Journal of Machine Learning Research 24 (2023) 1-59 Submitted 5/21; Revised 10/22; Published 1/23

Learning Mean-Field Games with Discounted and Average Costs

Berkay Anahtarci BERKAY.ANAHTARCI@OZYEGIN.EDU.TR
Department of Natural and Mathematical Sciences
Özyeğin University
İstanbul, Turkey

Can Deha Kariksiz DEHA.KARIKSIZ@OZYEGIN.EDU.TR
Department of Natural and Mathematical Sciences
Özyeğin University
İstanbul, Turkey

Naci Saldi NACI.SALDI@BILKENT.EDU.TR
Department of Mathematics
Bilkent University
Ankara, Turkey

Editor: Alekh Agarwal

Abstract

We consider learning approximate Nash equilibria for discrete-time mean-field games with stochastic nonlinear state dynamics subject to both average and discounted costs. To this end, we introduce a mean-field equilibrium (MFE) operator, whose fixed point is a mean-field equilibrium, i.e., equilibrium in the infinite population limit. We first prove that this operator is a contraction, and propose a learning algorithm to compute an approximate mean-field equilibrium by approximating the MFE operator with a random one. Moreover, using the contraction property of the MFE operator, we establish the error analysis of the proposed learning algorithm. We then show that the learned mean-field equilibrium constitutes an approximate Nash equilibrium for finite-agent games.

Keywords: Mean-field games, approximate Nash equilibrium, fitted Q-iteration algorithm, discounted-cost, average-cost.

1. Introduction

We consider learning approximate Nash equilibria in discrete-time stochastic dynamic games with a large population of identical agents in a mean-field interaction. The usual approach to analyse these game models is to study the infinite-population limit of the problem. This idea was utilized in the works of Huang et al. (2006), Lasry and P.Lions (2007), where mean-field games (MFGs) were introduced to obtain an approximate Nash equilibria for continuous-time differential games with a large number of agents interacting via a mean-field term (i.e., the empirical distribution of the local states). For studies of continuous-time mean-field games with various models and cost functions, see Huang et al. (2007); Tembine et al. (2014); Huang (2010); Bensoussan et al. (2013); Cardaliaguet (2011); Carmona and Delarue (2013); Gomes and Saúde (2014); Moon and Başar (2016a).

Our goal in this paper is to learn approximate Nash equilibria for a class of stochastic dynamic games by considering stationary mean-field games in the infinite population limit. In particular, we establish an algorithm to learn *stationary* or *oblivious* mean-field equilibrium (see Weintraub et al. (2005, 2008)) in the infinite population limit and make use of the learned equilibrium in the finite agent setting as an approximate Nash equilibrium.

In stationary mean-field games, a generic agent competes against a stationary mean-field term that time-homogeneously models the collective behaviour of other agents (Weintraub et al. (2005)), and therefore solves a Markov decision process (MDP) with a constraint on the stationary distribution of the state. The stationary mean-field equilibrium, which consists of a policy and a state measure, is the concept of equilibrium in the infinite-population limit. This pair must satisfy the Nash certainty equivalence (NCE) principle (Huang et al. (2006)), which states that the policy should be optimal under a given state measure, and when the generic agent applies this policy the resulting stationary distribution of the agent's state must be the same as the state measure. The existence of stationary mean-field equilibrium can be established via Kakutani's fixed point theorem under quite mild assumptions. Furthermore, it can be shown that when the number of agents is large enough, the policy in stationary mean-field equilibrium is an approximate Nash equilibrium for a finite-agent setting (Adlakha et al. (2015)).

In the literature, Weintraub et al. (2010) propose an algorithm for computing oblivious equilibrium in a stationary mean-field industry dynamics model. Adlakha et al. (2015) study a stationary mean-field game model with a countable state-space under an infinite-horizon discounted-cost criterion. Huang and Ma (2019) consider stationary mean-field games with binary action space, where they establish the existence and the uniqueness of the stationary mean-field equilibrium. Light and Weintraub (2022) consider stationary mean-field games with a continuum of states and actions, and establish a novel uniqueness result for stationary mean-field equilibrium Gomes et al. (2010) study both stationary and non-stationary mean-field games with a finite state space over a finite horizon and establish the existence and uniqueness of the mean-field equilibrium for both cases. Elliot et al. (2013); Moon and Başar (2015); Nourian and Nair (2013); Moon and Başar (2016b) consider discrete-time mean-field games with linear state dynamics. Saldi et al. (2018, 2019) consider a discrete-time non-stationary mean-field game with Polish state and action spaces under the discounted-cost optimality criterion for fully-observed case and partially-observed case, respectively. Saldi et al. (2020) consider a discrete-time risk-sensitive non-stationary mean-field game with Polish state and action spaces. Biswas (2015); Wiecek (2019); Wiecek and Altman (2015); Saldi (2020) study discrete-time non-stationary mean-field games subject to the average-cost optimality criterion.

We point out that, except the linear model and the paper of Weintraub et al. (2010), the studies mentioned above only establish the existence and uniqueness of the mean-field equilibrium, and they propose no algorithm with convergence guarantee to compute this mean-field equilibrium when the model is known. In our recent work (Anahtarci et al. (2020a)), we study this problem for a very general class of models, propose a value iteration algorithm, and prove the convergence of this algorithm to the stationary mean-field equilibrium. In this current paper, we generalize this algorithm to the model-free setting using fitted Q-iteration (see Antos et al. (2007a)); that is, we propose a learning algorithm

to compute an equilibrium solution for discrete-time stationary mean-field games under the discounted-cost and average-cost optimality criteria.

Learning in stationary mean-field games has become prominent in recent years. In the continuous-time setup, Yin et al. (2014) develop a learning algorithm for a mean-field oscillator game model to obtain approximate Nash equilibrium. In the discrete case, Kash et al. (2011) consider the learning of equilibrium policy in static anonymous games with countably many players. Yang et al. (2018b) establish a learning algorithm for classical stochastic games via mean-field approximation by factoring the Q-function in terms of actions. Subramanian and Mahajan (2018, 2019) consider learning gradient-based equilibria in stationary mean-field games and develop a two-time scale stochastic gradient ascent algorithm, respectively. Guo et al. (2019) develop a Q-learning algorithm to obtain stationary mean-field equilibria for finite state-action stationary mean-field games, where the convergence analysis depends on contractivity assumptions on the operators involved in the algorithm. Elie et al. (2019) establish a fictitious play iterative learning algorithm to compact state-action non-stationary mean-field games under finite-horizon discounted cost criterion, where the dynamics of the state and the one-stage cost function satisfy certain structure. They also propose an error analysis of the learning algorithm for the game model with deterministic state dynamics. Carmona et al. (2019a) study linear-quadratic mean-field control and establish the convergence of policy gradient algorithm. Fu et al. (2019) develop an actor-critic algorithm to learn mean-field equilibrium for linear-quadratic mean-field games. Yang et al. (2018a) consider a mean-field game in which agents can control their transition probabilities without any restriction. In this case, the action space becomes the set of distributions on the state space. Using this specific structure, they can transform a mean-field game into an equivalent deterministic Markov decision process by enlarging the state and action spaces, and then, apply classical reinforcement learning algorithms to compute mean-field equilibrium. Carmona et al. (2019b) apply a similar analysis to mean-field control problems, and the convergence of the Q-learning algorithm for deterministic systems is established.

In this paper, we develop a learning algorithm that guarantees convergence in a discrete-time stationary mean-field game with nonlinear stochastic state dynamics. We take into account the average cost criterion, in contrast to earlier research that mainly dealt with discounted cost or finite-horizon total cost criteria. It's also important to note that the majority of the aforementioned works with convergence guarantees focus on finite state and finite action settings, whereas we assume that the action space is a compact and uncountable subset of a finite dimensional Euclidean space. In general, it is more challenging to deal with this assumption. Furthermore, we prove that our algorithm converges to the global stationary mean-field equilibrium rather than local stationary mean-field equilibria. We also establish easily verifiable conditions on the system components for the convergence of the learning algorithm, which is lacking in some of the prior works mentioned above.

Our learning algorithm performs two-steps in each iteration. In the first step, given any mean-field term, an optimal policy is learned via a fitted Q-iteration algorithm. Then, using this optimal policy and the current mean-field term, the next mean-field term is computed in the second step by an empirical estimate of the transition probability, which is obtained via simulation. We prove that the policy obtained by this algorithm is close to the mean-field equilibrium policy, and can therefore be used as an approximate Nash equilibrium for a finite-agent game if the number of agents is sufficiently large.

The error analysis of our learning process depends crucially on the determination of the contraction of the operator providing the stationary mean-field equilibrium. It is necessary 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 through the value iteration algorithm. Although establishing the Lipschitz continuity of the optimal Q 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 problem, we assume that the function in the optimality equation is strongly convex and has Lipschitz continuous gradient. In our recent work (Anahtarci et al. (2020b)) we provide a different approach by introducing a strongly convex regularization function in the one-stage cost function that helps us to obtain Lipschitz continuity of the optimal policy with respect to the mean-field term via duality between strong convexity and smoothness, therefore eliminating 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. Although this regularization approach allows us to relax the assumptions on the system components, it adds bias to the equilibrium solution since regularization in general favours randomized policies over deterministic policies, and causes the regularized stationary mean-field equilibrium to deviate from the true stationary mean-field equilibrium as a result of the additional regularization term in the one-stage cost function.

The paper is organized as follows. In Section 2, we introduce the mean-field game and define the mean-field equilibrium. In Section 3 and Section 6, we introduce MFE operator when the model is known for discounted-cost and average-cost, respectively. In Section 4 and Section 7, we formulate the finite-agent version of the game problem for discounted-cost and average-cost, respectively. In Section 5 and Section 8, we propose and perform the error analysis of the learning algorithm for the unknown model and prove that learned mean-field equilibrium constitutes an approximate Nash equilibrium for finite-agent games for discounted-cost and average-cost, respectively. In Section 9, we propose a numerical example. Section 10 concludes the paper.

Notation. For a finite set E, we let $\mathcal{P}(\mathsf{E})$ and $M(\mathsf{E})$ denote the set of all probability distributions on E and the set of real-valued functions on E, respectively. In this paper, $\mathcal{P}(\mathsf{E})$ is always endowed with l_1-norm $\|\cdot\|_1$. 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 $B(a,\rho) \coloneqq \{b : \|a-b\| \leq \rho\}$, where $\|\cdot\|$ denotes the Euclidean norm. Let $Q : \mathsf{E}_1 \times \mathsf{E}_2 \to \mathbb{R}$, where E_1 and E_2 are two sets. Then, we define $Q_{\min}(e_1) \coloneqq \inf_{e_2 \in \mathsf{E}_2} Q(e_1, e_2)$. The notation $v \sim \nu$ means that the random element v has distribution ν.

2. Mean-field games and mean-field equilibria

In this paper, we consider a discrete-time mean-field game with state space X and action space A. Here, X is a finite set with the discrete metric $d_{\mathsf{X}}(x,y) = 1_{\{x \neq y\}}$ and A is a convex and compact subset of a finite dimensional Euclidean space $\mathbb{R}^{\dim_{\mathsf{A}}}$ equipped with the Euclidean norm $\|\cdot\|$[1]. The state dynamics evolve according to the transition probability

1. In this work, by updating several definitions appropriately, all results are still true if one metrizes the finite dimensional Euclidean space $\mathbb{R}^{\dim_{\mathsf{A}}}$ with l_p-norm for $p \geq 1$.

$p : \mathsf{X}\times\mathsf{A}\times\mathcal{P}(\mathsf{X}) \to \mathcal{P}(\mathsf{X})$; that is, given the current state $x(t)$, action $a(t)$, and state-measure μ, the next state $x(t+1)$ is distributed as follows:

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

In this model, 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. A policy π is deterministic if $\pi(\,\cdot\,|x) = \delta_{f(x)}(\,\cdot\,)$ for some $f : \mathsf{X} \to \mathsf{A}$. Let Π_d denote the set of all deterministic policies.

Although we name this model as mean-field game, it is indeed neither a game nor a Markov decision process (MDP) in the strict sense. This model is in between them. Similar to the MDP model, we have a single agent with Markovian dynamics that has an objective function to minimize. However, similar to the game model, this agent should also compete with the collective behaviour of other agents. We model this collective behaviour by an exogenous *state-measure* $\mu \in \mathcal{P}(\mathsf{X})$[2]. By law of large numbers, this measure μ should be consistent with the state distribution of this single agent when the agent applies its optimal policy. The precise mathematical description of the problem is given as follows.

If we fix a state-measure $\mu \in \mathcal{P}(\mathsf{X})$, which represents the collective behaviour of the other agents, the evolution of the state and action of a generic agent is governed by transition probability $p : \mathsf{X} \times \mathsf{A} \times \mathcal{P}(\mathsf{X}) \to \mathcal{P}(\mathsf{X})$, policy $\pi : \mathsf{X} \to \mathcal{P}(\mathsf{A})$, and initial distribution η_0 of the state; that is,

$$\begin{aligned} &x(0) \sim \eta_0,\ \ x(t) \sim p(\,\cdot\,|x(t-1), a(t-1), \mu),\ \ t \geq 1, \\ &a(t) \sim \pi(\,\cdot\,|x(t)),\ \ t \geq 0. \end{aligned}$$

For this model, a policy $\pi^* \in \Pi$ of a generic agent is optimal for μ if

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

where

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

or

$$J_\mu(\pi) = \limsup_{T\to\infty} \frac{1}{T} E^\pi\left[\sum_{t=0}^{T-1} c(x(t), a(t), \mu)\right].$$

Here, the first cost function is the discounted-cost with discount factor $\beta \in (0,1)$ and the second cost function is the average-cost. The measurable function $c : \mathsf{X}\times\mathsf{A}\times\mathcal{P}(\mathsf{X}) \to [0,\infty)$ is the one-stage cost function. With these definitions, to introduce the optimality criteria of the model, we need the following two set-valued mappings.

2. 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 and is a future research direction.

The first set-valued mapping $\Psi : \mathcal{P}(\mathsf{X}) \to 2^{\Pi}$ is defined as

$$\Psi(\mu) = \{\pi \in \Pi : \pi \text{ is optimal for } \mu \text{ when } \eta_0 = \mu\}.$$

Hence, $\Psi(\mu)$ is the set of optimal policies for a given state-measure μ when the initial distribution η_0 is equal to μ as well.

We define the second set-valued mapping $\Lambda : \Pi \to 2^{\mathcal{P}(\mathsf{X})}$ as follows: for any $\pi \in \Pi$, the state-measure $\mu_\pi \in \Lambda(\pi)$ if it satisfies the following fixed point equation:

$$\mu_\pi(\,\cdot\,) = \sum_{x \in \mathsf{X}} \int_{\mathsf{A}} p(\,\cdot\,|x, a, \mu_\pi)\, \pi(da|x)\, \mu_\pi(x).$$

Note that if $\mu_\pi \in \Lambda(\pi)$ and $\eta_0 = \mu_\pi$, then $x(t) \sim \mu_\pi$ for all $t \geq 0$ under policy π. If there is no assumption on the transition probability $p : \mathsf{X} \times \mathsf{A} \times \mathcal{P}(\mathsf{X}) \to \mathcal{P}(\mathsf{X})$, we may have $\Lambda(\pi) = \emptyset$ for some π. However, under Assumption 1, we always have $\Lambda(\pi)$ non-empty, which will be proved in Lemma 3.

We can now define the notion of equilibrium for mean-field games via the 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* $\mu_* \in \Lambda(\pi_*)$*; that is,* π_* *is an optimal policy for* μ_* *and* μ_* *is the stationary distribution of the states under policy* π_* *and initial distribution* μ_**.*

In the literature, the existence of mean-field equilibria has been established for the discounted-cost in Saldi et al. (2018) and for the average-cost in Wiecek (2019); Saldi (2020). Our aim in this paper is to develop a learning algorithm for computing an approximate mean-field equilibrium in the model-free setting. To that end, we define the following relaxed version of mean-field equilibrium.

