frac{K_1\,K_{H_1}}{2\rho}\right),\ C_2 := \left(L_1 + \frac{\beta K_1 K_{\text{Lip}}}{2}\right)\frac{K_1}{1-C_1}$$
$$C_3 := \left(L_1 + L_{\text{reg}} + \frac{\beta K_1 K_{\text{Lip}}}{2}\right).$$

Note that by Assumption 2, the constant C_1 is strictly less than 1.

Theorem 1 *Let π_ε be an ε-regularized-mean-field equilibrium policy for the mean-field equilibrium (π_*, μ_*). Let $\mu_0 \in \Lambda^{\text{reg}}(\pi_\varepsilon)$. Then, for any $\delta > 0$, there exists a positive integer $N(\delta)$, such that, for each $N \geq N(\delta)$, the N-tuple of policies $\boldsymbol{\pi}^{(N)} = \{\pi_\varepsilon, \pi_\varepsilon, \ldots, \pi_\varepsilon\}$ is an $(\tau\,\varepsilon + \delta)$-Nash equilibrium for the game with N agents, where $\tau := \frac{2C_2+C_3}{1-\beta}$.*

Proof The proof is in "Appendix 8.6". □

In the next section, we develop an algorithm for learning ε-regularized-mean-field equilibrium policy via fitted Q-iteration and empirical estimation of the transition probability. Therefore, if each agent in the finite-agent game model adopts this learned policy, then the resulting policy will be an approximate Nash equilibrium for finite-agent setup.

5 Q-Learning Algorithm

In this section, we establish an offline learning algorithm for obtaining approximate regularized mean-field equilibrium. We suppose that a generic agent has access to a simulator, which generates a new state $y \sim P(\cdot\,|x, u, \mu)$ and gives the reward $R(x, u, \mu)$ for any given state x, action u, and state measure μ. This is a typical assumption in offline reinforcement learning algorithms.

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 $Q_\mu^{\text{reg},*}$ for a given μ is learned via fitted Q-learning algorithm, which has been introduced in [4] 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 $\mathcal{F}$. 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 $\hat{H}_2$.

Remark 2 Note that this learning algorithm can be applied to finite-agent game problem as follows. First of all, we must assume that each agent has access to a simulator, which

generates a new state $y \sim P(\cdot|x,u,\mu)$ and gives the reward $R(x,u,\mu)$ for any given state x, action u, and state measure μ. This is a typical assumption in offline reinforcement learning algorithms. Using this simulator, each agent runs the proposed learning algorithm offline to compute an approximate regularized mean-field equilibrium policy. Agents then have to agree on learned approximate regularized mean-field equilibrium policies via running some consensus algorithm. Then, the resulting joint policy will be approximate Nash equilibrium by Theorem 1.

We now proceed by giving the description of $\hat{H}_1$ first. Let ν be a probability measure on X such that $\min_{x\in\mathsf{X}} \nu(x) > 0$. Define $\xi_0 := 1/\sqrt{\min_{x\in\mathsf{X}} \nu(x)}$. 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. With an abuse of notation, 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 1 Algorithm $\hat{H}_1$

Inputs $([N,L],\mu)$
generate i.i.d. samples $\{(x_t,u_t,r_t,y_{t+1})_{t=1}^N\}$ using

$$x_t \sim \nu,\ u_t \sim \pi_b(x_t)(\cdot),\ r_t = R^{\text{reg}}(x_t,u_t,\mu),\ y_{t+1} \sim P(\cdot|x_t,u_t,\mu)$$

Start with $Q_0 = 0$
for $l = 0,\dots,L-1$ **do**

$$Q_{l+1} = \arg\min_{f\in\mathcal{F}} \frac{1}{N}\sum_{t=1}^{N} \frac{1}{m(\mathsf{U})\,\pi_b(x_t,u_t)} \left| f(x_t,u_t) - \left[r_t + \beta \max_{u'\in\mathsf{U}} Q_l(y_{t+1},u') \right] \right|^2$$

end for
return Q_L

Remark 3 Note that in Algorithm 1, one can alternatively use sample path $\{x_t,u_t\}_{t=1}^N$ generated by the policy π_b instead of using i.i.d. samples. In this case, under π_b, the state process $\{x_t\}$ should assumed to be strictly stationary and exponentially β-mixing [4]. Since exponentially β-mixing stationary processes forget its past exponentially fast, when there is a sufficiently large time difference between two samples, they behave like i.i.d. processes. Therefore, this makes the error analysis of the exponential β-mixing case almost the same as the i.i.d. case. However, the main problem in β-mixing case is finding a policy π_b satisfying this mixing condition. We refer the reader to [4, 5] for the details of the error analysis of $\hat{H}_1$ in exponentially β-mixing case.

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 $m(B(u,\xi)\cap\mathsf{U}) \geq \min\{\alpha\, m(B(u,\xi)), m(\mathsf{U})\}$, where m is the Lebesgue measure on U when considered as a subset of $\mathbb{R}^{|\mathsf{A}|-1}$ (see [5]). For any $Q \in \mathcal{C}$, we define v-norm of Q as

$$\|Q\|_v := \left[\sum_x \int_{\mathsf{U}} |Q(x,u)|^2 \frac{m(du)}{m(\mathsf{U})}\, v(x) \right]^{1/2}.$$

We also define $r_{\mathbf{m}} := \sup_{(x,u,\mu)\in \mathsf{X}\times\mathsf{U}\times\mathcal{P}(\mathsf{X})} |R^{\mathrm{reg}}(x,u,\mu)|$ and $Q_{\mathbf{m}} := r_{\mathbf{m}}/(1-\beta)$. Using these, we need to define the following constants:

$$
\begin{aligned}
E(\mathcal{F}) &:= \sup_{\mu\in\mathcal{P}(\mathsf{X})} \sup_{Q\in\mathcal{F}} \inf_{Q'\in\mathcal{F}} \|Q' - H_\mu Q\|_v \\
L_{\mathbf{m}} &:= (1+\beta) Q_{\mathbf{m}} + r_{\mathbf{m}}, \quad C := \frac{L_{\mathbf{m}}^2}{m(\mathsf{U})\,\pi_0} \\
\Upsilon &= 8\,e^2\,(V_{\mathcal{F}}+1)\,(V_{\mathcal{F}_{\max}}+1)\left(\frac{64 e Q_{\mathbf{m}} L_{\mathbf{m}}(1+\beta)}{m(\mathsf{U})\pi_0}\right)^{V_{\mathcal{F}}+V_{\mathcal{F}_{\max}}} \\
V &= V_{\mathcal{F}} + V_{\mathcal{F}_{\max}}, \quad \gamma = 512 C^2 \\
\Delta &:= \frac{1}{1-\beta}\left[\frac{m(\mathsf{U})|\mathsf{A}|!\xi_0}{\alpha(2/(K_{\mathrm{Lip}}+L_{\mathrm{reg}}))^{|\mathsf{A}|-1}}\,E(\mathcal{F})\right]^{\frac{1}{|\mathsf{A}|}} \\
\Lambda &:= \frac{1}{1-\beta}\left[\frac{m(\mathsf{U})|\mathsf{A}|!\xi_0}{\alpha(2/(K_{\mathrm{Lip}}+L_{\mathrm{reg}}))^{|\mathsf{A}|-1}}\right]^{\frac{1}{|\mathsf{A}|}}.
\end{aligned}
$$

Here, $E(\mathcal{F})$ gives the representation error of the function class $\mathcal{F}$. This error in general is zero or very small, since any Q function in $\mathcal{C}$ can be approximated very well via, for instance, neural networks with some fixed architecture. Hence, we can think of the error due to $E(\mathcal{F})$ negligible. The following theorem gives the error analysis of the algorithm $\hat{H}_1$.

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

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

if $\frac{\beta^L}{1-\beta}\,Q_{\mathbf{m}} < \frac{\varepsilon}{2}$ and $N \ge m_1(\epsilon,\delta,L)$, 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, which is in general negligible.

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

Algorithm 2 Algorithm $\hat{H}_2$

Inputs (M,μ,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 ([2,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*

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

Algorithm 3 provides the overall description of the algorithm $\hat{H}$, which replaces the MFE operator H.

Algorithm 3 Algorithm $\hat{H}$

Inputs $\left(K, \{[N_k, L_k]\}_{k=0}^{K}, \{M_k\}_{k=0}^{K-1}, \mu_0\right)$
Start with μ_0
for $k = 0, \ldots, K-1$ **do**

$$\begin{aligned} \mu_{k+1} &= \hat{H}\left([N_k, L_k], M_k\right)(\mu_k) \\ &:= \hat{H}_2[M_k]\left(\mu_k, \hat{H}_1[N_k, L_k](\mu_k)\right) \end{aligned}$$

end for
return μ_K

Note that in Algorithm 3, for each stage $k = 0, \ldots, K-1$, the input is μ_k. In addition, we also pick integers N_k and L_k as inputs for the random operator $\hat{H}_1$ and pick integer M_k as an input for the random operator $\hat{H}_2$ at each stage. We first compute an approximate Q-function for μ_k via $\hat{H}_1[N_k, L_k](\mu_k)$ and we compute an approximate new mean-field term via $\hat{H}_2[M_k]\left(\mu_k, \hat{H}_1[N_k, L_k](\mu_k)\right)$. In the second stage of the iteration, since we are using an approximate Q-function instead of the exact Q-function, we also have an error due to $\hat{H}_1$ in addition to the error resulting from $\hat{H}_2$.

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*

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

where $\theta := \frac{4}{\rho}$*. Let* K, L *be such that*

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

Then, pick N, M *such that*

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

Let μ_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.

Proof Note that for any $\mu \in \mathcal{P}(\mathsf{X})$, $Q \in \mathcal{C}$,and $\hat{Q} \in \mathcal{F}$, we have

$$\begin{aligned}
&\|H_2(\mu, Q) - H_2(\mu, \hat{Q})\|_1 \\
&= \sum_{y\in\mathsf{X}} \left| \sum_{x\in\mathsf{X}} P(y|x, f_Q(x), \mu)\,\mu(x) - \sum_{x\in\mathsf{X}} P(y|x, f_{\hat{Q}}(x), \mu)\,\mu(x) \right| \\
&\leq \sum_{x\in\mathsf{X}} \|P(\cdot|x, f_Q(x), \mu) - P(\cdot|x, f_{\hat{Q}}(x), \mu)\|_1\,\mu(x) \\
&\leq \sum_{x\in\mathsf{X}} K_1\,\|f_Q(x) - f_{\hat{Q}}(x)\|_1\,\mu(x). \qquad (3)
\end{aligned}$$

Suppose that Q is of the following form:

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

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 maximizer of $Q(x,\cdot)$ and $f_{\hat{Q}}(x)$ is the maximizer of $\hat{Q}(x,\cdot)$. Let us set $f_Q(x) = u$ and $f_{\hat{Q}}(x) = u'$. Then, it follows that

$$\begin{aligned}
Q(x,u) - Q(x,u') &= \langle q_x^{\mu,v}, u\rangle - \Omega(u) - \langle q_x^{\mu,v}, u'\rangle + \Omega(u') \\
&= \langle q_x^{\mu,v}, u-u'\rangle + \Omega(u') - \Omega(u) \\
&\overset{(I)}{\geq} \langle q_x^{\mu,v}, u-u'\rangle + \langle \nabla\Omega(u), u'-u\rangle + \frac{\rho}{2}\,\|u-u'\|_1^2 \\
&= \langle \nabla Q(x,u), u-u'\rangle + \frac{\rho}{2}\,\|u-u'\|_1^2 \\
&\overset{(II)}{\geq} \frac{\rho}{2}\,\|u-u'\|_1^2,
\end{aligned}$$

where (I) follows from strong convexity of Ω with respect to l_1-norm and (II) follows from first-order optimality condition for differentiable concave functions. Now, we have

$$\begin{aligned}
&\|f_Q(x) - f_{\hat{Q}}(x)\|_1^2 \\
&\leq \frac{2}{\rho}\left(Q(x, f_Q(x)) - Q(x, f_{\hat{Q}}(x))\right)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{2}{\rho}\Big(Q(x, f_Q(x)) - \hat{Q}(x, f_{\hat{Q}}(x)) + \hat{Q}(x, f_{\hat{Q}}(x)) - Q(x, f_{\hat{Q}}(x))\Big) \\
&= \frac{2}{\rho}\Big(\max_{u\in \mathsf{U}} Q(x,u) - \max_{u\in \mathsf{U}} \hat{Q}(x,u) + \hat{Q}(x, f_{\hat{Q}}(x)) - Q(x, f_{\hat{Q}}(x))\Big) \\
&\le \frac{4}{\rho}\,\|Q-\hat{Q}\|_\infty =: \theta\,\|Q-\hat{Q}\|_\infty. \qquad (4)
\end{aligned}
$$

Hence, combining (3) and (4) yields

$$
\|H_2(\mu, Q) - H_2(\mu, \hat{Q})\|_1 \le \sqrt{\theta}\,K_1\sqrt{\|Q-\hat{Q}\|_\infty}. \qquad (5)
$$

Using (5), for any $k = 0, \ldots, K-1$, we have

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

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 Theorems 2 and 3. Therefore, with probability at least $1-\delta$