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

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

Note that in above definition, we require that π_* is deterministic. Indeed, this is the case under the assumptions stated below. Therefore, without loss of generality, we can place this restriction on π_*.

With this definition, our goal now is to learn an ε-mean-field equilibrium policy under the model-free set-up. To this end, we will impose certain assumptions on the components of the mean-field game model. Before doing this, we need to give some definitions. Let us define $M_\tau(\mathsf{X})$ as the set of real-valued functions on X bounded by $\|c\|_\infty/(1-\tau)$. Here, $\tau = \beta$ if the objective function is discounted-cost and $\tau = \beta^{\text{av}}$ (see Assumption 3) if the objective function is average-cost. Let $F : \mathsf{X} \times M_\tau(\mathsf{X}) \times \mathcal{P}(\mathsf{X}) \times \mathsf{A} \to \mathbb{R}$ be given by

$$F : \mathsf{X} \times M_\tau(\mathsf{X}) \times \mathcal{P}(\mathsf{X}) \times \mathsf{A} \ni (x, v, \mu, a) \mapsto c(x, a, \mu) + \xi \sum_{y \in \mathsf{X}} v(y)\, p(y|x, a, \mu) \in \mathbb{R},$$

where $\xi = \beta$ if the objective function is discounted-cost and $\xi = 1$ if the objective function is average-cost. We may now state our assumptions.

Assumption 1

(a) The one-stage cost function c satisfies the following Lipschitz bound:

$$|c(x,a,\mu) - c(\hat{x},\hat{a},\hat{\mu})| \leq L_1 \left(d_{\mathsf{X}}(x,\hat{x}) + \|a-\hat{a}\| + \|\mu-\hat{\mu}\|_1\right), \tag{1}$$

for every $x,\hat{x} \in \mathsf{X}$, $a,\hat{a} \in \mathsf{A}$, and $\mu,\hat{\mu} \in \mathcal{P}(\mathsf{X})$.

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

$$\|p(\cdot|x,a,\mu) - p(\cdot|\hat{x},\hat{a},\hat{\mu})\|_1 \leq K_1 \left(d_{\mathsf{X}}(x,\hat{x}) + \|a-\hat{a}\| + \|\mu-\hat{\mu}\|_1\right), \tag{2}$$

for every $x,\hat{x} \in \mathsf{X}$, $a,\hat{a} \in \mathsf{A}$, and $\mu,\hat{\mu} \in \mathcal{P}(\mathsf{X})$.

(c) There exists $\alpha > 0$ such that for any $a \in \mathsf{A}$ and $\delta > 0$, we have

$$m\left(B(a,\delta) \cap \mathsf{A}\right) \geq \min\left\{\alpha\, m(B(a,\delta)), m(\mathsf{A})\right\}.$$

(d) For any $x \in \mathsf{X}$, $v \in M_\tau(\mathsf{X})$, and $\mu \in \mathcal{P}(\mathsf{X})$, $F(x,v,\mu,\cdot)$ is ρ-strongly convex. Moreover, the gradient $\nabla F(x,v,\mu,a) : \mathsf{X}\times M_\tau(\mathsf{X})\times\mathcal{P}(\mathsf{X})\times\mathsf{A} \to \mathbb{R}^d$ of F with respect to a satisfies the following Lipschitz bound:

$$\sup_{a\in\mathsf{A}} \|\nabla F(x,v,\mu,a) - \nabla F(\hat{x},\hat{v},\hat{\mu},a)\| \leq K_F \left(d_{\mathsf{X}}(x,\hat{x}) + \|v-\hat{v}\|_\infty + \|\mu-\hat{\mu}\|_1\right),$$

for every $x,\hat{x} \in \mathsf{X}$, $v,\hat{v} \in M_\tau(\mathsf{X})$, and $\mu,\hat{\mu} \in \mathcal{P}(\mathsf{X})$.

Let us motivate these assumptions. First, assumptions (a) and (b) are standard conditions in stochastic control theory to obtain a rate of convergence bound for learning algorithms. Assumption (c) is needed to bound the l_∞-norm of Lipschitz continuous functions on A with their l_2-norm. Assumption (d) is imposed to guarantee Lipschitz continuity of the optimal policy with respect to the corresponding state-measure. Indeed, this condition is equivalent to the standard assumption that guarantees Lipschitz continuity, with respect to unknown parameters, of the optimal solutions of the convex optimization problems (Bonnans and Shapiro, 2000, Theorem 4.51).

Example 1 *Let us consider the industry dynamics model, introduced in Weintraub et al. (2005, 2008), where the state $x(t) \in \mathsf{X}$ of the system gives the quality level of the firm, and the state lives in the finite set $\mathsf{X} = \{0,\ldots,m\}$. Given the mean-field term μ, the state of the system evolves in the following form:*

$$x(t+1) = \min\{x(t) + h(a(t),\mu,w(t)), m\},$$

where $a(t) \in \mathsf{A} = [0,K]$ is the action, which denotes the investment of the agent to increase its quality, $h : \mathsf{A}\times\mathcal{P}(\mathsf{X})\times\mathsf{W} \to \mathsf{X}$, and $w(t) \in \mathsf{W}$ is the independent noise[3]. In this model, the problem is to maximize the discounted reward, which is equivalent to minimizing the

3. The state dynamics of the model in Weintraub et al. (2005, 2008) does not depend on the mean-field term μ. For generality, we assume that there is such a dependence in our example.

negative of the discounted reward. Therefore, in the minimization formulation, one-stage cost function has the following form:

$$c(x,a,\mu) = c_2(a) - c_1(x,\mu),$$

where $c_1(x,\mu)$ is the profit of the firm and $c_2(a)$ is the cost of the investment. For this model, Assumption 1-(c) is true with $\alpha = 1$. To have Assumption 1-(d), we need to assume that: (i) $c_2(a)$ is differentiable and ρ-strongly convex (this is true if, for instance, $d^2c_2/da^2 \geq \rho$), (ii) $\mathbb{P}[h(a,\mu,w) = l]$ is convex in a and continuously differentiable in (a,μ), for all $l \in \mathsf{X}$. Indeed, for any x, v, and μ, the function $F(x,v,\mu,a)$ has the following form:

$$\begin{aligned}
F(x,v,\mu,a) &= c_2(a) - c_1(x,\mu) + \xi \sum_{y\in\mathsf{X}} v(y)\,\mathbb{P}[\min\{x + h(a,\mu,w), m\} = y] \\
&= c_2(a) - c_1(x,\mu) + \xi \left[\sum_{x\leq y<m} v(y)\,\mathbb{P}[h(a,\mu,w) = y - x] + v(m)\,\mathbb{P}[h(a,\mu,w) \geq m - x] \right] \\
&= c_2(a) - c_1(x,\mu) + \xi \left[\sum_{0\leq y<m} v(y)\,\mathbb{P}[h(a,\mu,w) = y - x] + \sum_{l=m-x}^{m} v(m)\,\mathbb{P}[h(a,\mu,w) = l] \right],
\end{aligned}$$

where the last equality is true since $\mathbb{P}[h(a,\mu,w) = y - x] = 0$ if $y < x$. Since $c_2(a)$ is ρ-strongly convex and $\mathbb{P}[h(a,\mu,w) = l]$ is convex in a, the function $F(x,v,\mu,a)$ is ρ-strongly convex in a. Moreover, for every $x, \hat{x} \in \mathsf{X}$, $v, \hat{v} \in M_\tau(\mathsf{X})$, and $\mu, \hat{\mu} \in \mathcal{P}(\mathsf{X})$, we have

$$\begin{aligned}
\sup_{a\in\mathsf{A}} |\nabla F(x,v,\mu,a) - \nabla F(\hat{x},\hat{v},\hat{\mu},a)| &\leq \sup_{a\in\mathsf{A}} |\nabla F(x,v,\mu,a) - \nabla F(\hat{x},v,\mu,a)| \\
&+ \sup_{a\in\mathsf{A}} |\nabla F(\hat{x},v,\mu,a) - \nabla F(\hat{x},\hat{v},\mu,a)| + \sup_{a\in\mathsf{A}} |\nabla F(\hat{x},\hat{v},\mu,a) - \nabla F(\hat{x},\hat{v},\hat{\mu},a)|. \quad (3)
\end{aligned}$$

To bound the terms in the sum (3), let us define the following constants:

$$\begin{aligned}
\Theta_1 &\coloneqq \sup_{\mu,a,l} |\nabla_a \mathbb{P}[h(a,\mu,w) = l]| \\
\Theta_2 &\coloneqq \sup_{a,\mu} \sum_{l=0}^{m-1} |\nabla_a \mathbb{P}[h(a,\mu,w) = l] - \nabla_a \mathbb{P}[h(a,\mu,w) = l+1]| \\
\Theta_3 &\coloneqq \sup_{a} \sum_{l=0}^{m} \sup_{\mu} \|\nabla_{a,\mu} \mathbb{P}[h(a,\mu,w) = l]\| \\
\Theta_4 &\coloneqq \sup_{a,\mu} \sum_{l=0}^{m} |\nabla_a \mathbb{P}[h(a,\mu,w) = l]|,
\end{aligned}$$

where $\nabla_a \mathbb{P}[h(a,\mu,w) = l]$ and $\nabla_{a,\mu} \mathbb{P}[h(a,\mu,w) = l]$ are the gradients of $\mathbb{P}[h(a,\mu,w) = l]$ with respect to a and (a,μ), respectively. Since $\mathbb{P}[h(a,\mu,w) = l]$ is continuously differentiable with respect to (a,μ), and the sets A and $\mathcal{P}(\mathsf{X})$ are compact, the constants above are well-defined.

Without loss of generality, suppose $x \leq \hat{x}$. Considering the first term in the sum (3), for all $a \in \mathsf{A}$ we have

$$|\nabla F(x,v,\mu,a) - \nabla F(\hat{x},v,\mu,a)|$$

$$
\begin{aligned}
&\leq \xi \Bigg| \sum_{0\leq y<m} v(y)\, \nabla_a \mathbb{P}[h(a,\mu,w) = y - x] + \sum_{m-x\leq l\leq m} v(m)\, \nabla_a \mathbb{P}[h(a,\mu,w) = l] \\
&\qquad\qquad \sum_{0\leq y<m} v(y)\, \nabla_a \mathbb{P}[h(a,\mu,w) = y - \hat{x}] + \sum_{m-\hat{x}\leq l\leq m} v(m)\, \nabla_a \mathbb{P}[h(a,\mu,w) = l] \Bigg| \\
&\leq \xi \Bigg| \sum_{0\leq y<m} v(y) \sum_{l=y-\hat{x}}^{y-x-1} \left(\nabla_a \mathbb{P}[h(a,\mu,w) = l] - \nabla_a \mathbb{P}[h(a,\mu,w) = l+1]\right) \Bigg| \\
&\qquad\qquad + \xi \left| \sum_{m-\hat{x}\leq y<m-x} v(m)\, \nabla_a \mathbb{P}[h(a,\mu,w) = l] \right| \\
&\leq \xi\, \frac{\|c\|_\infty}{1-\tau}\, (\Theta_2 + \Theta_1)\, |x - \hat{x}|. \qquad (4)
\end{aligned}
$$

For the second term in the sum (3), for all $a \in \mathsf{A}$ we have

$$
\begin{aligned}
&|\nabla F(\hat{x}, v, \mu, a) - \nabla F(\hat{x}, \hat{v}, \mu, a)| \\
&\leq \xi \Bigg| \sum_{\hat{x}\leq y<m} v(y)\, \nabla_a \mathbb{P}[h(a,\mu,w) = y - \hat{x}] + \sum_{m-\hat{x}\leq l\leq m} v(m)\, \nabla_a \mathbb{P}[h(a,\mu,w) = l] \\
&\qquad\qquad \sum_{\hat{x}\leq y<m} \hat{v}(y)\, \nabla_a \mathbb{P}[h(a,\mu,w) = y - \hat{x}] + \sum_{m-\hat{x}\leq l\leq m} \hat{v}(m)\, \nabla_a \mathbb{P}[h(a,\mu,w) = l] \Bigg| \\
&\leq \xi\, \Theta_4\, \|v - \hat{v}\|_\infty. \qquad (5)
\end{aligned}
$$

Finally, for the third term in the sum (3), for all $a \in \mathsf{A}$, we have

$$
\begin{aligned}
&|\nabla F(\hat{x}, \hat{v}, \mu, a) - \nabla F(\hat{x}, \hat{v}, \hat{\mu}, a)| \\
&\leq \xi \Bigg| \sum_{\hat{x}\leq y<m} \hat{v}(y)\, \nabla_a \mathbb{P}[h(a,\mu,w) = y - \hat{x}] + \sum_{m-\hat{x}\leq l\leq m} \hat{v}(m)\, \nabla_a \mathbb{P}[h(a,\mu,w) = l] \\
&\qquad\qquad \sum_{\hat{x}\leq y<m} \hat{v}(y)\, \nabla_a \mathbb{P}[h(a,\hat{\mu},w) = y - \hat{x}] + \sum_{m-\hat{x}\leq l\leq m} \hat{v}(m)\, \nabla_a \mathbb{P}[h(a,\hat{\mu},w) = l] \Bigg| \\
&\leq \xi\, \frac{\|c\|_\infty}{1-\tau} \left| \sum_{0\leq l\leq m} \Big(\nabla_a \mathbb{P}[h(a,\mu,w) = l] - \nabla_a \mathbb{P}[h(a,\hat{\mu},w) = l] \Big) \right|. \qquad (6)
\end{aligned}
$$

By the mean-value theorem, there exists $\tilde{\mu}$ such that

$$
\nabla_a \mathbb{P}[h(a,\mu,w) = l] - \nabla_a \mathbb{P}[h(a,\hat{\mu},w) = l] = \nabla_{a,\mu} \mathbb{P}[h(a,\tilde{\mu},w) = l] \cdot (\mu - \hat{\mu})
$$

Hence, (6) can be bounded from above as follows:

$$
(6) \leq \xi\, \frac{\|c\|_\infty}{1-\tau}\, \Theta_3\, \|\mu - \hat{\mu}\|_1. \qquad (7)
$$

Bringing together the upper bounds in (4), (5), and (7), we get

$$
\sup_{a\in\mathsf{A}} |\nabla F(x, v, \mu, a) - \nabla F(\hat{x}, \hat{v}, \hat{\mu}, a)|
$$

$$\leq \max\left\{\xi\,\frac{\|c\|_\infty m}{1-\tau}\,(\Theta_2+\Theta_1), \xi\,\Theta_4, \xi\,\frac{\|c\|_\infty}{1-\tau}\,\Theta_3\right\}(d_{\mathsf{X}}(x,\hat{x})+\|v-\hat{v}\|_\infty+\|\mu-\hat{\mu}\|_1) \quad (8)$$
$$=: K_F\,(d_{\mathsf{X}}(x,\hat{x})+\|v-\hat{v}\|_\infty+\|\mu-\hat{\mu}\|_1)\,.$$

since $|x-\hat{x}| \leq m\,d_{\mathsf{X}}(x,\hat{x})$. Hence, Assumption 1-(d) is true for this model under the conditions (i) and (ii). Note that the bound in (8) is fairly crude. By using further properties of the transition probability and the one-stage cost function in addition to conditions (i) and (ii) in specific examples, one can significantly improve this bound. Moreover, instead of the l_1-norm on the set of distributions on the state space X, if we use Wasserstein distance of order 1, it is also possible to improve the bound in (8) by altering the related analysis according to this distance.

In this paper, we first consider learning in discounted-cost MFGs. Then, we turn our attention to the average-cost case. In the sequel, we first introduce a mean-field equilibrium (MFE) operator for discounted-cost, which can be used to compute mean-field equilibrium when the model is known. We prove that this operator is a contraction. Then, under model-free setting, we approximate this MFE operator with a random one and establish a learning algorithm. Using this learning algorithm, we obtain ε-mean-field equilibrium policy with high confidence. To obtain the last result, it is essential that MFE operator is contraction. After we complete the analysis for discounted-cost, we study average-cost setting by applying the same strategy.

Before we move on to the next section, for completeness, let us prove the following result.

Lemma 3 *Under Assumption 1, for any π, the set $\Lambda(\pi)$ is non-empty.*

Proof Recall that for any $\pi \in \Pi$, the state-measure $\mu_\pi \in \Lambda(\pi)$ if it satisfies the following fixed-point equation:

$$\mu_\pi(\,\cdot\,) = \sum_{x\in\mathsf{X}} \int_{\mathsf{A}} p(\,\cdot\,|x,a,\mu_\pi)\,\pi(da|x)\,\mu_\pi(x). \quad (9)$$

Let us define the set-valued mapping $L_\pi : \mathcal{P}(\mathsf{X}) \to 2^{\mathcal{P}(\mathsf{X})}$ as follows: given $\mu \in \mathcal{P}(\mathsf{X})$, a probability measure $\hat{\mu} \in L_\pi(\mu)$ if it is an invariant distribution of the transition probability $\int_{\mathsf{A}} p(\,\cdot\,|x,a,\mu)\,\pi(da|x)$; that is

$$\hat{\mu}(\,\cdot\,) = \sum_{x\in\mathsf{X}} \int_{\mathsf{A}} p(\,\cdot\,|x,a,\mu)\,\pi(da|x)\,\hat{\mu}(x).$$

Note that the transition probability $\int_{\mathsf{A}} p(\,\cdot\,|x,a,\mu)\,\pi(da|x)$ is Feller continuous, and since X is finite, the sequence of n-step transition probabilities are tight for any $x \in \mathsf{X}$. Therefore, we can apply Krylov-Bogoliubov theorem (Hairer, 2006, Theorem 4.17), and obtain that $L_\pi(\mu)$ is non-empty for each $\mu \in \mathcal{P}(\mathsf{X})$. Moreover, $L_\pi(\mu)$ is also convex for each $\mu \in \mathcal{P}(\mathsf{X})$. If we can prove that L_π has a closed graph, by Kakutani's fixed point theorem (Aliprantis and Border, 2006, Corollary 17.55), we can conclude that L_π has a fixed point $\hat{\mu}$; that is, $\hat{\mu}$ satisfies (9). Hence, $\hat{\mu} \in \Lambda(\pi)$.

To this end, let $(\mu_n, \hat{\mu}_n) \to (\mu, \hat{\mu})$, where $\hat{\mu}_n \in L_\pi(\mu_n)$ for each n. Note that L_π has a closed graph if $\hat{\mu} \in L_\pi(\mu)$. For each n, we have

$$\hat{\mu}_n(y) = \sum_{x \in \mathsf{X}} \int_{\mathsf{A}} p(y|x, a, \mu_n)\, \pi(da|x)\, \hat{\mu}_n(x), \ \forall y \in \mathsf{X}.$$

For all $y \in \mathsf{X}$, the left part of the above equation converges to $\mu(y)$ since $\hat{\mu}_n \to \hat{\mu}$ and the right part of the same equation converges to

$$\sum_{x \in \mathsf{X}} \int_{\mathsf{A}} p(y|x, a, \mu)\, \pi(da|x)\, \hat{\mu}(x)$$

since $\hat{\mu}_n \to \hat{\mu}$ and $\int_{\mathsf{A}} p(y|\cdot, a, \mu_n)\, \pi(da|\cdot)$ converges to $\int_{\mathsf{A}} p(y|\cdot, a, \mu)\, \pi(da|\cdot)$ continuously by Assumption 1 (Langen, 1981, Theorem 3.5)[4]. Hence, we have

$$\hat{\mu}(\cdot) = \sum_{x \in \mathsf{X}} \int_{\mathsf{A}} p(\cdot|x, a, \mu)\, \pi(da|x)\, \hat{\mu}(x).$$

In other words, $\hat{\mu} \in L_\pi(\mu)$, and so, L_π has a closed graph. This completes the proof. ■

3. Mean-field equilibrium operator for discounted-cost

In this section, we introduce a mean-field equilibrium (MFE) operator for discounted-cost, whose fixed point is a mean-field equilibrium. We prove that this operator is a contraction. Using this result, we then establish the convergence of the learning algorithm that gives approximate mean-field equilibrium policy. To that end, in addition to Assumption 1, we impose the assumption below. But, before that, let us define the constants:

$$c_{\mathbf{m}} \coloneqq \|c\|_\infty, \ Q_{\mathbf{m}} \coloneqq \frac{c_{\mathbf{m}}}{1-\beta}, \ Q_{\text{Lip}} \coloneqq \frac{L_1}{1 - \beta K_1/2}. \tag{10}$$

Assumption 2 *We assume that*

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

where $Q_{\text{Lip}} > 0$.

This assumption is used to ensure that the MFE operator is a contraction, which is crucial to establish the error analysis of the learning algorithm. Note that using Banach fixed point theorem, one can also compute the mean-field equilibrium by applying the MFE

4. Suppose g, g_n ($n \geq 1$) are uniformly bounded measurable functions on metric space E. The sequence of functions g_n is said to converge to g continuously if $\lim_{n\to\infty} g_n(e_n) = g(e)$ for any $e_n \to e$ where $e \in \mathsf{E}$. In this case, (Langen, 1981, Theorem 3.5) states that if $\mu_n \to \mu$ weakly, then $\int_{\mathsf{E}} g_n(e)\, \mu_n(de) \to \int_{\mathsf{E}} g(e)\, \mu(de)$. If E is finite with discrete metric, then weak convergence of probability measures on E is equivalent to l_1-convergence.

operator recursively to obtain successive approximations (i.e., Picard iteration). However, even if it is restrictive, it is not possible to prove convergence of the learning algorithm without a contraction condition. In addition, imposing a contraction condition is a common method in learning mean-field games (see Guo et al. (2019); Fu et al. (2019)).

Note that, given any state-measure μ, the value function J_μ of policy π with initial state x is given by

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

Then, the optimal value function is defined as $J^*_\mu(x) := \inf_{\pi\in\Pi} J_\mu(\pi, x)$ for all $x \in \mathsf{X}$. Using J^*_μ, we can characterize the set of optimal policies $\Psi(\mu)$ for μ as follows. Firstly, $J^*_\mu(x)$ is the unique fixed point of the Bellman optimality operator T_μ, which is a β-contraction with respect to the $\|\cdot\|_\infty$-norm:

$$J^*_\mu(x) = \min_{a\in\mathsf{A}}\left[c(x, a, \mu) + \beta \sum_{y\in\mathsf{X}} J^*_\mu(y)\, p(y|x, a, \mu)\right] =: T_\mu J^*_\mu(x).$$

Additionally, if $f^* : \mathsf{X} \to \mathsf{A}$ attains the minimum in the equation above for all $x \in \mathsf{X}$ as follows

$$\min_{a\in\mathsf{A}}\left[c(x, a, \mu) + \beta \sum_{y\in\mathsf{X}} J^*_\mu(y)\, p(y|x, a, \mu)\right] = c(x, f^*(x), \mu) + \beta \sum_{y\in\mathsf{X}} J^*_\mu(y)\, p(y|x, f^*(x), \mu),$$

then the policy $\pi^*(a|x) = \delta_{f^*(x)}(a) \in \Pi_d$ is optimal for μ and for any initial distribution η_0. We refer the reader to (Hernández-Lerma and Lasserre, 1996, Chapter 4) and (Hernández-Lerma and Lasserre, 1999, Chapter 8) for these classical results in MDP theory.

We can also obtain a similar characterization by using the optimal Q-function instead of the optimal value function J^*_μ. Indeed, we define the optimal Q-function as

$$Q^*_\mu(x, a) = c(x, a, \mu) + \beta \sum_{y\in\mathsf{X}} J^*_\mu(y)\, p(y|x, a, \mu).$$

Note that $Q^*_{\mu,\min}(x) := \min_{a\in\mathsf{A}} Q^*_\mu(x, a) = J^*_\mu(x)$ for all $x \in \mathsf{X}$, and so, we have

$$Q^*_\mu(x, a) = c(x, a, \mu) + \beta \sum_{y\in\mathsf{X}} Q^*_{\mu,\min}(y)\, p(y|x, a, \mu) =: H_\mu Q^*_\mu(x, a),$$

where H_μ is the Bellman optimality operator for Q-functions. It is straightforward to prove that H_μ is a $\|\cdot\|_\infty$-contraction with modulus β and the unique fixed point of H_μ is Q^*_μ. Hence, we can develop a Q-iteration algorithm to compute Q^*_μ, and the policy $\pi^*(a|x) = \delta_{f^*(x)}(a) \in \Pi_d$ is optimal for μ and for any initial distribution η_0, if $Q^*_\mu(x, f^*(x)) = Q^*_{\mu,\min}(x)$ for all $x \in \mathsf{X}$. The advantage of Q-iteration algorithm is that one can adapt this algorithm to the model-free setting via Q-learning.

Let us recall the following fact about l_1 norm on the set probability distributions on finite sets (Georgii, 2011, p. 141). Suppose that there exists a real valued function g on a finite set E. Then, for any pair of probability distributions μ, ν on E, we have

$$\left|\sum_e g(e)\,\mu(e) - \sum_e g(e)\,\nu(e)\right| \le \frac{\operatorname{span}(g)}{2}\,\|\mu - \nu\|_1, \tag{11}$$

where $\mathrm{span}(g) \coloneqq \sup_{e\in\mathsf{E}} g(e) - \inf_{e\in\mathsf{E}} g(e)$ is the span-seminorm. Using this result, we can prove the following fact about optimal value functions.

Lemma 4 *For any $\mu \in \mathcal{P}(\mathsf{X})$, the optimal value function $Q^*_{\mu,\min}$ is Q_{Lip}-Lipschitz continuous; that is,*

$$|Q^*_{\mu,\min}(x) - Q^*_{\mu,\min}(y)| \le Q_{\mathrm{Lip}}\, d_{\mathsf{X}}(x,y).$$

Proof Fix any $\mu \in \mathcal{P}(\mathsf{X})$. Let $u : \mathsf{X} \to \mathbb{R}$ be K-Lipschitz continuous for some $0 < K < L_1$. Then $g = u/K$ is 1-Lipschitz continuous and therefore, for all $a \in \mathsf{A}$ and $z, y \in \mathsf{X}$ we have

$$\begin{aligned}\left|\sum_x u(x)p(x|z,a,\mu) - \sum_x u(x)p(x|y,a,\mu)\right| &= K\left|\sum_x g(x)p(x|z,a,\mu) - \sum_x g(x)p(x|y,a,\mu)\right| \\ &\le \frac{K}{2}\,\|p(\,\cdot\,|z,a,\mu) - p(\,\cdot\,|y,a,\mu)\|_1 \quad \text{(by (11))} \\ &\le \frac{KK_1}{2}\, d_{\mathsf{X}}(z,y), \quad \text{(by Assumption 1)}\end{aligned}$$

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

$$\begin{aligned}&|T_\mu u(z) - T_\mu u(y)| \\ &\quad\le \sup_a\left\{|c(z,a,\mu) - c(y,a,\mu)| + \beta\left|\sum_x u(x)p(x|z,a,\mu) - \sum_x u(x)p(x|y,a,\mu)\right|\right\} \\ &\quad\le L_1 d_{\mathsf{X}}(z,y) + \beta\frac{KK_1}{2}d_{\mathsf{X}}(z,y) = \left(L_1 + \beta\frac{KK_1}{2}\right)d_{\mathsf{X}}(z,y).\end{aligned}$$

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

Before introducing the mean-field equilibrium (MFE) operator, we first define the set $\mathcal{C}$ on which the Q-functions live. We let $\mathcal{C}$ consist of all Q-functions $Q : \mathsf{X}\times\mathsf{A} \to \mathbb{R}$ such that $Q(x,\cdot)$ is Q_{Lip}-Lipschitz and ρ-strongly convex for every $x \in \mathsf{X}$ with $\|Q\|_\infty \le Q_{\mathbf{m}}$, and the gradient ∇Q of Q with respect to a satisfies the bound

$$\sup_{a\in\mathsf{A}} \|\nabla Q(x,a) - \nabla Q(\hat{x},a)\| \le K_F\, d_{\mathsf{X}}(x,\hat{x}),$$

for every $x, \hat{x} \in \mathsf{X}$.

The MFE operator defined as a composition of the operators H_1 and H_2, where $H_1 : \mathcal{P}(\mathsf{X}) \to \mathcal{C}$ is defined as $H_1(\mu) = Q^*_\mu$ (optimal Q-function for μ), and $H_2 : \mathcal{P}(\mathsf{X})\times\mathcal{C} \to \mathcal{P}(\mathsf{X})$ is defined as

$$H_2(\mu, Q)(\cdot) \coloneqq \sum_{x\in\mathsf{X}} p(\cdot|x, f_Q(x), \mu)\,\mu(x), \tag{12}$$

where $f_Q(\,\cdot\,) \coloneqq \arg\min_{a\in\mathsf{A}} Q(\,\cdot\,,a)$ is the unique minimizer of $Q \in \mathcal{C}$ by ρ-strong convexity. Here, H_1 computes the optimal Q-function given the current state-measure, and H_2 computes the next state-measure given the current state-measure and the corresponding optimal Q-function. Therefore, the MFE operator is given by

$$H : \mathcal{P}(\mathsf{X}) \ni \mu \mapsto H_2\left(\mu, H_1(\mu)\right) \in \mathcal{P}(\mathsf{X}). \tag{13}$$

In this section, our goal is to prove that H is a contraction.

Remark 5 *Note that we can alternatively define the operator H_2 as a mapping from $\mathcal{C}$ to $\mathcal{P}(\mathsf{X})$ as follows: $H_2(Q) = \mu$ if μ satisfies the following fixed point equation:*

$$\mu(\cdot) \coloneqq \sum_{x\in\mathsf{X}} p(\cdot|x, f_Q(x), \mu)\, \mu(x).$$

Notice that H_2 is a well-defined operator since such a state-measure $\mu \in \mathcal{P}(\mathsf{X})$ exists for any Q by Lemma 3. Hence, we may define the MFE operator as $H(\mu) \coloneqq H_2(H_1(\mu))$. In this case, one can prove that this operator has the same contraction coefficient as the original MFE operator given in (13). However, although the original H_2 operator in (12) can effortlessly be approximated via computing the empirical estimate of $p(\cdot|x, f_Q(x), \mu)$ for each $x \in \mathsf{X}$, which is possible since $|\mathsf{X}| < \infty$, approximating the new H_2 operator is quite costly. Indeed, we need to compute a fixed point of some equation in this case. Therefore, there is no advantage to replace original H_2 with the new one.

In the following lemma, we prove that H_1 is Lipschitz continuous, which will later used to prove that H operator is a contraction.

Lemma 6 *The mapping H_1 is Lipschitz continuous on $\mathcal{P}(\mathsf{X})$ with the Lipschitz constant K_{H_1}, where*

$$K_{H_1} \coloneqq \frac{Q_{\mathrm{Lip}}}{1-\beta}$$

.