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

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 policy with high confidence. Since approximate regularized mean-field equilibrium policy 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 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

$$\pi_K(x) := \arg\max_{u \in U} Q_K(x, u),$$

where $Q_K = \hat{H}_1([N, L])(\mu_K)$. *Then, with probability at least* $1 - \delta(1 + \frac{1}{2K})$, *the policy* π_K *is a* $\kappa(\varepsilon, \Delta)$*-regularized mean-field equilibrium policy, where*

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

that is

$$\sup_{x \in X} \|\pi_K(x) - \pi_*(x)\|_1 \leq \kappa(\varepsilon, \Delta).$$

Therefore, with probability at least $1 - \delta(1 + \frac{1}{2K})$, *by Theorem 1, an* N*-tuple of policies* $\mathbf{\mathfrak{B}}^{(N)} = \{\pi_K, \pi_K, \ldots, \pi_K\}$ *is an* $\tau\, \kappa(\varepsilon, \Delta) + \sigma$*-Nash equilibrium for the regularized game with* $N \geq N(\sigma)$ *agents if* $\mu_0 \in \Lambda^{\text{reg}}(\pi_K)$.

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

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

Let $\pi_K(x) := \arg\max_{u \in U} Q_K(x, u)$. Using the same analysis that leads to (4), we can obtain the following bound:

$$\sup_{x \in X} \|\pi_K(x) - \pi_*(x)\|_1^2 \leq \theta \|Q_K - H_1(\mu_*)\|_\infty.$$

Hence, with probability at least $1 - \delta(1 + \frac{1}{2K})$, the policy π_K is a $\kappa(\varepsilon, \Delta)$-regularized mean-field equilibrium, where

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

This completes the proof. □

Remark 4 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:

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

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

6 Numerical Example

In this section, we show the effectiveness of the learning algorithm with a numerical example. In this example, we consider a mean-field game with a binary state space $\mathsf{X} = \{0, 1\}$ and a binary action space $\mathsf{A} = \{0, 1\}$. The transition probability $p : \mathsf{X} \times \mathsf{A} \to \mathcal{P}(\mathsf{X})$ is independent of the mean-field term and is given by

$$p(1|0, 0) = \eta, \;\; p(1|1, 0) = 1 - \alpha,$$
$$p(1|0, 1) = \kappa, \;\; p(1|1, 1) = 1 - \xi.$$

The one-stage reward function $r : \mathsf{X} \times \mathsf{A} \times \mathcal{P}(\mathsf{X}) \to [0, \infty)$ depends on the mean-field term and is defined as

$$r(x, a, \mu) = \tau(1 - \langle\mu\rangle)(1 - x) + \lambda\langle\mu\rangle(1 - a),$$

where $\langle\mu\rangle$ is the mean of the distribution μ on X. This model satisfies Assumption 1 with

$$L_1 = \max\left\{\tau, \frac{\tau + \lambda}{2}\right\}$$
$$K_1 = \max\left\{2|1 - \alpha - \eta|, |\eta - \kappa|, \frac{2}{3}|1 - \xi - \eta|, \frac{2}{3}|1 - \alpha - \kappa|, 2|1 - \xi - \kappa|, |\xi - \alpha|\right\}.$$

In the equivalent game model, the action space becomes $\mathsf{U} = \mathcal{P}(\mathsf{A})$. With this new action space, the new transition probability $P : \mathsf{X} \times \mathsf{U} \to \mathcal{P}(\mathsf{X})$ is given by

$$P(\cdot|x, u) = p(\cdot|x, 0)\, u(0) + p(\cdot|x, 1)\, u(1),$$

and the new one-stage reward function $R : \mathsf{X} \times \mathsf{U} \times \mathcal{P}(\mathsf{X}) \to [0, \infty)$ is given by

$$R(x, u, \mu) = \tau(1 - \langle\mu\rangle)(1 - x) + \lambda\langle\mu\rangle u(0).$$

The regularization function $\Omega : \mathcal{P}(\mathsf{X}) \to \mathbb{R}$ is taken as the weighted negative binary entropy:

$$\Omega(u) = \gamma \;\left(\log(u(0))\, u(0) + \log(u(1))\, u(1)\right).$$

Therefore, the regularized one-stage reward function is

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

Note that Ω is a γ-strongly convex function with respect to the l_1-norm.

For numerical results, we use the following values of the parameters:

$$\eta = 0.6,\; \alpha = 0.3,\; \kappa = 0.7,\; \xi = 0.2$$
$$\tau = 0.2,\; \lambda = 0.2,\; \gamma = 0.15,\; \beta = 0.2.$$

With these parameters, Lipschitz constants in Assumption 1 become $L_1 = 0.2$ and $K_1 = 0.2$. Using these constants, $\rho = \gamma = 0.15$, and $\beta = 0.2$, one can also verify that Assumption 2 holds. We run the learning algorithm 20 times using the following parameters: $N = 1000$, $L = 10$, $M = 1000$, $K = 20$ and take the average of the outputs. Here, output of the learning algorithm contains the mean-field term, mean-field policy, and corresponding value function. In fitted Q-learning algorithm, we pick the function class $\mathcal{F}$ as two-layer neural networks with 20 hidden units. We use neural network fitting tool of MATLAB. In particular, we use "fitnet", "train", and "net" functions of MATLAB, where "Levenberg-Marquardt" is picked as the training algorithm and the transfer function is chosen as "hyperbolic tangent sigmoid transfer function". The parameters of the neural network fitting tool of MATLAB are