Proof First of all, H_1 is well-defined; that is, it maps any $\mu \in \mathcal{P}(\mathsf{X})$ into $\mathcal{C}$. Indeed, recall that Q^*_μ is the fixed point of the contractive operator H_μ:

$$Q^*_\mu(x,a) = c(x,a,\mu) + \beta \sum_{y\in\mathsf{X}} Q^*_{\mu,\min}(y)\, p(y|x,a,\mu).$$

Then, using Assumption 1-(a),(b),(d), it is straightforward to prove that $H_1(\mu) \in \mathcal{C}$. Indeed, the only non-trivial fact is the Q_{Lip}-Lipschitz continuity of Q^*_μ on $\mathsf{X}\times\mathsf{A}$. To this end, let $(x,a),(\hat{x},\hat{a}) \in \mathsf{X}\times\mathsf{A}$ be arbitrary. Then,

$$\begin{aligned}
&|Q^*_\mu(x,a) - Q^*_\mu(\hat{x},\hat{a})| \\
&\qquad = |c(x,a,\mu) + \beta \sum_y Q^*_{\mu,\min}(y) p(y|x,a,\mu) - c(\hat{x},\hat{a},\mu) - \beta \sum_y Q^*_{\mu,\min}(y) p(y|\hat{x},\hat{a},\mu)| \\
&\qquad \leq L_1(d_{\mathsf{X}}(x,\hat{x}) + \|a-\hat{a}\|) + \beta \frac{K_1 Q_{\mathrm{Lip}}}{2}\left(d_{\mathsf{X}}(x,\hat{x}) + \|a-\hat{a}\|\right),
\end{aligned}$$

where the last inequality follows from (11) and Lemma 4. Hence, Q^*_μ is Q_{Lip}-Lipschitz continuous.

Now, we prove that H_1 is K_{H_1}-Lipschitz on $\mathcal{P}(\mathsf{X})$. For any $\mu, \hat{\mu} \in \mathcal{P}(\mathsf{X})$, we have

$$\begin{aligned}
&\|H_1(\mu) - H_1(\hat{\mu})\|_\infty = \|Q^*_\mu - Q^*_{\hat{\mu}}\|_\infty \\
&= \sup_{x,a} \left| c(x,a,\mu) + \beta \sum_y Q^*_{\mu,\min}(y) p(y|x,a,\mu) - c(x,a,\hat{\mu}) - \beta \sum_y Q^*_{\hat{\mu},\min}(y) p(y|x,a,\hat{\mu}) \right| \\
&\leq L_1 \|\mu - \hat{\mu}\|_1 + \beta \left| \sum_y Q^*_{\mu,\min}(y) p(y|x,a,\mu) - \sum_y Q^*_{\mu,\min}(y) p(y|x,a,\hat{\mu}) \right| \\
&\qquad\qquad + \beta \left| \sum_y Q^*_{\mu,\min}(y) p(y|x,a,\hat{\mu}) - \sum_y Q^*_{\hat{\mu},\min}(y) p(y|x,a,\hat{\mu}) \right| \\
&\leq L_1 \|\mu - \hat{\mu}\|_1 + \beta \frac{K_1 Q_{\text{Lip}}}{2} \|\mu - \hat{\mu}\|_1 + \beta \|Q^*_\mu - Q^*_{\hat{\mu}}\|_\infty,
\end{aligned}$$

where the last inequality follows from (11) and Lemma 4. ■

Now, using Lemma 6, we can prove that H is a contraction on $\mathcal{P}(\mathsf{X})$.

Proposition 7 *The mapping H is a contraction with contraction on $\mathcal{P}(\mathsf{X})$ constant K_H, where*

$$K_H \coloneqq \frac{3K_1}{2}\left(1 + \frac{K_F}{\rho}\right) + \frac{K_1 K_F K_{H_1}}{\rho}.$$

Proof Fix any $\mu, \hat{\mu} \in \mathcal{P}(\mathsf{X})$. Note that, since $Q^*_\mu = H_\mu Q^*_\mu$, the mapping $f_{Q^*_\mu}(x)$ is the unique minimizer of $F(x, Q^*_{\mu,\min}, \mu, \cdot)$. Similarly, $f_{Q^*_{\hat{\mu}}}(y)$ is the unique minimizer of $F(y, Q^*_{\hat{\mu},\min}, \hat{\mu}, \cdot)$. For any $x, y \in \mathsf{X}$, we define $a = f_{Q^*_\mu}(x)$ and $r = f_{Q^*_{\hat{\mu}}}(y) - f_{Q^*_\mu}(x)$. As a is the unique minimizer of a strongly convex function $F(x, Q^*_{\mu,\min}, \mu, \cdot)$, by the first-order optimality condition, we have

$$\nabla F\left(x, Q^*_{\mu,\min}, \mu, a\right) \cdot r \geq 0.$$

For $a + r$ and $F(y, Q^*_{\hat{\mu},\min}, \hat{\mu}, \cdot)$, by first-order optimality condition, we also have

$$\nabla F\left(y, Q^*_{\hat{\mu},\min}, \hat{\mu}, a + r\right) \cdot r \leq 0.$$

Therefore, by ρ-strong convexity of F in Assumption 1-(d) and (Hajek and Raginsky, 2019, Lemma 3.2), we have

$$\begin{aligned}
-\nabla F(y, Q^*_{\hat{\mu},\min}, \hat{\mu}, a) \cdot r &\geq -\nabla F(y, Q^*_{\hat{\mu},\min}, \hat{\mu}, a) \cdot r + \nabla F(y, Q^*_{\hat{\mu},\min}, \hat{\mu}, a + r) \cdot r \\
&\geq \rho \|r\|^2. \qquad (14)
\end{aligned}$$

Similarly, by Assumption 1-(d), we also have

$$\begin{aligned}
-\nabla F(y, Q^*_{\hat{\mu},\min}, \hat{\mu}, a) \cdot r &\leq -\nabla F(y, Q^*_{\hat{\mu},\min}, \hat{\mu}, a) \cdot r + \nabla F(x, Q^*_{\mu,\min}, \mu, a) \cdot r \\
&\leq \|r\| \, \|\nabla F(x, Q^*_{\mu,\min}, \mu, a) - \nabla F(y, Q^*_{\hat{\mu},\min}, \hat{\mu}, a)\|
\end{aligned}$$

$$\leq K_F \|r\| \left(d_{\mathsf{X}}(x,y) + \|Q^*_{\mu,\min} - Q^*_{\hat{\mu},\min}\|_\infty + \|\mu - \hat{\mu}\|_1\right)$$
$$\leq K_F \|r\| \left(d_{\mathsf{X}}(x,y) + \|Q^*_{\mu} - Q^*_{\hat{\mu}}\|_\infty + \|\mu - \hat{\mu}\|_1\right). \tag{15}$$

Combining (14) and (15) yields

$$\rho \|r\|^2 \leq K_F \|r\| \left(d_{\mathsf{X}}(x,y) + \|Q^*_{\mu} - Q^*_{\hat{\mu}}\|_\infty + \|\mu - \hat{\mu}\|_1\right).$$

Since $r = f_{Q^*_{\hat{\mu}}}(y) - f_{Q^*_{\mu}}(x)$, we obtain

$$\begin{aligned}
\|f_{Q^*_{\hat{\mu}}}(y) - f_{Q^*_{\mu}}(x)\| &\leq \frac{K_F}{\rho} \left(d_{\mathsf{X}}(x,y) + \|Q^*_{\mu} - Q^*_{\hat{\mu}}\|_\infty + \|\mu - \hat{\mu}\|_1\right) \\
&= \frac{K_F}{\rho} \left(d_{\mathsf{X}}(x,y) + \|H_1(\mu) - H_1(\hat{\mu})\|_\infty + \|\mu - \hat{\mu}\|_1\right) \\
&\leq \frac{K_F}{\rho} \left(d_{\mathsf{X}}(x,y) + K_{H_1}\|\mu - \hat{\mu}\|_1 + \|\mu - \hat{\mu}\|_1\right). \qquad (16)
\end{aligned}$$

Therefore, $f_{Q^*_{\mu}}(x)$ is Lipschitz continuous with respect to (x,μ).

Now, using (16), we have

$$\begin{aligned}
\|H_2(\mu, H_1(\mu)) - H_2(\hat{\mu}, H_1(\hat{\mu}))\|_1 &= \sum_y \Bigg| \sum_x p(y|x, f_{Q^*_{\mu}}(x), \mu)\, \mu(x) \\
&\qquad - \sum_x p(y|x, f_{Q^*_{\hat{\mu}}}(x), \hat{\mu})\, \hat{\mu}(x) \Bigg| \\
&\leq \sum_y \Bigg| \sum_x p(y|x, f_{Q^*_{\mu}}(x), \mu)\, \mu(x) \\
&\qquad - \sum_x p(y|x, f_{Q^*_{\hat{\mu}}}(x), \hat{\mu})\, \mu(x) \Bigg| \\
&\quad + \sum_y \Bigg| \sum_x p(y|x, f_{Q^*_{\hat{\mu}}}(x), \hat{\mu})\, \mu(x) \\
&\qquad - \sum_x p(y|x, f_{Q^*_{\hat{\mu}}}(x), \hat{\mu})\, \hat{\mu}(x) \Bigg| \\
&\overset{(I)}{\leq} \sum_x \left\| p(\cdot|x, f_{Q^*_{\mu}}(x), \mu) - p(\cdot|x, f_{Q^*_{\hat{\mu}}}(x), \hat{\mu}) \right\|_1 \mu(x) \\
&\qquad + \frac{K_1}{2} \left(1 + \frac{K_F}{\rho}\right) \|\mu - \hat{\mu}\|_1 \\
&\leq K_1 \left(\|f_{Q^*_{\mu}}(x) - f_{Q^*_{\hat{\mu}}}(x)\| + \|\mu - \hat{\mu}\|_1 \right) \\
&\qquad + \frac{K_1}{2} \left(1 + \frac{K_F}{\rho}\right) \|\mu - \hat{\mu}\|_1 \\
&\leq K_H \|\mu - \hat{\mu}\|_1. \qquad (17)
\end{aligned}$$

Note that (16) and Assumption 1-(b) lead to

$$\|p(\cdot|x, f_{Q^*_{\hat{\mu}}}(x), \hat{\mu}) - p(\cdot|y, f_{Q^*_{\hat{\mu}}}(y), \hat{\mu})\|_1 \leq K_1 \left(1 + \frac{K_F}{\rho}\right).$$

Hence, (I) follows from Lemma 32. This completes the proof. ■

Remark 8 *Note that in the MDP theory, it is normally not required to establish the Lipschitz continuity of the optimal policy. Indeed, the Lipschitz continuity of the optimal value function is in general needed, which can be established straightforward as in Lemma 4. However, in mean-field games, since the optimal policy $f_{Q^*_\mu}$ directly affects the behaviour of the next state-measure through*

$$H_2(\mu, Q^*_\mu)(\cdot) = \sum_x p(\cdot|x, f_{Q^*_\mu}, \mu)\, \mu(x),$$

one must also establish the Lipschitz continuity of the optimal policy $f_{Q^*_\mu}$ in this case. This is indeed the key point in the proof of Proposition 7.

In Anahtarci et al. (2020b), we establish the Lipschitz continuity of the optimal policy by introducing a regularization term into a one-stage cost function. This significantly relaxes conditions on the system components in Assumption 1-(d) and simplifies the analysis. However, the regularization term adds some bias to the equilibrium solution (i.e., it in general favours randomized policies over deterministic policies) and also causes the regularized stationary mean-field equilibrium to deviate from true stationary mean-field equilibrium as a result of the additional regularization term in the one-stage cost function.

Under Assumption 1 and Assumption 2, H is a contraction mapping. Therefore, by the Banach fixed point theorem, H has a unique fixed point. Let $\mu_* \in \mathcal{P}(\mathsf{X})$ be this unique fixed point and let $Q^*_{\mu_*} = H_1(\mu_*)$. Define the policy $\pi_*(\cdot\,|x) = \delta_{f_{Q^*_{\mu_*}}(x)}(\cdot\,)$. Then, one can prove that the pair (π_*, μ_*) is a mean-field equilibrium. Indeed, note that $(\mu_*, Q^*_{\mu_*})$ satisfies the following equations

$$\mu_*(\cdot) = \sum_{x\in\mathsf{X}} p(\cdot|x, a, \mu_*)\, \pi_*(a|x)\, \mu_*(x), \tag{18}$$

$$Q^*_{\mu_*}(x, a) = c(x, a, \mu_*) + \beta \sum_{y\in\mathsf{X}} Q^*_{\mu_*,\min}(y)\, p(y|x, a, \mu_*). \tag{19}$$

Here, (19) implies that $\pi_* \in \Psi(\mu_*)$ and (18) implies that $\mu_* \in \Lambda(\pi_*)$. Hence, (π_*, μ_*) is a mean-field equilibrium. Hence, we can compute this mean-field equilibrium via applying H recursively starting from arbitrary state-measure. 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 cost function c 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, as stated before.

4. Finite-Agent Game for Discounted-cost

The mean-field game model defined in Section 2 is indeed the infinite-population version of the finite-agent game model with mean-field interactions, which will be described in this

section. In this model, there are N-agents and for every time step $t \in \{0, 1, 2, \dots\}$ and every agent $i \in \{1, 2, \dots, N\}$, $x_i^N(t) \in \mathsf{X}$ and $a_i^N(t) \in \mathsf{A}$ denote the state and the action of Agent i at time t, respectively. Moreover,

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

denote the empirical distribution of the agents' states at time t. For each $t \geq 0$, next states $(x_1^N(t+1), \dots, x_N^N(t+1))$ of agents have the following conditional distribution given current states $(x_1^N(t), \dots, x_N^N(t))$ and actions $(a_1^N(t), \dots, a_N^N(t))$:

$$(x_1^N(t+1), \dots, x_N^N(t+1)) \sim \bigotimes_{i=1}^{N} p\big(\,\cdot\,\big|x_i^N(t), a_i^N(t), e_t^{(N)}\big).$$

A *policy* π for a generic agent in this model is a conditional distribution on A given X; that is, agents can only use their individual states to design their actions. The set of all policies for Agent i is denoted by Π_i. Hence, under $\pi \in \Pi_i$, the conditional distribution of the action $a_i^N(t)$ of Agent i at time t given its state $x_i^N(t)$ is

$$a_i^N(t) \sim \pi(\,\cdot\,|x_i^N(t)).$$

Therefore, the information structure of the problem is decentralized. The initial states $\{x_i^N(0)\}_{i=1}^N$ are independent and identically distributed according to the initial distribution η_0.

We let $\boldsymbol{\pi}^{(N)} := (\pi^1, \dots, \pi^N)$, $\pi^i \in \Pi_i$, denote an N-tuple of policies for all the agents in the game. Under such an N-tuple of policies, for Agent i, the discounted-cost is given by

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

Since agents are coupled through their dynamics and cost functions via the empirical distribution of the states, the problem is indeed a classical game problem. Therefore, the standard notion of optimality is a player-by-player one.

Definition 9 *An N-tuple of policies $\boldsymbol{\pi}^{(N*)} = (\pi^{1*}, \dots, \pi^{N*})$ constitutes a* Nash equilibrium *if*

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

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

We note that obtaining a Nash equilibria is in general prohibitive for finite-agent game model due to the decentralized nature of the information structure of the problem and the large number of agents (see (Saldi et al., 2018, pp. 4259)). Therefore, it is of interest to seek an approximate Nash equilibrium, whose definition is given below.