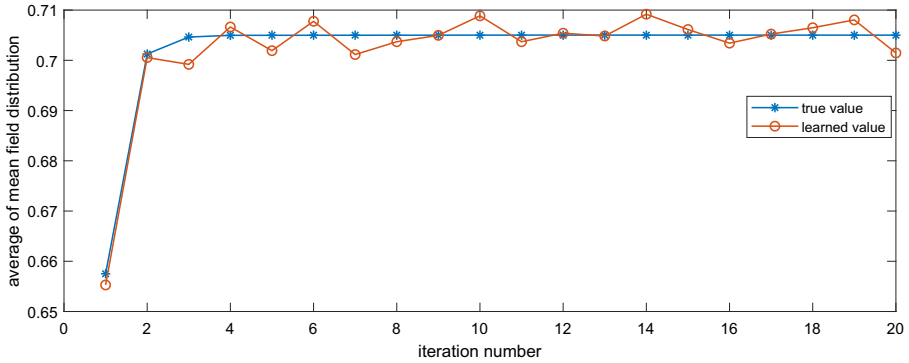

Fig. 1 Comparison of mean-field terms

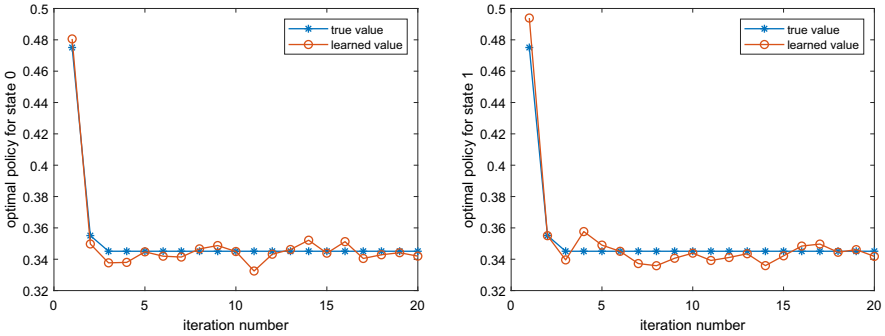

Fig. 2 Comparison of policies

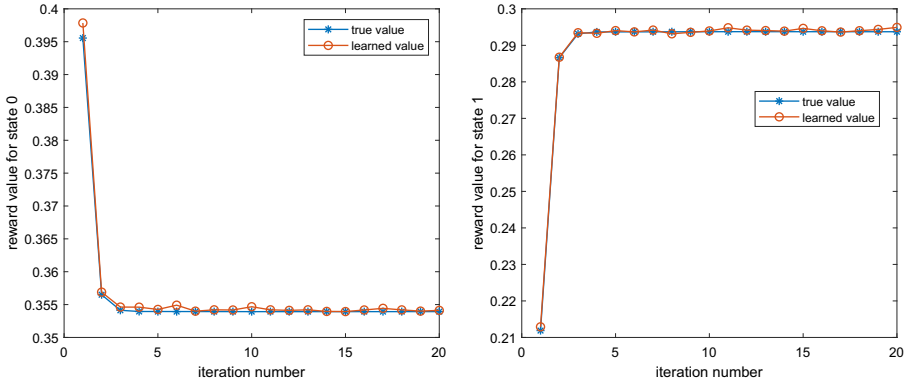

Fig. 3 Comparison of rewards

set to default values. We also run the value iteration algorithm using MFE operator H to find the correct mean-field term, mean-field policy, and corresponding value function. Then, we compare the learned outputs with correct outputs. Figures 1, 2, and 3 show this comparison. It can be seen that learned outputs converge to the true outputs.

7 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 supposed 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.

One interesting future direction is to study learning regularized mean-field games with abstract observation and action spaces. In this case, to obtain similar results, one needs to extend duality of strong convexity and smoothness to the functions defined on infinite dimensional spaces such as the set of probability measures on abstract spaces.

Acknowledgements This work was partly supported by the BAGEP Award of the Science Academy.

Funding Funding was provided by Bilim Akademisi (Grant No. BAGEP 2021).

8 Appendix

8.1 Duality of Strong Convexity and Smoothness

Suppose that $\mathsf{E} = \mathbb{R}^d$ for some $d \geq 1$ with an inner product $\langle \cdot, \cdot \rangle$. We denote $\mathbb{R}^* = \mathbb{R} \bigcup \{\infty\}$. Let $f : \mathsf{E} \to \mathbb{R}^*$ be a differentiable convex function with the domain $S := \{x \in \mathsf{E} : f(x) \in \mathbb{R}\}$, which is necessarily convex subset of E. The Fenchel conjugate of f is a convex function $f^* : \mathsf{E} \to \mathbb{R}^*$ that is defined as

$$f^*(y) := \sup_{x \in S} \langle x, y \rangle - f(x).$$

Now, we will state duality result between strong convexity and smoothness. To this end, we suppose that f is ρ-strongly convex with respect to a norm $\| \cdot \|$ on E (not necessarily Euclidean norm); that is, for all $x, y \in S$, we have

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2}\rho \|y - x\|^2.$$

To state the result, we need to define the dual norm of $\| \cdot \|$. The dual norm $\| \cdot \|_*$ of $\| \cdot \|$ on E is defined as

$$\|z\|_* := \sup\{\langle z, x \rangle : \|x\| \leq 1\}.$$

For example, $\| \cdot \|_\infty$ is the dual norm of $\| \cdot \|_1$.

Proposition 3 ([36,Lemma 15]) *Let $f : E \to \mathbb{R}^*$ be a differentiable ρ-strongly convex function with respect to the norm $\| \cdot \|$ and let S denote its domain. Then,*

1. *f^* is differentiable on E.*
2. *$\nabla f^*(y) = \arg\max_{x \in S} \langle x, y \rangle - f(x)$.*
3. *f^* is $\frac{1}{\rho}$-smooth with respect to the norm $\| \cdot \|_*$; that is,*

$$\|\nabla f^*(y_1) - \nabla f^*(y_2)\| \leq \frac{1}{\rho}\|y_1 - y_2\|_* \text{ for all } y_1, y_2 \in \mathsf{E}.$$

In the paper, we make use of the properties 2 and 3 of Proposition 3 to establish the Lipschitz continuity of the optimal policies, which enables us to prove the main results of our paper.

8.2 Proof of Proposition 1

Fix any $x, \hat{x}, u, \hat{u}, \mu, \hat{\mu}$. Let us recall the following fact about l_1 norm on the set probability distributions on finite sets [15,p. 141]. Suppose that there exists a real valued function F on a finite set E. Let $\lambda(F) := \sup_{e\in\mathsf{E}} F(e) - \inf_{e\in\mathsf{E}} F(e)$. Then, for any pair of probability distributions μ, ν on E, we have

$$\left|\sum_e F(e)\,\mu(e) - \sum_e F(e)\,\nu(e)\right| \le \frac{\lambda(F)}{2}\,\|\mu-\nu\|_1. \tag{6}$$

Using this fact, we now have

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

where the last inequality follows from the following fact in view of (6):

$$\begin{aligned}
\sup_a r(x,a,\mu) - \inf_a r(x,a,\mu) &:= r(x,a_{\max},\mu) - r(x,a_{\min},\mu) \\
&\le 2L_1\,1_{\{a_{\max}\neq a_{\min}\}} = 2L_1.
\end{aligned}$$

Similarly, we have

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

To show that (I) 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

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

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 $\nu(\cdot) := \sum_{(a,\hat{a}) \in \mathsf{A} \times \mathsf{A}} K(\cdot|a, \hat{a})\, \xi(a, \hat{a})$, 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, (I) follows. This completes the proof.

8.3 Proof of Lemma 1

Fix any μ. If a function $f : \mathsf{X} \to \mathbb{R}$ is K-Lipschitz continuous for some K, then $g = \frac{f}{K}$ is 1-Lipschitz continuous. Hence, for all $u \in \mathsf{U}$ and $z, y \in \mathsf{X}$ we have

$$\begin{aligned}
&\left| \sum_x f(x) P(x|z, u, \mu) - \sum_x f(x) P(x|y, u, \mu) \right| \\
&= K \left| \sum_x g(x) P(x|z, u, \mu) - \sum_x g(x) P(x|y, u, \mu) \right| \\
&\leq \frac{K}{2}\, \|P(\cdot|z, u, \mu) - P(\cdot|y, u, \mu)\|_1 \text{ (by (6))} \\
&\leq \frac{K K_1}{2}\, 1_{\{z \neq y\}}, \text{ (by Proposition 1)}
\end{aligned}$$

since $\sup_x g(x) - \inf_x g(x) \le 1$. Hence, the contraction operator T_μ maps K-Lipschitz functions to $L_1 + \beta K K_1/2$-Lipschitz functions, since, for all $z, y \in \mathsf{X}$

$$\begin{aligned}|T_\mu f(z) - T_\mu f(y)| &\le \sup_u \Big\{ |R(z,u,\mu) - R(y,u,\mu)| \\ &\quad + \beta \Big| \sum_x f(x) P(x|z,u,\mu) - \sum_x f(x) P(x|y,u,\mu) \Big| \Big\} \\ &\le L_1 1_{\{z \neq y\}} + \beta \frac{K K_1}{2} 1_{\{z \neq y\}} = \Big(L_1 + \beta \frac{K K_1}{2} \Big) 1_{\{z \neq y\}}.\end{aligned}$$

Now we apply T_μ recursively to obtain the sequence $\{T_\mu^n f\}$ by letting $T_\mu^n f = T_\mu(T_\mu^{n-1} f)$, which converges to the value function $Q_{\mu,\max}^{\mathrm{reg},*}$ by the Banach fixed point theorem. Clearly, by mathematical induction, we have for all $n \ge 1$, $T_\mu^n f$ is K_n-Lipschitz continuous, where $K_n = L_1 \sum_{i=0}^{n-1} (\beta K_1/2)^i + K(\beta K_1/2)^n$. If we choose $K < L_1$, then $K_n \le K_{n+1}$ for all n and therefore, $K_n \uparrow \frac{L_1}{1-\beta K_1/2}$. Hence, $T_\mu^n f$ is $\frac{L_1}{1-\beta K_1/2}$-Lipschitz continuous for all n, and therefore, $Q_{\mu,\max}^{\mathrm{reg},*}$ is also $\frac{L_1}{1-\beta K_1/2}$-Lipschitz continuous.

8.4 Proof of Lemma 2

Under Assumption 1, it is straightforward to prove that H_1 maps $\mathcal{P}(\mathsf{X})$ into $\mathcal{C}$. Indeed, the only non-trivial fact is the $\big(K_{\mathrm{Lip}} + L_{\mathrm{reg}}\big)$-Lipschitz continuity of $H_1(\mu) =: Q_\mu^{\mathrm{reg},*}$. This can be proved as follows: For any (x,u) and $(\hat{x},\hat{u})$, we have

$$\begin{aligned}|Q_\mu^{\mathrm{reg},*}(x,u) - Q_\mu^{\mathrm{reg},*}(\hat{x},\hat{u})| &= |R(x,u,\mu) - \Omega(u) + \beta \sum_y Q_{\mu,\max}^{\mathrm{reg},*}(y) P(y|x,u,\mu) \\ &\quad - R(\hat{x},\hat{u},\mu) + \Omega(\hat{u}) - \beta \sum_y Q_{\mu,\max}^{\mathrm{reg},*}(y) P(y|\hat{x},\hat{u},\mu)| \\ &\le L_1 (1_{\{x \neq \hat{x}\}} + \|u - \hat{u}\|_1) + L_{\mathrm{reg}} \|u - \hat{u}\|_1 \\ &\quad + \beta \frac{K_1 K_{\mathrm{Lip}}}{2} (1_{\{x \neq \hat{x}\}} + \|u - \hat{u}\|_1),\end{aligned}$$

where the last inequality follows from (6) and Lemma 1. Hence, $Q_\mu^{\mathrm{reg},*}$ is $\big(K_{\mathrm{Lip}} + L_{\mathrm{reg}}\big)$-Lipschitz continuous.

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

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

$$+\beta\left|\sum_y Q^{\mathrm{reg},*}_{\mu,\max}(y)P(y|x,u,\hat{\mu})-\sum_y Q^{\mathrm{reg},*}_{\hat{\mu},\max}(y)P(y|x,u,\hat{\mu})\right|$$
$$\le L_1\|\mu-\hat{\mu}\|_1+\frac{\beta K_1 K_{\mathrm{Lip}}}{2}\|\mu-\hat{\mu}\|_1+\beta\|Q^{\mathrm{reg},*}_{\mu}-Q^{\mathrm{reg},*}_{\hat{\mu}}\|_\infty,$$

where the last inequality follows from (6) and Lemma 1. This completes the proof.

8.5 Proof of Lemma 3

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

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

where $q^\mu_x(\cdot):=r(x,\cdot,\mu)+\beta\sum_{y\in\mathsf{X}}Q^{\mathrm{reg},*}_{\mu,\max}(y)\,p(y|x,\cdot,\mu)$. By ρ-strong convexity of Ω, $Q^{\mathrm{reg},*}_{\mu}(x,\cdot)$ has a unique maximizer $f_\mu(x)\in\mathsf{U}$ for any $x\in\mathsf{X}$, which is the optimal policy for μ. By Property 2 of Proposition 3, we have

$$f_\mu(x)=\nabla\Omega^*(q^\mu_x),$$

where Ω^* is the Fenchel conjugate of Ω, and $\Omega^*(q^\mu_x)=Q^{\mathrm{reg},*}_{\mu,\max}(x)$.