Definition 10 *An N-tuple of policies $\boldsymbol{\pi}^{(N*)} = (\pi^{1*}, \dots, \pi^{N*})$ constitutes a* δ-Nash equilibrium *if*

$$J_i^{(N)}(\boldsymbol{\pi}^{(N*)}) \leq \inf_{\pi^i \in \Pi_i} J_i^{(N)}(\boldsymbol{\pi}_{-i}^{(N*)}, \pi^i) + \delta$$

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

In finite-agent mean-field game model, if the number of agents is large enough, one can obtain a δ-Nash equilibrium by studying the infinite-population limit $N \to \infty$ of the game (i.e., mean-field game). In the infinite-agent case, the empirical distribution of the states can be modelled as an exogenous state-measure, which should be consistent with the distribution of a generic agent by the law of large numbers (i.e., mean-field equilibrium); that is, a generic agent should solve the mean-field game that is introduced in the preceding section. Then, it is possible to prove that if each agent in the finite-agent N game problem adopts the policy in mean-field equilibrium, the resulting N-tuple of policies will be an approximate Nash equilibrium for all sufficiently large N. This was indeed proved in Saldi et al. (2018).

Note that it is also possible to prove that if each agent in the finite-agent game model adopts the ε-mean-field equilibrium policy, the resulting policy will be also an approximate Nash equilibrium for all sufficiently large N-agent game models. Indeed, this is the statement of the next theorem.

But before, let us define the following constants:

$$C_1 \coloneqq \left(\frac{3\,K_1}{2} + \frac{K_1\,K_F}{2\rho}\right),\; C_2 \coloneqq \left(L_1 + \frac{\beta K_1 Q_{\mathrm{Lip}}}{2}\right)\frac{K_1}{1-C_1},\; C_3 \coloneqq \left(L_1 + \frac{\beta K_1 Q_{\mathrm{Lip}}}{2}\right).$$

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

Theorem 11 *Let π_ε be an ε-mean-field equilibrium policy for the mean-field equilibrium $(\pi_*, \mu_*) \in \Pi_d \times \mathcal{P}(\mathsf{X})$ given by the unique fixed point of the MFE operator H. Let $\eta_0 \in \Lambda(\pi_\varepsilon)$. 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 $\boldsymbol{\pi}^{(N)} = \{\pi_\varepsilon, \pi_\varepsilon, \dots, \pi_\varepsilon\}$ is a $(\delta + \tau\varepsilon)$-Nash equilibrium for the game with N agents, where $\tau \coloneqq \dfrac{2C_2 + C_3}{1-\beta}$.*

Proof By an abuse of notation, we denote the deterministic mappings from X to A that induce policies π_* and π_ε as π_* and π_ε as well, respectively. Note that in view of (16), one can prove that

$$\|\pi_*(x) - \pi_*(y)\| \leq \frac{K_F}{\rho}\, d_{\mathsf{X}}(x, y). \tag{20}$$

Let $\mu_\varepsilon \in \Lambda(\pi_\varepsilon)$. Then, we have

$$\|\mu_\varepsilon - \mu_*\|_1 = \sum_y \left| \sum_x p(y|x, \pi_\varepsilon(x), \mu_\varepsilon)\, \mu_\varepsilon(x) - \sum_x p(y|x, \pi_*(x), \mu_*)\, \mu_*(x) \right|$$

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

Note that (20) and Assumption 1 lead to

$$
\|p(\cdot|x,\pi_*(x),\mu_*) - p(\cdot|y,\pi_*(y),\mu_*)\|_1 \leq K_1\left(1+\frac{K_F}{\rho}\right) d_{\mathsf{X}}(x,y).
$$

Hence, (I) follows from Lemma 32. Therefore, we have

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

Now, fix any policy $\pi \in \Pi_d$. For any state-measure μ, it is a well-known fact in MDP theory that the value function $J_\mu(\pi,\cdot)$ of π satisfies the following fixed point equation:

$$
J_\mu(\pi,x) = c(x,\pi(x),\mu) + \beta \sum_y J_\mu(\pi,y)\,p(y|x,\pi(x),\mu),
$$

for every $x \in \mathsf{X}$. Therefore, we have

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

Here (II) follows from (11) and the fact that $J_{\mu_*}(\pi,\cdot)$ is Q_{Lip}-Lipschitz continuous, which can be proved as in Lemma 4. Therefore, we obtain

$$\|J_{\mu_*}(\pi,\cdot) - J_{\mu_\varepsilon}(\pi,\cdot)\|_\infty \le \frac{C_2\,\varepsilon}{1-\beta}. \tag{21}$$

Similarly, we also have

$$\begin{aligned}
\|J_{\mu_*}(\pi_*,\cdot) - J_{\mu_*}(\pi_\varepsilon,\cdot)\|_\infty &= \sup_x \Big| c(x,\pi_*(x),\mu_*) + \beta \sum_y J_{\mu_*}(\pi_*,y)\, p(y|x,\pi_*(x),\mu_*) \\
&\qquad - c(x,\pi_\varepsilon(x),\mu_*) - \beta \sum_y J_{\mu_*}(\pi_\varepsilon,y)\, p(y|x,\pi_\varepsilon(x),\mu_*)\Big| \\
&\le L_1 \sup_x \|\pi_*(x) - \pi_\varepsilon(x)\| + \beta \sup_x \Big| \sum_y J_{\mu_*}(\pi_*,y)\, p(y|x,\pi_*(x),\mu_*) \\
&\qquad - \sum_y J_{\mu_*}(\pi_*,y)\, p(y|x,\pi_\varepsilon(x),\mu_*)\Big| \\
&\quad + \beta \sup_x \Big| \sum_y J_{\mu_*}(\pi_*,y)\, p(y|x,\pi_\varepsilon(x),\mu_*) \\
&\qquad - \sum_y J_{\mu_*}(\pi_\varepsilon,y)\, p(y|x,\pi_\varepsilon(x),\mu_*)\Big| \\
&\overset{(III)}{\le} \left(L_1 + \frac{\beta K_1 Q_{\text{Lip}}}{2}\right) \sup_x \|\pi_*(x) - \pi_\varepsilon(x)\| \\
&\qquad + \beta \|J_{\mu_*}(\pi_*,\cdot) - J_{\mu_*}(\pi_\varepsilon,\cdot)\|_\infty \\
&\le \left(L_1 + \frac{\beta K_1 Q_{\text{Lip}}}{2}\right) \varepsilon + \beta \|J_{\mu_*}(\pi_*,\cdot) - J_{\mu_*}(\pi_\varepsilon,\cdot)\|_\infty.
\end{aligned}$$

Here (III) follows from (11) and the fact that $J_{\mu_*}(\pi_*,\cdot)$ is Q_{Lip}-Lipschitz continuous, which can be proved as in Lemma 4. Therefore, we obtain

$$\|J_{\mu_*}(\pi_*,\cdot) - J_{\mu_*}(\pi_\varepsilon,\cdot)\|_\infty \le \frac{C_3\varepsilon}{1-\beta}, \tag{22}$$

where $C_3 \coloneqq \left(L_1 + \frac{\beta K_1 Q_{\text{Lip}}}{2}\right)$.

Note that we must prove that

$$J_i^{(N)}(\boldsymbol{\pi}^{(N)}) \le \inf_{\pi^i \in \Pi_i} J_i^{(N)}(\boldsymbol{\pi}_{-i}^{(N)}, \pi^i) + \tau\,\varepsilon + \delta \tag{23}$$

for each $i = 1,\dots,N$, when N is sufficiently large. As the transition probabilities and the one-stage cost functions are the same for all agents, it is sufficient to prove (23) for Agent 1 only. Given $\delta > 0$, for each $N \ge 1$, let $\tilde{\pi}^{(N)} \in \Pi_1$ be a deterministic policy such that

$$J_1^{(N)}(\tilde{\pi}^{(N)}, \pi_\varepsilon, \dots, \pi_\varepsilon) < \inf_{\pi' \in \Pi_1} J_1^{(N)}(\pi', \pi_\varepsilon, \dots, \pi_\varepsilon) + \frac{\delta}{3}.$$

On the other hand, by (Saldi et al., 2018, Theorem 4.10)

$$\begin{aligned}
\lim_{N\to\infty} J_1^{(N)}(\tilde{\pi}^{(N)}, \pi_\varepsilon, \dots, \pi_\varepsilon) &= \lim_{N\to\infty} J_{\mu_\varepsilon}(\tilde{\pi}^{(N)}) \\
&\geq \lim_{N\to\infty} J_{\mu_*}(\tilde{\pi}^{(N)}) - \frac{C_2\,\varepsilon}{1-\beta} \quad \text{(by (21))} \\
&\geq \inf_{\pi'\in\Pi_d} J_{\mu_*}(\pi') - \frac{C_2\,\varepsilon}{1-\beta} \\
&= J_{\mu_*}(\pi_*) - \frac{C_2\,\varepsilon}{1-\beta} \\
&\geq J_{\mu_*}(\pi_\varepsilon) - \frac{C_2\,\varepsilon}{1-\beta} - \frac{C_3\,\varepsilon}{1-\beta} \quad \text{(by (22))} \\
&\geq J_{\mu_\varepsilon}(\pi_\varepsilon) - \frac{2\,C_2\,\varepsilon}{1-\beta} - \frac{C_3\,\varepsilon}{1-\beta} \quad \text{(by (21))} \\
&=: J_{\mu_\varepsilon}(\pi_\varepsilon) - \tau\,\varepsilon.
\end{aligned}$$

Note that by (Saldi et al., 2018, Theorem 4.10), we also have

$$\lim_{N\to\infty} J_1^{(N)}(\pi_\varepsilon, \pi_\varepsilon, \dots, \pi_\varepsilon) = J_{\mu_\varepsilon}(\pi_\varepsilon).$$

Hence, there exists $N(\delta)$ such that for all $N \geq N(\delta)$, we have

$$\begin{aligned}
J_1^{(N)}(\tilde{\pi}^{(N)}, \pi_\varepsilon, \dots, \pi_\varepsilon) + \frac{\delta}{3} &\geq J_{\mu_\varepsilon}(\pi_\varepsilon) - \tau\,\varepsilon \\
J_{\mu_\varepsilon}(\pi_\varepsilon) + \frac{\delta}{3} &\geq J_1^{(N)}(\pi_\varepsilon, \pi_\varepsilon, \dots, \pi_\varepsilon).
\end{aligned}$$

Therefore, for all $N \geq N(\delta)$, we obtain

$$\begin{aligned}
\inf_{\pi'\in\Pi_1} J_1^{(N)}(\pi', \pi_\varepsilon, \dots, \pi_\varepsilon) + \delta + \tau\,\varepsilon &\geq J_1^{(N)}(\tilde{\pi}^{(N)}, \pi_\varepsilon, \dots, \pi_\varepsilon) + \frac{2\delta}{3} + \tau\,\varepsilon \\
&\geq J_{\mu_\varepsilon}(\pi_\varepsilon) + \frac{\delta}{3} \\
&\geq J_1^{(N)}(\pi_\varepsilon, \pi_\varepsilon, \dots, \pi_\varepsilon).
\end{aligned}$$

■

Theorem 11 implies that, by learning ε-mean-field equilibrium policy in the infinite-population limit, one can obtain an approximate Nash equilibrium for the finite-agent game problem for which computing or learning the exact Nash equilibrium is in general prohibitive.

In the next section, we approximate the MFE operator H introduced in Section 3 via the random operator $\hat{H}$ to develop an algorithm for learning a ε-mean-field equilibrium policy in the model-free setting.

5. Learning Algorithm for Discounted-cost

In this section, we develop an offline learning algorithm to learn an approximate mean-field equilibrium policy. To this end, we suppose that a generic agent has access to a simulator, which generates a new state $y \sim p(\cdot\,|x,a,\mu)$ and gives the cost $c(x,a,\mu)$ for any given state x, action a, and state measure μ. This is a typical assumption in offline reinforcement learning algorithms.

Each iteration of our learning algorithm has two stages. Using a fitted Q-iteration algorithm, we learn the optimal Q-function Q^*_μ for a given state-measure μ in the first stage by replacing H_1 with a random operator $\hat{H}_1$. The Q-functions are selected from a fixed function class $\mathcal{F}$ which can be defined as a collection of neural networks with a specific architecture or a linear span of a finite number of basis functions. There will be an additional representation error in the learning algorithm depending on this choice, which is generally negligible since Q-functions in $\mathcal{C}$ can be well approximated by functions from $\mathcal{F}$.

In the second stage, we update the state-measure by approximating the transition probability via its empirical estimate by replacing H_2 with a random operator $\hat{H}_2$. It should be noted that if the alternative H_2 operator mentioned in Remark 5 is used, the random operator that approximates this alternative H_2 operator would be more complicated than $\hat{H}_2$. Indeed, in this case, an empirical estimation of the transition probability might be insufficient.

We proceed by introducing the random operator $\hat{H}_1$. To describe $\hat{H}_1$, we need to give some definitions. Let $m_{\mathsf{A}}(\cdot) \coloneqq m(\cdot)/m(\mathsf{A})$ be the uniform probability measure on A. Let us choose a probability measure ν on X such that $\min_x \nu(x) > 0$. For instance, one can choose ν as the uniform distribution over X. Define $\zeta_0 \coloneqq 1/\sqrt{\min_x \nu(x)}$. We also choose some policy π_b such that, for any $x \in \mathsf{X}$, the distribution $\pi_b(\cdot|x)$ on A has density with respect to Lebesgue measure m. To simplify the notation, we denote this density by $\pi_b(a|x)$, and assume that it satisfies $\pi_0 \coloneqq \inf_{(x,a)\in\mathsf{X}\times\mathsf{A}} \pi_b(a|x) > 0$. Note that the randomized policy π_b is used to generate data for the learning algorithm below. In general, given any mean-field term, it is enough to consider deterministic policies for optimality. However, as is typical in reinforcement learning, we employ randomized policies to explore the action space in the training stage. Because of this, π_b is introduced in a stochastic manner. We can now define the random operator $\hat{H}_1$.

Algorithm 1 Algorithm $\hat{H}_1$

Input μ, Data size N, Number of iterations L
Generate i.i.d. samples $\{(x_t, a_t, c_t, y_{t+1})_{t=1}^N\}$ using

$$x_t \sim \nu,\ a_t \sim \pi_b(\cdot|x_t),\ c_t = c(x_t, a_t, \mu),\ y_{t+1} \sim p(\cdot|x_t, a_t, \mu)$$

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

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

end for
return Q_L

Remark 12 *Notice that in Algorithm $\hat{H}_1$, we use the distribution ν and policy π_b to build an i.i.d. dataset. In fact, instead of using i.i.d. samples, one can use a sample path $\{x_t, a_t\}_{t=1}^N$ generated by the policy π_b instead of using i.i.d. samples by setting $c_t = c(x_t, a_t, \mu)$ and $y_{t+1} = x_{t+1}$. Then, in order to establish the error analysis, we have to assume that under π_b, the state process $\{x_t\}$ must be strictly stationary and exponentially β-mixing (see Antos et al. (2007a)). The main issue in this case, however, is finding a policy π_b that meets the mixing condition. Indeed, since exponentially β-mixing stationary processes forget their history exponentially fast, they behave like i.i.d. processes when there is a sufficiently large time gap between two samples. As a result, the error analysis of the exponential β-mixing case is very close to that of the i.i.d. case. For more information on the error analysis of $\hat{H}_1$ in the exponentially β-mixing case, see Antos et al. (2007b,a).*

We perform an error analysis of the algorithm $\hat{H}_1$ before defining the second stage $\hat{H}_2$. To that end, we define the l_2-norm of any $g : \mathsf{X} \times \mathsf{A} \to \mathbb{R}$ as

$$\|g\|_\nu^2 \coloneqq \sum_{x \in \mathsf{X}} \int_{\mathsf{A}} g(x,a)^2\, m_{\mathsf{A}}(da)\, \nu(x),$$

and introduce the constants

$$E(\mathcal{F}) \coloneqq \sup_{\mu \in \mathcal{P}(\mathsf{X})} \sup_{Q \in \mathcal{F}} \inf_{Q' \in \mathcal{F}} \|Q' - H_\mu Q\|_\nu, \tag{24}$$

and

$$L_{\mathbf{m}} \coloneqq (1+\beta) Q_{\mathbf{m}} + c_{\mathbf{m}},\ \ C \coloneqq \frac{L_{\mathbf{m}}^2}{m(\mathsf{A})\, \pi_0},\ \ \gamma = 512 C^2.$$

Here $E(\mathcal{F})$ describes the representation error of the function class $\mathcal{F}$. This error is generally small since every Q function in $\mathcal{C}$ can be very well approximated using, for example, neural networks with a fixed architecture. As a result, we may consider the error caused

by $E(\mathcal{F})$ to be negligible. The error analysis of the algorithm $\hat{H}_1$ is given by the following theorem. We define $\mathcal{F}_{\min} \coloneqq \{Q_{\min} : Q \in \mathcal{F}\}$ and let

$$\Upsilon = 8\, e^2\, (V_{\mathcal{F}} + 1)\, (V_{\mathcal{F}_{\min}} + 1)\, \left(\frac{64 e Q_{\mathbf{m}} L_{\mathbf{m}} (1+\beta)}{m(\mathsf{A})\pi_0}\right)^{V_{\mathcal{F}} + V_{\mathcal{F}_{\min}}},\ V = V_{\mathcal{F}} + V_{\mathcal{F}_{\min}}.$$

Theorem 13 *For any $(\varepsilon, \delta) \in (0,1)^2$ and $N \geq m_1(\epsilon, \delta, L)$, with probability at least $1-\delta$ we have*

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

if $\frac{\beta^L}{1-\beta}\, Q_{\mathbf{m}} < \frac{\varepsilon}{2}$, where

$$m_1(\varepsilon, \delta, L) \coloneqq \frac{\gamma (2\Lambda)^{4(\dim_{\mathsf{A}}+1)}}{\varepsilon^{4(\dim_{\mathsf{A}}+1)}}\, \ln\left(\frac{\Upsilon (2\Lambda)^{2V(\dim_{\mathsf{A}}+1)} L}{\delta \varepsilon^{2V(\dim_{\mathsf{A}}+1)}}\right),$$

and

$$\Delta \coloneqq \frac{1}{1-\beta}\left[\frac{m(\mathsf{A})(\dim_{\mathsf{A}}+1)!\zeta_0}{\alpha(2/Q_{\mathrm{Lip}})^{\dim_{\mathsf{A}}}}\, E(\mathcal{F})\right]^{\frac{1}{\dim_{\mathsf{A}}+1}},\ \Lambda \coloneqq \frac{1}{1-\beta}\left[\frac{m(\mathsf{A})(\dim_{\mathsf{A}}+1)!\zeta_0}{\alpha(2/Q_{\mathrm{Lip}})^{\dim_{\mathsf{A}}}}\right]^{\frac{1}{\dim_{\mathsf{A}}+1}}.$$

The constant error Δ is due to the algorithm's representation error $E(\mathcal{F})$, which is generally negligible.

Proof For any real-valued function $Q(x,a)$, recall the definition

$$\|Q\|_\nu^2 \coloneqq \sum_{x \in \mathsf{X}} \int_{\mathsf{A}} Q(x,a)^2\, m_{\mathsf{A}}(da)\, \nu(x).$$

Let Q_l be the random Q-function at the l^{th}-step of the algorithm. First, we find an upper bound to the following probability

$$P_0 \coloneqq \mathbb{P}\left(\|Q_{l+1} - H_\mu Q_l\|_\nu^2 > E(\mathcal{F})^2 + \varepsilon'\right),$$

for a given $\varepsilon' > 0$. To that end, we define

$$\hat{L}_N(f; Q) \coloneqq \frac{1}{N}\sum_{t=1}^{N} \frac{1}{m(\mathsf{A})\, \pi_b(a_t|x_t)} \left| f(x_t, a_t) - \left[c_t + \beta \min_{a' \in \mathsf{A}} Q(y_{t+1}, a')\right]\right|^2.$$

The normalization with $\pi_b(a_t|x_t)$ is used here to avoid assigning more weight to the actions that are preferred by the policy, and $m(\mathsf{A})$ is introduced for mathematical convenience.

One can show that (see (Antos et al., 2007b, Lemma 4.1))

$$\mathbb{E}\left[\hat{L}_N(f; Q)\right] = \|f - H_\mu Q\|_\nu^2 + L^*(Q) =: L(f; Q),$$

where $L^*(Q)$ is some quantity independent of f. Since we need a similar equation for the average-cost, let us prove it in detail so that we can refer to this proof in the future. Indeed, for each $t = 1, \ldots, N$, define

$$\hat{Q}_t \coloneqq c_t + \beta \min_{a' \in \mathsf{A}} Q(y_{t+1}, a').$$

Then,

$$\mathbb{E}\left[\hat{Q}_t \,\middle|\, x_t, a_t\right] = H_\mu Q(x_t, a_t).$$

Note that we can write

$$\begin{aligned}\mathbb{E}\left[\left(f(x_t,a_t) - \left[c_t + \beta \min_{a'\in\mathsf{A}} Q(y_{t+1},a')\right]\right)^2 \,\middle|\, x_t, a_t\right] &= \mathbb{E}\left[\left(f(x_t,a_t) - \hat{Q}_t\right)^2 \,\middle|\, x_t, a_t\right] \\ &= \mathbb{E}\left[\left(\hat{Q}_t - H_\mu Q(x_t,a_t)\right)^2 \,\middle|\, x_t, a_t\right] + \left(f(x_t,a_t) - H_\mu Q(x_t,a_t)\right)^2.\end{aligned}$$

Dividing each term by $m(\mathsf{A})\,\pi_b(a_t|x_t)$, taking the expectation of both sides with respect to a and x, and using the *law of iterated expectation* we get

$$\begin{aligned}&\mathbb{E}\left[\frac{\left(f(x_t,a_t) - [c_t + \beta \min_{a'\in\mathsf{A}} Q(y_{t+1},a')]\right)^2}{m(\mathsf{A})\,\pi_b(a_t|x_t)}\right] \\ &\quad= \mathbb{E}\left[\frac{\left(\hat{Q}_t - H_\mu Q(x_t,a_t)\right)^2}{m(\mathsf{A})\,\pi_b(a_t|x_t)}\right] + \sum_{x_t\in\mathsf{X}}\int_{\mathsf{A}} \frac{\left(f(x_t,a_t) - H_\mu Q(x_t,a_t)\right)^2}{m(\mathsf{A})\,\pi_b(a_t|x_t)}\,\pi_b(a_t|x_t)\,m(da_t)\,\nu(x_t) \\ &\quad= \mathbb{E}\left[\frac{\left(\hat{Q}_t - H_\mu Q(x_t,a_t)\right)^2}{m(\mathsf{A})\,\pi_b(a_t|x_t)}\right] + \sum_{x_t\in\mathsf{X}}\int_{\mathsf{A}} \left(f(x_t,a_t) - H_\mu Q(x_t,a_t)\right)^2\, m_{\mathsf{A}}(da_t)\,\nu(x_t) \\ &\quad=: L^*(Q) + \|f - H_\mu Q\|_\nu^2 =: L(f;Q).\end{aligned}$$

Since the samples are i.i.d., this establishes the fact.

Using above discussion, we can obtain the following bound

$$\begin{aligned}\|Q_{l+1} - H_\mu Q_l\|_\nu^2 - E(\mathcal{F})^2 &\le \|Q_{l+1} - H_\mu Q_l\|_\nu^2 - \inf_{f\in\mathcal{F}} \|f - H_\mu Q_l\|_\nu^2 \\ &= L(Q_{l+1};Q_l) - \inf_{f\in\mathcal{F}} L(f;Q_l) \\ &= L(Q_{l+1};Q_l) - \hat{L}_N(Q_{l+1};Q_l) + \hat{L}_N(Q_{l+1};Q_l) - \inf_{f\in\mathcal{F}} L(f;Q_l) \\ &= L(Q_{l+1};Q_l) - \hat{L}_N(Q_{l+1};Q_l) + \inf_{f\in\mathcal{F}} \hat{L}_N(f;Q_l) - \inf_{f\in\mathcal{F}} L(f;Q_l) \\ &\le 2\sup_{f\in\mathcal{F}} \left|L(f;Q_l) - \hat{L}_N(f;Q_l)\right| \\ &\le 2\sup_{f,Q\in\mathcal{F}} \left|L(f;Q) - \hat{L}_N(f;Q)\right|.\end{aligned}$$

This implies that

$$P_0 \le \mathbb{P}\left(\sup_{f,Q\in\mathcal{F}} \left|L(f;Q) - \hat{L}_N(f;Q)\right| > \frac{\varepsilon'}{2}\right). \tag{25}$$

For any $f, Q \in \mathcal{F}$, we define

$$l_{f,Q}(x,a,c,y) \coloneqq \frac{1}{m(\mathsf{A})\pi_b(a|x)}\left|f(x,a) - c - \beta \min_{a'\in\mathsf{A}} Q(y,a')\right|^2.$$

Let $\mathcal{L}_{\mathcal{F}} \coloneqq \{l_{f,Q} : f, Q \in \mathcal{F}\}$. Note that $\{z_t\}_{t=1}^N \coloneqq \{(x_t, a_t, c_t, y_{t+1})\}_{t=1}^N$ are i.i.d. and

$$\frac{1}{N}\sum_{t=1}^{N} l_{f,Q}(z_t) = \hat{L}_N(f;Q) \text{ and } \mathbb{E}[l_{f,Q}(z_1)] = L(f;Q).$$

Recall the constant $L_{\mathbf{m}} \coloneqq (1+\beta)Q_{\mathbf{m}} + c_{\mathbf{m}}$. One can prove that $0 \le l_{f,Q} \le \frac{L_{\mathbf{m}}^2}{m(\mathsf{A})\,\pi_0} =: C$. Then, by Pollard's Tail Inequality (Pollard, 1984, Theorem 24, p. 25), we have

$$\begin{aligned} P_0 &\le \mathbb{P}\left(\sup_{f,Q\in\mathcal{F}}\left|\frac{1}{N}\sum_{t=1}^{N} l_{f,Q}(z_t) - \mathbb{E}[l_{f,Q}(z_1)]\right| > \frac{\varepsilon'}{2}\right) \\ &\le 8\,\mathbb{E}\left[N_1\left(\frac{\varepsilon'}{16}, \{z_t\}_{t=1}^N, \mathcal{L}_{\mathcal{F}}\right)\right] e^{\frac{-N\,\varepsilon'^2}{512\,C^2}}. \end{aligned}$$

For any $l_{f,Q}$ and $l_{g,T}$, we also have (see (Antos et al., 2007b, pp. 18))

$$\begin{aligned} \frac{1}{N}\sum_{t=1}^{N} |l_{f,Q}(z_t) - l_{g,T}(z_t)| \le \frac{2L_{\mathbf{m}}}{m(\mathsf{A})\,\pi_0}\Bigg(&\frac{1}{N}\sum_{t=1}^{N} |f(x_t,a_t) - g(x_t,a_t)| \\ &+ \beta\frac{1}{N}\sum_{t=1}^{N}\left|\min_{b\in\mathsf{A}} Q(y_{t+1},b) - \min_{b\in\mathsf{A}} T(y_{t+1},b)\right|\Bigg). \end{aligned}$$

This implies that, for any $\epsilon > 0$, we have

$$\begin{aligned} N_1\left(\frac{2L_{\mathbf{m}}}{m(\mathsf{A})\,\pi_0}(1+\beta)\,\epsilon, \{z_t\}_{t=1}^N, \mathcal{L}_{\mathcal{F}}\right) &\le N_1\left(\epsilon, \{(x_t,a_t)\}_{t=1}^N, \mathcal{F}\right) N_1\left(\epsilon, \{y_{t+1}\}_{t=1}^N, \mathcal{F}_{\min}\right) \\ &\overset{(I)}{\le} e\,(V_{\mathcal{F}}+1)\left(\frac{2eQ_{\mathbf{m}}}{\epsilon}\right)^{V_{\mathcal{F}}} e\,(V_{\mathcal{F}_{\min}}+1)\left(\frac{2eQ_{\mathbf{m}}}{\epsilon}\right)^{V_{\mathcal{F}_{\min}}}, \end{aligned} \tag{26}$$

where (I) follows from Lemma 31. Therefore, we have the following bound on the probability P_0:

$$P_0 \le 8\left\{e^2\,(V_{\mathcal{F}}+1)\,(V_{\mathcal{F}_{\min}}+1)\left(\frac{64eQ_{\mathbf{m}}L_{\mathbf{m}}(1+\beta)}{m(\mathsf{A})\pi_0\varepsilon'}\right)^{V_{\mathcal{F}}+V_{\mathcal{F}_{\min}}}\right\} e^{\frac{-N\,\varepsilon'^2}{512\,C^2}}. \tag{27}$$

Recall the constants

$$\Upsilon = 8\,e^2\,(V_{\mathcal{F}}+1)\,(V_{\mathcal{F}_{\min}}+1)\left(\frac{64eQ_{\mathbf{m}}L_{\mathbf{m}}(1+\beta)}{m(\mathsf{A})\pi_0}\right)^{V_{\mathcal{F}}+V_{\mathcal{F}_{\min}}}, V \quad = V_{\mathcal{F}} + V_{\mathcal{F}_{\min}},\ \gamma = 512C^2.$$

Then, we can write (27) as follows

$$P_0 \coloneqq \mathbb{P}\left(\|Q_{l+1} - H_\mu Q_l\|_\nu^2 > E(\mathcal{F})^2 + \varepsilon'\right) \le \Upsilon\,\varepsilon'^{-V}\,e^{\frac{-N\varepsilon'^2}{\gamma}} =: \frac{\delta'}{L}. \tag{28}$$

Hence, for each $l = 0, \ldots, L-1$, with probability at most $\frac{\delta'}{L}$

$$\|Q_{l+1} - H_\mu Q_l\|_\nu^2 > \varepsilon' + E(\mathcal{F})^2.$$

This implies that with probability at most $\frac{\delta'}{L}$

$$\|Q_{l+1} - H_\mu Q_l\|_\nu > \sqrt{\varepsilon'} + E(\mathcal{F}).$$

Using this, we can conclude that with probability at least $1-\delta'$

$$\begin{aligned}
\|Q_L - H_1(\mu)\|_\infty &\leq \sum_{l=0}^{L-1} \beta^{L-(l+1)} \|Q_{l+1} - H_\mu Q_l\|_\infty + \|H_\mu^L Q_0 - H_1(\mu)\|_\infty \\
&\overset{(II)}{\leq} \sum_{l=0}^{L-1} \beta^{L-(l+1)} \left[\frac{m(\mathsf{A})(\dim_{\mathsf{A}}+1)!\zeta_0}{\alpha(2/Q_{\mathrm{Lip}})^{\dim_{\mathsf{A}}}} \|Q_{l+1} - H_\mu Q_l\|_\nu\right]^{\frac{1}{\dim_{\mathsf{A}}+1}} + \frac{\beta^L}{1-\beta} Q_{\mathbf{m}} \\
&\leq \sum_{l=0}^{L-1} \beta^{L-(l+1)} \left[\frac{m(\mathsf{A})(\dim_{\mathsf{A}}+1)!\zeta_0}{\alpha(2/Q_{\mathrm{Lip}})^{\dim_{\mathsf{A}}}} (\sqrt{\varepsilon'} + E(\mathcal{F}))\right]^{\frac{1}{\dim_{\mathsf{A}}+1}} + \frac{\beta^L}{1-\beta} Q_{\mathbf{m}} \\
&\leq \frac{1}{1-\beta}\left(\left[\frac{m(\mathsf{A})(\dim_{\mathsf{A}}+1)!\zeta_0}{\alpha(2/Q_{\mathrm{Lip}})^{\dim_{\mathsf{A}}}} E(\mathcal{F})\right]^{\frac{1}{\dim_{\mathsf{A}}+1}} + \left[\frac{m(\mathsf{A})(\dim_{\mathsf{A}}+1)!\zeta_0}{\alpha(2/Q_{\mathrm{Lip}})^{\dim_{\mathsf{A}}}}\right]^{\frac{1}{\dim_{\mathsf{A}}+1}} {\varepsilon'}^{\frac{1}{2(\dim_{\mathsf{A}}+1)}}\right) \\
&\qquad + \frac{\beta^L}{1-\beta} Q_{\mathbf{m}},
\end{aligned}$$

where (II) follows from Lemma 29. Then, with probability at least $1-\delta'$, we have

$$\|Q_L - H_1(\mu)\|_\infty \leq \Lambda {\varepsilon'}^{\frac{1}{2(\dim_{\mathsf{A}}+1)}} + \Delta + \frac{\beta^L}{1-\beta} Q_{\mathbf{m}}. \tag{29}$$