Moreover, for any $\mu,\hat{\mu}\in\mathcal{P}(\mathsf{X})$ and $x,\hat{x}\in\mathsf{X}$, by property 3 of Proposition 3 and by noting the fact that $\|\cdot\|_\infty$ is the dual norm of $\|\cdot\|_1$ on U, we obtain the following bound:

$$\|f_\mu(x)-f_{\hat{\mu}}(\hat{x})\|_1\le\frac{1}{\rho}\|q^\mu_x-q^{\hat{\mu}}_{\hat{x}}\|_\infty.$$

Note that we have

$$\begin{aligned}\|q^\mu_x-q^{\hat{\mu}}_{\hat{x}}\|_\infty&=\sup_{a\in\mathsf{A}}\Big|r(x,a,\mu)+\beta\sum_{y\in\mathsf{X}}Q^{\mathrm{reg},*}_{\mu,\max}(y)\,p(y|x,a,\mu)\\&\quad-r(\hat{x},a,\hat{\mu})-\beta\sum_{y\in\mathsf{X}}Q^{\mathrm{reg},*}_{\hat{\mu},\max}(y)\,p(y|\hat{x},a,\hat{\mu})\Big|\\&\le L_1(1_{\{x\neq\hat{x}\}}+\|\mu-\hat{\mu}\|_1)\\&\quad+\beta\sup_{a\in\mathsf{A}}\left|\sum_y Q^{\mathrm{reg},*}_{\mu,\max}(y)p(y|x,a,\mu)-\sum_y Q^{\mathrm{reg},*}_{\hat{\mu},\max}(y)p(y|x,a,\mu)\right|\\&\quad+\beta\sup_{a\in\mathsf{A}}\left|\sum_y Q^{\mathrm{reg},*}_{\hat{\mu},\max}(y)p(y|x,a,\mu)-\sum_y Q^{\mathrm{reg},*}_{\hat{\mu},\max}(y)p(y|\hat{x},a,\hat{\mu})\right|\\&\le L_1(1_{\{x\neq\hat{x}\}}+\|\mu-\hat{\mu}\|_1)+\beta\|Q^{\mathrm{reg},*}_{\mu}-Q^{\mathrm{reg},*}_{\hat{\mu}}\|_\infty\\&\quad+\beta\frac{K_1K_{\mathrm{Lip}}}{2}(1_{\{x\neq\hat{x}\}}+\|\mu-\hat{\mu}\|_1)\\&\le K_{\mathrm{Lip}}(1_{\{x\neq\hat{x}\}}+\|\mu-\hat{\mu}\|_1)+\beta K_{H_1}\|\mu-\hat{\mu}\|_1\\&\le K_{H_1}(1_{\{x\neq\hat{x}\}}+\|\mu-\hat{\mu}\|_1).\end{aligned}$$

Therefore, we obtain

$$\|f_\mu(x) - f_{\hat{\mu}}(\hat{x})\|_1 \leq \frac{1}{\rho} K_{H_1}(1_{\{x \neq \hat{x}\}} + \|\mu - \hat{\mu}\|_1).$$

8.6 Proof of Theorem 1

Let $\mu_\varepsilon \in \Lambda^{\text{reg}}(\pi_\varepsilon)$. Then, we have

$$\begin{aligned}
\|\mu_\varepsilon - \mu_*\|_1 &= \sum_y \left| \sum_x P(y|x, \pi_\varepsilon, \mu_\varepsilon)\, \mu_\varepsilon(x) - \sum_x P(y|x, \pi_*(x), \mu_*)\, \mu_*(x) \right| \\
&\leq \sum_y \left| \sum_x P(y|x, \pi_\varepsilon, \mu_\varepsilon)\, \mu_\varepsilon(x) - \sum_x P(y|x, \pi_*(x), \mu_*)\, \mu_\varepsilon(x) \right| \\
&\quad + \sum_y \left| \sum_x P(y|x, \pi_*(x), \mu_*)\, \mu_\varepsilon(x) - \sum_x P(y|x, \pi_*(x), \mu_*)\, \mu_*(x) \right| \\
&\overset{(I)}{\leq} \sum_x \|P(\cdot|x, \pi_\varepsilon(x), \mu_\varepsilon) - P(\cdot|x, \pi_*(x), \mu_*)\|_1\, \mu_\varepsilon(x) \\
&\quad + \frac{K_1}{2}\left(1 + \frac{K_{H_1}}{\rho}\right) \|\mu_\varepsilon - \mu_*\|_1 \\
&\leq K_1 \left(\sup_x \|\pi_\varepsilon(x) - \pi_*(x)\|_1 + \|\mu_\varepsilon - \mu_*\|_1 \right) \\
&\quad + \frac{K_1}{2}\left(1 + \frac{K_{H_1}}{\rho}\right) \|\mu_\varepsilon - \mu_*\|_1 \\
&\leq K_1\, \varepsilon + \left(\frac{3\, K_1}{2} + \frac{K_1\, K_{H_1}}{2\rho} \right) \|\mu_\varepsilon - \mu_*\|_1.
\end{aligned}$$

Note that Lemma 3 and Proposition 1 lead to

$$\|P(\cdot|x, \pi_*(x), \mu_*) - P(\cdot|y, \pi_*(y), \mu_*)\|_1 \leq K_1 \left(1 + \frac{K_{H_1}}{\rho}\right) 1_{\{x \neq y\}}.$$

Hence, (I) follows from [24,Lemma A2]. Therefore, we have:

$$\|\mu_\varepsilon - \mu_*\|_1 \leq \frac{K_1\, \varepsilon}{1 - C_1},$$

where $C_1 := \left(\frac{3\, K_1}{2} + \frac{K_1\, K_{H_1}}{2\rho} \right)$. Note that by Assumption 2, $C_1 < 1$. Now, fix any policy $\pi \in \Pi$. Then, we have

$$\begin{aligned}
&\|J^{\text{reg}}_{\mu_*}(\pi, \cdot) - J^{\text{reg}}_{\mu_\varepsilon}(\pi, \cdot)\|_\infty \\
&= \sup_x \Big| R^{\text{reg}}(x, \pi(x), \mu_*) + \beta \sum_y J^{\text{reg}}_{\mu_*}(\pi, y)\, p(y|x, \pi(x), \mu_*) \\
&\quad - R^{\text{reg}}(x, \pi(x), \mu_\varepsilon) - \beta \sum_y J^{\text{reg}}_{\mu_\varepsilon}(\pi, y)\, p(y|x, \pi(x), \mu_\varepsilon) \Big| \\
&\leq L_1\, \|\mu_* - \mu_\varepsilon\|_1
\end{aligned}$$

$$
\begin{aligned}
&+\beta \sup_x \left| \sum_y J^{\mathrm{reg}}_{\mu_*}(\pi, y)\, p(y|x, \pi(x), \mu_*) - \sum_y J^{\mathrm{reg}}_{\mu_*}(\pi, y)\, p(y|x, \pi(x), \mu_\varepsilon) \right| \\
&+\beta \sup_x \left| \sum_y J^{\mathrm{reg}}_{\mu_*}(\pi, y)\, p(y|x, \pi(x), \mu_\varepsilon) - \sum_y J^{\mathrm{reg}}_{\mu_\varepsilon}(\pi, y)\, p(y|x, \pi(x), \mu_\varepsilon) \right| \\
&\overset{(II)}{\leq} \left(L_1 + \frac{\beta K_1 K_{\mathrm{Lip}}}{2} \right) \|\mu_* - \mu_\varepsilon\|_1 + \beta \|J^{\mathrm{reg}}_{\mu_*}(\pi, \cdot) - J^{\mathrm{reg}}_{\mu_\varepsilon}(\pi, \cdot)\|_\infty \\
&\leq \left(L_1 + \frac{\beta K_1 K_{\mathrm{Lip}}}{2} \right) \frac{K_1 \varepsilon}{1 - C_1} + \beta \|J^{\mathrm{reg}}_{\mu_*}(\pi, \cdot) - J^{\mathrm{reg}}_{\mu_\varepsilon}(\pi, \cdot)\|_\infty .
\end{aligned}
$$