The result follows by picking $\delta = \delta' \coloneqq L\,\Upsilon\,{\varepsilon'}^{-V}\, e^{\frac{-N{\varepsilon'}^2}{\gamma}}$ in (28), choosing $\Lambda {\varepsilon'}^{\frac{1}{2(\dim_{\mathsf{A}}+1)}} = \varepsilon/2$, and $\frac{\beta^L}{1-\beta} Q_{\mathbf{m}} = \varepsilon/2$. ■

Remark 14 *We use the $\|\cdot\|_\nu$-norm on Q-functions until a certain stage in the proof of Theorem 13, and then we use Lemma 29 to go back to the $\|\cdot\|_\infty$-norm. Notice that Assumption 1-(c) on A is needed to accomplish this because the operator H_1 becomes a β-contraction only in terms of the $\|\cdot\|_\infty$-norm. However, without switching from $\|\cdot\|_\nu$-norm to $\|\cdot\|_\infty$-norm, a similar error analysis in terms of $\|\cdot\|_\nu$-norm can be formed by replacing Assumption 1-(c) with a concentrability assumption (see Munos and Szepesvári (2008); Agarwal et al. (2019)). To that end, let us define the state-action visitation probability of any policy π as*

$$d^\pi(x, da) \coloneqq (1-\beta)\sum_{t=0}^{\infty} \mathbb{P}^\pi(x(t) = x, a(t) \in da).$$

The concentrability assumption states that the state-action visitation probability d^π is absolutely continuous with respect to $\nu(x) \otimes m_{\mathsf{A}}(da)$ for any $\pi \in \Pi$, and the corresponding densities are uniformly bounded, i.e.,

$$\sup_{\pi\in\Pi} \left\| \frac{d^\pi}{\nu \otimes m_{\mathsf{A}}} \right\|_\infty \leq C,$$

for some C. Under this assumption, the final part of Theorem 13 can be handled with the $\|\cdot\|_\nu$-norm instead of the $\|\cdot\|_\infty$-norm using performance difference lemma (Agarwal et al., 2019, Theorem 15.4). However, it is not possible to establish the overall error analysis of the learning algorithm using the $\|\cdot\|_\nu$-norm on Q-functions under the same set of assumptions on the system components without strengthening Assumption 1-(d).

We now give the description of the random operator $\hat{H}_2$, and then, do the error analysis. In this algorithm, the goal is to replace the operator H_2, which gives the next state-measure, with $\hat{H}_2$. We achieve this by simulating the transition probability $p(\cdot|x,a,\mu)$ for certain state-measure μ and policy π. This is possible since $|\mathsf{X}|$ is finite.

Algorithm 2 Algorithm $\hat{H}_2$

Inputs (μ, Q), Data size M, Number of iterations $|\mathsf{X}|$
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 $\sum_{x\in\mathsf{X}} p_M(\cdot|x, f_Q(x), \mu)\,\mu(x)$

This is the error analysis of the random operator $\hat{H}_2$.

Theorem 15 *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

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

Proof By Hoeffding Inequality (Hajek and Raginsky, 2019, Theorem 2.1), for any $x, y \in \mathsf{X}$, we have

$$\mathbb{P}\left(|p_M(y|x, f_Q(x), \mu) - p(y|x, f_Q(x), \mu)| > \frac{\varepsilon}{|\mathsf{X}|}\right) \leq 2e^{\frac{-M\varepsilon^2}{|\mathsf{X}|^2}}.$$

Hence, we have

$$\begin{aligned}
&\mathbb{P}\left(\left\|\hat{H}_2[M](\mu,Q) - H_2(\mu,Q)\right\|_1 \leq \varepsilon\right) \\
&\geq \mathbb{P}\left(\sum_{x,y\in\mathsf{X}} |p_M(y|x, f_Q(x), \mu) - p(y|x, f_Q(x), \mu)|\;\mu(x) \leq \varepsilon\right)
\end{aligned}$$

$$\geq 1 - \mathbb{P}\left(\exists x, y \in \mathsf{X} \text{ s.t. } |p_M(y|x, f_Q(x), \mu) - p(y|x, f_Q(x), \mu)| > \frac{\varepsilon}{|\mathsf{X}|},\right)$$
$$\geq 1 - 2\,|\mathsf{X}|^2\, e^{\frac{-M\varepsilon^2}{|\mathsf{X}|^2}}.$$

The result follows by picking $\delta = 2\,|\mathsf{X}|^2\, e^{\frac{-M\varepsilon^2}{|\mathsf{X}|^2}}$. ■

The overall description of the learning algorithm is given below. In this algorithm, to achieve an approximate mean-field equilibrium policy, we successively apply the random operator $\hat{H}$ which replaces the MFE operator H.

Algorithm 3 Learning Algorithm

Input μ_0, Number of iterations K, Parameters of $\hat{H}_1$ and $\hat{H}_2$ $\left(\{[N_k, L_k]\}_{k=0}^{K-1}, \{M_k\}_{k=0}^{K-1}\right)$

Start with μ_0
for $k = 0, \ldots, K-1$ **do**
$\mu_{k+1} = \hat{H}\left([N_k, L_k], M_k\right)(\mu_k) \coloneqq \hat{H}_2[M_k]\left(\mu_k, \hat{H}_1[N_k, L_k](\mu_k)\right)$
end for
return μ_K

The current state-measure μ_k is the input for each iteration $k = 0, \ldots, K-1$. In addition, for the random operator $\hat{H}_1$, we choose integers N_k and L_k as the data size and the number of iterations, respectively, and for the random operator $\hat{H}_2$, we choose an integer M_k as the data size. We first compute an approximate Q-function for μ_k by applying $\hat{H}_1[N_k, L_k](\mu_k)$, and then we compute an approximate next state-measure by applying $\hat{H}_2[M_k](\mu_k, \hat{H}_1[N_k, L_k](\mu_k))$. Since an approximate Q-function is used instead of the exact Q-function in the second stage of the iteration, there will be an error due to $\hat{H}_1$ in addition to the error resulting from $\hat{H}_2$.

The error analyses of the algorithms $\hat{H}_1$ and $\hat{H}_2$ have been completed in Theorem 13 and Theorem 15, respectively. The error analysis for the learning algorithm for the random operator $\hat{H}$, which is a combination of $\hat{H}_1$ and $\hat{H}_2$, is given below. We state the key result of this section as a corollary after the proof of the following theorem.

Theorem 16 *Fix any $(\varepsilon, \delta) \in (0,1)^2$. Define*

$$\varepsilon_1 \coloneqq \frac{\rho\,(1-K_H)^2\,\varepsilon^2}{64(K_1)^2},\ \varepsilon_2 \coloneqq \frac{(1-K_H)\,\varepsilon}{4}.$$

Let K, L be such that

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

Then, pick N, M such that

$$N \geq m_1\left(\varepsilon_1, \frac{\delta}{2K}, L\right),\ M \geq m_2\left(\varepsilon_2, \frac{\delta}{2K}\right). \tag{30}$$

Let μ_K be the output of the learning algorithm with parameters

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

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

$$\|\mu_K - \mu_*\|_1 \leq \frac{2K_1\sqrt{\Delta}}{\sqrt{\rho}(1-K_H)} + \varepsilon,$$

where μ_* is the state-measure in mean-field equilibrium given by the MFE operator H.

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

$$\begin{aligned}
\|H_2(\mu, Q) - H_2(\mu, \hat{Q})\|_1 &= \sum_{y\in\mathsf{X}} \left| \sum_{x\in\mathsf{X}} p(y|x, f_Q(x), \mu)\, \mu(x) - \sum_{x\in\mathsf{X}} p(y|x, f_{\hat{Q}}(x), \mu)\, \mu(x) \right| \\
&\leq \sum_{x\in\mathsf{X}} \|p(\cdot|x, f_Q(x), \mu) - p(\cdot|x, f_{\hat{Q}}(x), \mu)\|_1\, \mu(x) \\
&\leq \sum_{x\in\mathsf{X}} K_1 \|f_Q(x) - f_{\hat{Q}}(x)\|\, \mu(x). \qquad (31)
\end{aligned}$$

For all $x \in \mathsf{X}$, note that the mapping $f_Q(x)$ is the minimizer of $Q(x, \cdot)$ and the mapping $f_{\hat{Q}}(x)$ is the unique minimizer of $\hat{Q}(x, \cdot)$ by strong convexity. Let us set $a = f_{\hat{Q}}(x)$ and $r = f_Q(x) - f_{\hat{Q}}(x)$. As a is the unique minimizer of a strongly convex function $\hat{Q}(x, \cdot)$, by first-order optimality condition, we have

$$\nabla \hat{Q}\,(x, a) \cdot r \geq 0.$$

Hence, by strong convexity

$$\begin{aligned}
\hat{Q}(x, a+r) - \hat{Q}(x, a) &\geq \nabla \hat{Q}(x, a) \cdot r + \frac{\rho}{2}\|r\|^2 \\
&\geq \frac{\rho}{2}\|r\|^2 \qquad (32)
\end{aligned}$$

For all $x \in \mathsf{X}$, this leads to

$$\begin{aligned}
\|f_Q(x) - f_{\hat{Q}}(x)\|^2 &\leq \frac{2}{\rho}\left(\hat{Q}(x, f_Q(x)) - \hat{Q}(x, f_{\hat{Q}}(x))\right) \\
&= \frac{2}{\rho}\left(\hat{Q}(x, f_Q(x)) - Q(x, f_Q(x)) + Q(x, f_Q(x)) - \hat{Q}(x, f_{\hat{Q}}(x))\right) \\
&= \frac{2}{\rho}\left(\hat{Q}(x, f_Q(x)) - Q(x, f_Q(x)) + \min_{a\in\mathsf{A}} Q(x, a) - \min_{a\in\mathsf{A}} \hat{Q}(x, a)\right) \\
&\leq \frac{4}{\rho}\,\|Q - \hat{Q}\|_\infty. \qquad (33)
\end{aligned}$$

Hence, combining (31) and (33) yields

$$\|H_2(\mu, Q) - H_2(\mu, \hat{Q})\|_1 \leq \frac{2K_1}{\sqrt{\rho}}\,\sqrt{\|Q - \hat{Q}\|_\infty}. \qquad (34)$$

Using (34) and the fact that $H_1(\mu_k) \in \mathcal{C}$ and $\hat{H}_1[N,L](\mu_k) \in \mathcal{F}$, for any $k = 0, \dots, K-1$, we have

$$\begin{aligned}\|H(\mu_k) - \hat{H}([N,L],M)(\mu_k)\|_1 &\leq \|H_2(\mu_k, H_1(\mu_k)) - H_2(\mu_k, \hat{H}_1[N,L](\mu_k))\|_1 \\ &\quad + \|H_2(\mu_k, \hat{H}_1[N,L](\mu_k)) - \hat{H}_2[M](\mu_k, \hat{H}_1[N,L](\mu_k))\|_1 \\ &\leq \frac{2K_1}{\sqrt{\rho}} \sqrt{\|H_1(\mu_k) - \hat{H}_1[N,L](\mu_k)\|_\infty} \\ &\quad + \|H_2(\mu_k, \hat{H}_1[N,L](\mu_k)) - \hat{H}_2[M](\mu_k, \hat{H}_1[N,L](\mu_k))\|_1.\end{aligned}$$

The last term is upper bounded by

$$\frac{2K_1\sqrt{\varepsilon_1 + \Delta}}{\sqrt{\rho}} + \varepsilon_2$$

with probability at least $1 - \frac{\delta}{K}$ by Theorem 13 and Theorem 15. Therefore, with probability at least $1 - \delta$

$$\begin{aligned}\|\mu_K - \mu_*\|_1 &\leq \sum_{k=0}^{K-1} K_H^{K-(k+1)} \|\hat{H}([N,L],M)(\mu_k) - H(\mu_k)\|_1 + \|H^K(\mu_0) - \mu_*\|_1 \\ &\leq \sum_{k=0}^{K-1} K_H^{K-(k+1)} \left(\frac{2K_1\sqrt{\varepsilon_1 + \Delta}}{\sqrt{\rho}} + \varepsilon_2 \right) + \frac{(K_H)^K}{1 - K_H} \\ &\leq \frac{2K_1\sqrt{\Delta}}{\sqrt{\rho}(1 - K_H)} + \varepsilon.\end{aligned}$$

This completes the proof. ■

Now, we state the main result of this section. It basically states that, by using the learning algorithm, one can learn an approximate mean-field equilibrium policy. By Theorem 11, this gives an approximate Nash-equilibrium for the finite-agent game.

Corollary 17 *Fix any $(\varepsilon, \delta) \in (0,1)^2$. Suppose that K, L, N, M satisfy the conditions in Theorem 16. Let μ_K be the output of the learning algorithm with parameters*

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

Define $\pi_K(x) \coloneqq \arg\min_{a \in \mathsf{A}} Q_K(x,a)$, where $Q_K \coloneqq \hat{H}_1([N,L])(\mu_K)$. Then, with probability at least $1 - \delta(1 + \frac{1}{2K})$, the policy π_K is a $\kappa(\varepsilon, \Delta)$-mean-field equilibrium policy, where

$$\kappa(\varepsilon, \Delta) = \sqrt{\frac{4}{\rho} \left(\frac{\rho^2 (1 - K_H)^2 \varepsilon^2}{64 (K_1)^2} + \Delta + K_{H_1} \left(\frac{2K_1\sqrt{\Delta}}{\sqrt{\rho}(1 - K_H)} + \varepsilon \right) \right)}.$$

Therefore, by Theorem 11, an N-tuple of policies $\pi^{(N)} = \{\pi_K, \pi_K, \dots, \pi_K\}$ is an $\tau\kappa(\varepsilon, \Delta) + \sigma$-Nash equilibrium for the game with $N \geq N(\sigma)$ agents.