Here, (II) follows from (6) and the fact that $J^{\mathrm{reg}}_{\mu_*}(\pi, \cdot)$ is K_{Lip}-Lipschitz continuous, which can be proved as in Lemma 1. Therefore, we obtain

$$
\|J^{\mathrm{reg}}_{\mu_*}(\pi, \cdot) - J^{\mathrm{reg}}_{\mu_\varepsilon}(\pi, \cdot)\|_\infty \leq \frac{C_2\, \varepsilon}{1 - \beta}, \tag{7}
$$

where $C_2 := \left(L_1 + \frac{\beta K_1 K_{\mathrm{Lip}}}{2} \right) \frac{K_1}{1 - C_1}$.

Note that we also have

$$
\begin{aligned}
&\|J^{\mathrm{reg}}_{\mu_*}(\pi_*, \cdot) - J^{\mathrm{reg}}_{\mu_*}(\pi_\varepsilon, \cdot)\|_\infty \\
&= \sup_x \Big| R^{\mathrm{reg}}(x, \pi_*(x), \mu_*) + \beta \sum_y J^{\mathrm{reg}}_{\mu_*}(\pi_*, y)\, p(y|x, \pi_*(x), \mu_*) \\
&\quad - R^{\mathrm{reg}}(x, \pi_\varepsilon(x), \mu_*) - \beta \sum_y J^{\mathrm{reg}}_{\mu_*}(\pi_*, y)\, p(y|x, \pi_\varepsilon(x), \mu_*) \Big| \\
&\leq (L_1 + L_{\mathrm{reg}}) \sup_x \|\pi_*(x) - \pi_\varepsilon(x)\|_1 \\
&\quad + \beta \sup_x \left| \sum_y J^{\mathrm{reg}}_{\mu_*}(\pi_*, y)\, p(y|x, \pi_*(x), \mu_*) - \sum_y J^{\mathrm{reg}}_{\mu_*}(\pi_*, y)\, p(y|x, \pi_\varepsilon(x), \mu_*) \right| \\
&\quad + \beta \sup_x \left| \sum_y J^{\mathrm{reg}}_{\mu_*}(\pi_*, y)\, p(y|x, \pi_\varepsilon(x), \mu_*) - \sum_y J^{\mathrm{reg}}_{\mu_*}(\pi_\varepsilon, y)\, p(y|x, \pi_\varepsilon(x), \mu_*) \right| \\
&\overset{(III)}{\leq} \left(L_1 + L_{\mathrm{reg}} + \frac{\beta K_1 K_{\mathrm{Lip}}}{2} \right) \sup_x \|\pi_*(x) - \pi_\varepsilon(x)\|_1 \\
&\quad + \beta \|J^{\mathrm{reg}}_{\mu_*}(\pi_*, \cdot) - J^{\mathrm{reg}}_{\mu_*}(\pi_\varepsilon, \cdot)\|_\infty \\
&\leq \left(L_1 + L_{\mathrm{reg}} + \frac{\beta K_1 K_{\mathrm{Lip}}}{2} \right) \varepsilon + \beta \|J^{\mathrm{reg}}_{\mu_*}(\pi_*, \cdot) - J^{\mathrm{reg}}_{\mu_*}(\pi_\varepsilon, \cdot)\|_\infty .
\end{aligned}
$$

Here, (III) follows from (6) and the fact that $J^{\mathrm{reg}}_{\mu_*}(\pi_*, \cdot)$ is K_{Lip}-Lipschitz continuous, which can be proved as in Lemma 1. Therefore, we obtain

$$
\|J^{\mathrm{reg}}_{\mu_*}(\pi_*, \cdot) - J^{\mathrm{reg}}_{\mu_*}(\pi_\varepsilon, \cdot)\|_\infty \leq \frac{C_3 \varepsilon}{1 - \beta}, \tag{8}
$$

where $C_3 := \left(L_1 + L_{\mathrm{reg}} + \frac{\beta K_1 K_{\mathrm{Lip}}}{2} \right)$.

Note that we must prove that

$$
J^{(N)}_i(\boldsymbol{\pi}^{(N)}) \geq \sup_{\pi^i \in \Pi_i} J^{(N)}_i(\boldsymbol{\pi}^{(N)}_{-i}, \pi^i) - \tau\, \varepsilon - \delta \tag{9}
$$

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 (9) for Agent 1 only. Given $\delta > 0$, for each $N \geq 1$, let $\tilde{\pi}^{(N)} \in \Pi_1$ be such that

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

Then, by [33,Theorem 4.10], we have

$$\begin{aligned}
&\lim_{N\to\infty} J_1^{(N)}(\tilde{\pi}^{(N)}, \pi_\varepsilon, \ldots, \pi_\varepsilon) = \lim_{N\to\infty} J_{\mu_\varepsilon}^{\mathrm{reg}}(\tilde{\pi}^{(N)}) \\
&\leq \lim_{N\to\infty} J_{\mu_*}^{\mathrm{reg}}(\tilde{\pi}^{(N)}) + \frac{C_2\,\varepsilon}{1-\beta} \text{ (by (7))} \\
&\leq \sup_{\pi'} J_{\mu_*}^{\mathrm{reg}}(\pi') + \frac{C_2\,\varepsilon}{1-\beta} \\
&= J_{\mu_*}^{\mathrm{reg}}(\pi_*) + \frac{C_2\,\varepsilon}{1-\beta} \\
&\leq J_{\mu_*}^{\mathrm{reg}}(\pi_\varepsilon) + \frac{C_2\,\varepsilon}{1-\beta} + \frac{C_3\,\varepsilon}{1-\beta} \text{ (by (8))} \\
&\leq J_{\mu_\varepsilon}^{\mathrm{reg}}(\pi_\varepsilon) + \frac{2\,C_2\,\varepsilon}{1-\beta} + \frac{C_3\,\varepsilon}{1-\beta} \text{ (by (7))} \\
&= \lim_{N\to\infty} J_1^{(N)}(\pi_\varepsilon, \pi_\varepsilon, \ldots, \pi_\varepsilon) + \tau\,\varepsilon.
\end{aligned}$$

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

$$\begin{aligned}
&\sup_{\pi' \in \Pi_1} J_1^{(N)}(\pi', \pi_\varepsilon, \ldots, \pi_\varepsilon) - \delta - \tau\,\varepsilon \\
&\leq J_1^{(N)}(\tilde{\pi}^{(N)}, \pi_\varepsilon, \ldots, \pi_\varepsilon) - \frac{2\delta}{3} - \tau\,\varepsilon \\
&\leq J_{\mu_*}^{\mathrm{reg}}(\pi_\varepsilon) - \frac{\delta}{3} \\
&\leq J_1^{(N)}(\pi_\varepsilon, \pi_\varepsilon, \ldots, \pi_\varepsilon).
\end{aligned}$$

for all $N \geq N(\delta)$.

References

1. Adlakha S, Johari R, Weintraub G (2015) Equilibria of dynamic games with many players: existence, approximation, and market structure. J Econ Theory 156:269–316
2. Anahtarci B, Kariksiz C, Saldi N (2019) Fitted Q-learning in mean-field games. arXiv:1912.13309
3. Anahtarci B, Kariksiz C, Saldi N (2020) Value iteration algorithm for mean field games. Syst Control Lett 143
4. Antos A, Munos R, Szepesvári C (2007) Fitted Q-iteration in continuous action-space MDPs. In: Proceedings of the 20th international conference on neural information processing systems, pp 9–16
5. Antos A, Munos R, Szepesvári C (2007) Fitted Q-iteration in continuous action-space MDPs. Tech. rep. inria-00185311v1
6. Bensoussan A, Frehse J, Yam P (2013) Mean field games and mean field type control theory. Springer, New York
7. Biswas A (2015) Mean field games with ergodic cost for discrete time Markov processes. arXiv:1510.08968
8. Cardaliaguet P (2011) Notes on mean-field games. Technical report, p 120

9. Carmona R, Delarue F (2013) Probabilistic analysis of mean-field games. SIAM J Control Optim 51(4):2705–2734
10. Carmona R, Lauriere M, Tan Z (2019) Linear-quadratic mean-field reinforcement learning: convergence of policy gradient methods. arXiv:1910.04295
11. Elie R, Perolat J, Lauriere M, Geist M, Pietquin O (2019) Approximate fictitious play for mean-field games. arXiv:1907.02633
12. Elliot R, Li X, Ni Y (2013) Discrete time mean-field stochastic linear-quadratic optimal control problems. Automatica 49:3222–3233
13. Fu Z, Yang Z, Chen Y, Wang Z (2019) Actor-critic provably finds Nash equilibria of linear-quadratic mean-field games. arXiv:1910.07498
14. Geist M, Scherrer B, Pietquin O (2019) A theory of regularized Markov decision processes. arXiv:1901.11275
15. Georgii H (2011) Gibbs Measures and Phase Transitions. De Gruyter studies in mathematics. De Gruyter
16. Gomes D, Mohr J, Souza R (2010) Discrete time, finite state space mean field games. J Math Pures Appl 93:308–328
17. Gomes D, Saúde J (2014) Mean field games models: a brief survey. Dyn Games Appl 4(2):110–154
18. Guo X, Hu A, Xu R, Zhang J (2019) Learning mean-field games. arXiv:1901.09585
19. Huang M (2010) Large-population LQG games involving major player: the nash certainty equivalence principle. SIAM J Control Optim 48(5):3318–3353
20. Huang M, Caines P, Malhamé R (2007) Large-population cost coupled LQG problems with nonuniform agents: individual-mass behavior and decentralized ϵ-Nash equilibria. IEEE Trans Autom Control 52(9):1560–1571
21. Huang M, Malhamé R, Caines P (2006) Large population stochastic dynamic games: closed loop McKean-Vlasov systems and the Nash certainty equivalence principle. Commun Inform Syst 6:221–252
22. Kara AD, Yüksel S (2019) Robustness to incorrect priors in partially observed stochastic control. SIAM J Control Optim 57(3):1929–1964
23. Kara AD, Yüksel S (2020) Robustness to incorrect system models in stochastic control. SIAM J Control Optim 58(2):1144–1182
24. Kontorovich L, Ramanan K (2008) Concentration inequalities for dependent random variables via the martingale method. Ann Probab 36(6):2126–2158
25. Lasry J, Lions P (2007) Mean field games. Japan J Math 2:229–260
26. Mehta P, Meyn S (2009) Q-learning and Pontryagin's minimum principle. In: Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference, pp 3598–3605
27. Moon J, Başar T (2015) Discrete-time decentralized control using the risk-sensitive performance criterion in the large population regime: a mean field approach. In: ACC 2015. Chicago
28. Moon J, Başar T (2016) Discrete-time mean field Stackelberg games with a large number of followers. In: CDC 2016. Las Vegas
29. Moon J, Başar T (2016) Robust mean field games for coupled Markov jump linear systems. Int J Control 89(7):1367–1381
30. Neu G, Jonsson A, Gomez V (2017) A unified view of entropy-regularized Markov decision processes. arXiv:1705.07798
31. Nourian M, Nair G (2013) Linear-quadratic-Gaussian mean field games under high rate quantization. In: CDC 2013. Florence
32. Saldi N (2019) Discrete-time average-cost mean-field games on Polish spaces. arXiv:1908.08793 (accepted to Turkish Journal of Mathematics)
33. Saldi N, Başar T, Raginsky M (2018) Markov-Nash equilibria in mean-field games with discounted cost. SIAM J Control Optim 56(6):4256–4287
34. Saldi N, Başar T, Raginsky M (2019) Approximate Markov-Nash equilibria for discrete-time risk-sensitive mean-field games. to appear in Mathematics of Operations Research
35. Saldi N, Başar T, Raginsky M (2019) Approximate Nash equilibria in partially observed stochastic games with mean-field interactions. Math Oper Res 44(3):1006–1033
36. Shalev-Shwartz S (2007) Online learning: theory, algorithms, and applications. Ph.D. thesis, The Hebrew University of Jerusalem
37. Tembine H, Zhu Q, Başar T (2014) Risk-sensitive mean field games. IEEE Trans Autom Control 59(4):835–850
38. Vidyasagar M (2010) Learning and generalization: with applications to neural networks, 2nd edn. Springer, New York
39. Wiecek P (2020) Discrete-time ergodic mean-field games with average reward on compact spaces. Dyn Games Appl 10:222–256

40. Wiecek P, Altman E (2015) Stationary anonymous sequential games with undiscounted rewards. J Optim Theory Appl 166(2):686–710
41. Yang J, Ye X, Trivedi R, Hu X, Zha H (2018) Learning deep mean field games for modelling large population behaviour. arXiv:1711.03156
42. Yin H, Mehta P, Meyn S, Shanbhag U (2014) Learning in mean-field games. IEEE Trans Autom Control 59:629–644