Proof By Theorem 13 and Theorem 16, with probability at least $1-\delta(1+\frac{1}{2K})$, we have

$$\begin{aligned}\|Q_K - H_1(\mu_*)\|_\infty &\le \|Q_K - H_1(\mu_K)\|_\infty + \|H_1(\mu_K) - H_1(\mu_*)\|_\infty \\ &\le \varepsilon_1 + \Delta + K_{H_1}\|\mu_K - \mu_*\|_1 \\ &\le \varepsilon_1 + \Delta + K_{H_1}\left(\frac{2K_1\sqrt{\Delta}}{\sqrt{\rho}(1-K_H)} + \varepsilon\right) \\ &= \frac{\rho\,(1-K_H)^2\,\varepsilon^2}{64(K_1)^2} + \Delta + K_{H_1}\left(\frac{2K_1\sqrt{\Delta}}{\sqrt{\rho}(1-K_H)} + \varepsilon\right).\end{aligned}$$

Let $\pi_K(x) \coloneqq \arg\min_{a\in\mathsf{A}} Q_K(x,a)$. Using the same analysis that leads to (33), we can obtain the following bound since $Q_K \in \mathcal{F}$ and $H_1(\mu_*) \in \mathcal{C}$:

$$\sup_{x\in\mathsf{X}} \|\pi_K(x) - \pi_*(x)\|^2 \le \frac{4}{\rho}\,\|Q_K - H_1(\mu_*)\|_\infty.$$

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

$$\kappa(\varepsilon,\Delta) = \sqrt{\frac{4}{\rho}\left(\frac{\rho\,(1-K_H)^2\,\varepsilon^2}{64(K_1)^2} + \Delta + K_{H_1}\left(\frac{2K_1\sqrt{\Delta}}{\sqrt{\rho}(1-K_H)} + \varepsilon\right)\right)}.$$

This completes the proof. ■

Remark 18 *Note that, in Corollary 17, there is a constant Δ, which depends on the representation error $E(\mathcal{F})$. In general, $E(\mathcal{F})$ is very small since any Q function in $\mathcal{C}$ can be approximated quite well by functions in $\mathcal{F}$. Therefore, Δ is negligible. In this case, we have the following error bound:*

$$\kappa(\varepsilon,0) = \sqrt{\frac{4}{\rho}\left(\frac{\rho\,(1-K_H)^2\,\varepsilon^2}{64(K_1)^2} + K_{H_1}\varepsilon\right)}.$$

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

6. Mean-field Equilibrium Operator for Average-cost

In this section, we introduce the corresponding MFE operator for average-cost mean-field games. For the purpose of keeping the notation similar to the discounted-cost case while making the distinctions more apparent, we use the 'av' superscript to denote the related quantities in the average-cost setting. For instance, to denote the average-cost of any policy π with initial state x under state-measure μ, we use $J_\mu^{\mathrm{av}}(\pi,x)$ instead of $J_\mu(\pi,x)$. Now, let us state the extra conditions imposed for the average-cost in addition to Assumption 1.

Assumption 3

(a) There exists a sub-probability measure λ on X such that

$$p(\,\cdot\,|x,a,\mu) \geq \lambda(\,\cdot\,)$$

for all x, a, μ.

(b) Let $\beta^{\mathrm{av}} := 1-\lambda(\mathsf{X})$ and $Q^{\mathrm{av}}_{\mathrm{Lip}} := \dfrac{L_1}{1-K_1/2} > 0$. We assume that

$$\frac{3K_1}{2}\left(1+\frac{K_F}{\rho}\right)+\frac{K_1K_FQ^{\mathrm{av}}_{\mathrm{Lip}}}{\rho(1-\beta^{\mathrm{av}})}<1.$$

Note that Assumption 3-(a) is the so-called 'minorization' condition. Minorization condition was used in the literature for studying the geometric ergodicity of Markov chains (see (Hernández-Lerma, 1989, Section 3.3)). The minorization condition is true when the transition probability satisfies conditions R0, R1(a) and R1(b) in Hernández-Lerma et al. (1991) (see also (Hernández-Lerma et al., 1991, Remark 3.3) and references therein for further conditions). In general, this condition is restrictive for unbounded state spaces, but it is quite general for compact or finite state spaces. Indeed, the minorization condition was used to study average-cost mean-field games with a compact state space in (Wiecek, 2019, Assumption A.3). Note that Assumption 3-(b) is used to ensure that MFE operator is contraction, which is crucial to establish the error analysis of the learning algorithm, and so, cannot be relaxed.

Recall that for the average-cost, given any state-measure μ, the value function J^{av}_{μ} of policy π with initial state x is given by

$$J^{\mathrm{av}}_{\mu}(\pi,x) := \limsup_{T\to\infty}\frac{1}{T}E^{\pi}\left[\sum_{t=0}^{T-1}c(x(t),a(t),\mu)\,\middle|\,x(0)=x\right].$$

Then, the optimal value function is defined as

$$J^{\mathrm{av},*}_{\mu}(x) := \inf_{\pi\in\Pi}J^{\mathrm{av}}_{\mu}(\pi,x).$$

Under Assumption 1 and Assumption 3, it can be proved that

$$J^{\mathrm{av},*}_{\mu}(x) = J^{\mathrm{av},*}_{\mu}(y) =: J^{\mathrm{av},*}_{\mu}$$

for all $x, y \in \mathsf{X}$, for some constant $J^{\mathrm{av},*}_{\mu}$; that is, the optimal value function does not depend on the initial state. Furthermore, let $h^*_{\mu}(x)$ be the unique fixed point of the β-contraction operator T^{av}_{μ} with respect to $\|\cdot\|_{\infty}$-norm:

$$h^*_{\mu}(x) = \min_{a\in\mathsf{A}}\left[c(x,a,\mu)+\sum_{y\in\mathsf{X}}h^*_{\mu}(y)q(y|x,a,\mu)\right] =: T^{\mathrm{av}}_{\mu}h^*_{\mu}(x),$$

where

$$q(\,\cdot\,|x,a,\mu) := p(\,\cdot\,|x,a,\mu)-\lambda(\,\cdot\,).$$

Then, the pair $\left(h^*_\mu, \sum_{y\in\mathsf{X}} h^*_\mu(y)\,\lambda(y)\right)$ satisfies the average-cost optimality equation (ACOE):

$$h^*_\mu(x) + \sum_{y\in\mathsf{X}} h^*_\mu(y)\,\lambda(y) = \min_{a\in\mathsf{A}}\left[c(x,a,\mu) + \sum_{y\in\mathsf{X}} h^*_\mu(y)\,p(y|x,a,\mu)\right].$$

Therefore, $J^{\mathrm{av},*}_\mu = \sum_{y\in\mathsf{X}} h^*_\mu(y)\,\lambda(y)$. Additionally, if $f^*:\mathsf{X}\to\mathsf{A}$ attains the minimum in the ACOE, that is,

$$\min_{a\in\mathsf{A}}\left[c(x,a,\mu) + \sum_{y\in\mathsf{X}} h^*_\mu(y)\,p(y|x,a,\mu)\right] = c(x,f^*(x),\mu) + \sum_{y\in\mathsf{X}} h^*_\mu(y)\,p(y|x,f^*(x),\mu) \quad (35)$$

for all $x\in\mathsf{X}$, then the policy $\pi^*(a|x) = \delta_{f^*(x)}(a) \in \Pi_d$ is optimal for any initial distribution. We refer the reader to (Hernández-Lerma, 1989, Chapter 3) for basics of average-cost Markov decision processes, where these classical results can be found.

We can also obtain a similar characterization by using a Q-function instead of h^*_μ. Indeed, we define the Q-function as

$$Q^{\mathrm{av},*}_\mu(x,a) = c(x,a,\mu) + \sum_{y\in\mathsf{X}} h^*_\mu(y)\,q(y|x,a,\mu).$$

Note that $Q^{\mathrm{av},*}_{\mu,\min}(x) \coloneqq \min_{a\in\mathsf{A}} Q^{\mathrm{av},*}_\mu(x,a) = h^*_\mu(x)$ for all $x\in\mathsf{X}$, and so, we have

$$Q^{\mathrm{av},*}_\mu(x,a) = c(x,a,\mu) + \sum_{y\in\mathsf{X}} Q^{\mathrm{av},*}_{\mu,\min}(y)\,q(y|x,a,\mu) =: H^{\mathrm{av}}_\mu Q^{\mathrm{av},*}_\mu(x,a),$$

where H^{av}_μ is the corresponding operator on Q-functions. Hence, the policy $\pi^*(a|x) = \delta_{f^*(x)}(a) \in \Pi_d$ is optimal for μ and for any initial distribution, if $Q^{\mathrm{av},*}_\mu(x,f^*(x)) = Q^{\mathrm{av},*}_{\mu,\min}(x)$ for all $x\in\mathsf{X}$. One can prove that H^{av}_μ is a $\|\cdot\|_\infty$-contraction with modulus β^{av}, and so, the unique fixed point of H^{av}_μ is $Q^{\mathrm{av},*}_\mu$. Indeed, let Q and $\hat{Q}$ be two different Q-functions. Then, we have

$$\begin{aligned}
\|H^{\mathrm{av}}_\mu Q - H^{\mathrm{av}}_\mu \hat{Q}\|_\infty &\le \sup_{(x,a)\in\mathsf{X}\times\mathsf{A}} \sum_{y\in\mathsf{X}} |Q_{\min}(y) - \hat{Q}_{\min}(y)|\,q(y|x,a,\mu) \\
&\le \|Q_{\min} - \hat{Q}_{\min}\|_\infty \sup_{(x,a)\in\mathsf{X}\times\mathsf{A}} q(\mathsf{X}|x,a,\mu) \\
&= \beta^{\mathrm{av}}\,\|Q_{\min} - \hat{Q}_{\min}\|_\infty.
\end{aligned}$$

Hence, using the Banach fixed point theorem, we can develop a Q-iteration algorithm to compute $Q^{\mathrm{av},*}_\mu$, the minimum of which gives the optimal policy. The benefit of this algorithm, as in the discounted case, is that it can be adapted to a model-free setting via Q-learning.

Using (11), we now prove the following result.

Lemma 19 *For any μ, $Q^{\mathrm{av},*}_{\mu,\min}$ is $Q^{\mathrm{av}}_{\mathrm{Lip}}$-Lipschitz continuous; that is,*

$$|Q^{\mathrm{av},*}_{\mu,\min}(x) - Q^{\mathrm{av},*}_{\mu,\min}(y)| \le Q^{\mathrm{av}}_{\mathrm{Lip}}\,d_{\mathsf{X}}(x,y).$$

Proof The proof is exactly the same with the proof of Lemma 4. The only difference is the absence of the discount factor β. ■

Before we define MFE operator, let us describe the set of possible Q-functions. This set $\mathcal{C}^{\text{av}}$ is the set of all Q-functions $Q : \mathsf{X} \times \mathsf{A} \to \mathbb{R}$ such that $\|Q\|_\infty \leq Q^{\text{av}}_{\mathbf{m}} := c_{\mathbf{m}}/(1-\beta^{\text{av}})$, $Q(x,\cdot)$ is $Q^{\text{av}}_{\text{Lip}}$-Lipschitz and ρ-strongly convex for all x, and the gradient $\nabla Q(x,a)$ of Q with respect to a satisfies the bound

$$\sup_{a\in\mathsf{A}} \|\nabla Q(x,a) - \nabla Q(\hat{x},a)\| \leq K_F,\ \forall x,\hat{x}.$$

Now, we can define the MFE operator H^{av}. The operator H^{av} is very similar to H; that is, it is a composition of two operators, where the first operator $H^{\text{av}}_1 : \mathcal{P}(\mathsf{X}) \to \mathcal{C}^{\text{av}}$ is defined as $H^{\text{av}}_1(\mu) = Q^{\text{av},*}_\mu$ (the unique fixed point of the operator H^{av}_μ). The second operator $H^{\text{av}}_2 : \mathcal{P}(\mathsf{X}) \times \mathcal{C}^{\text{av}} \to \mathcal{P}(\mathsf{X})$ is defined as

$$H^{\text{av}}_2(\mu, Q)(\cdot) := \sum_{x\in\mathsf{X}} p(\cdot|x, f_Q(x), \mu)\, \mu(x),$$

where $f_Q(x) := \arg\min_{a\in\mathsf{A}} Q(x,a)$ is the unique minimizer by ρ-strong convexity of Q, for any $Q \in \mathcal{C}^{\text{av}}$. Note that we indeed have $H^{\text{av}}_2 = H_2$, where H_2 is the operator that computes the new state-measure in discounted-cost. However, to be consistent with the notation used in this section, we keep H^{av}_2 as it is. Using these operators, let us define the MFE operator as a composition:

$$H^{\text{av}} : \mathcal{P}(\mathsf{X}) \ni \mu \mapsto H^{\text{av}}_2\left(\mu, H^{\text{av}}_1(\mu)\right) \in \mathcal{P}(\mathsf{X}).$$

Our goal is to establish that H^{av} is contraction. Using (11) and Lemma 19, we can first prove that H_1 is Lipschitz continuous.

Lemma 20 *The mapping H^{av}_1 is $K_{H^{\text{av}}_1}$-Lipschitz, where $K_{H^{\text{av}}_1} := \dfrac{Q^{\text{av}}_{\text{Lip}}}{1-\beta^{\text{av}}}$.*

Proof The proof can be done as in the proof of Lemma 6 by making appropriate modifications. Note that since $H^{\text{av}}_1(\mu) := Q^{\text{av},*}_\mu$ is the fixed point of the contraction operator H^{av}_μ, where H^{av}_μ is given by

$$H^{\text{av}}_\mu Q(x,a) = c(x,a,\mu) + \sum_{y\in\mathsf{X}} Q_{\min}(y)\, q(y|x,a,\mu),$$

by Assumption 1-(a),(b),(d), H^{av}_μ maps any continuous $Q : \mathsf{X}\times\mathsf{A} \to \mathbb{R}$ into $\mathcal{C}^{\text{av}}$. Hence, the fixed point $Q^{\text{av},*}_\mu$ of H^{av}_μ must be in $\mathcal{C}^{\text{av}}$ (see the proof of Lemma 6). Therefore, H^{av}_μ is well-defined.

Let us now prove that H^{av}_1 is $K_{H^{\text{av}}_1}$-Lipschitz. For any $\mu, \hat{\mu} \in \mathcal{P}(\mathsf{X})$, we have

$$\begin{aligned}
&\|H^{\text{av}}_1(\mu) - H^{\text{av}}_1(\hat{\mu})\|_\infty \\
&\qquad = \sup_{x,a} \left| c(x,a,\mu) + \sum_y Q^{\text{av},*}_{\mu,\min}(y) q(y|x,a,\mu) - c(x,a,\hat{\mu}) - \sum_y Q^{\text{av},*}_{\hat{\mu},\min}(y) q(y|x,a,\hat{\mu}) \right|
\end{aligned}$$

$$
\begin{aligned}
&\leq L_1 \left\|\mu - \hat{\mu}\right\|_1 \\
&+ \left| \sum_y Q^{\mathrm{av},*}_{\mu,\min}(y) q(y|x,a,\mu) - \sum_y Q^{\mathrm{av},*}_{\mu,\min}(y) q(y|x,a,\hat{\mu}) \right| \\
&+ \left| \sum_y Q^{\mathrm{av},*}_{\mu,\min}(y) q(y|x,a,\hat{\mu}) - \sum_y Q^{\mathrm{av},*}_{\hat{\mu},\min}(y) q(y|x,a,\hat{\mu}) \right| \\
&\leq L_1 \left\|\mu - \hat{\mu}\right\|_1 + Q^{\mathrm{av}}_{\mathrm{Lip}} K_1/2 \left\|\mu - \hat{\mu}\right\|_1 + \beta^{\mathrm{av}} \left\|Q^{\mathrm{av},*}_{\mu} - Q^{\mathrm{av},*}_{\hat{\mu}}\right\|_\infty,
\end{aligned}
$$

where the last inequality follows from (11), Lemma 19, and the fact $q(\mathsf{X}|x,a,\mu) = \beta^{\mathrm{av}}$ for all x, a, μ. ■

Now, using Lemma 20, we can prove that H^{av} is contraction.

Proposition 21 *The mapping H^{av} is a contraction with contraction modulus $K_{H^{\mathrm{av}}}$, where*

$$
K_{H^{\mathrm{av}}} \coloneqq \frac{3K_1}{2}\left(1 + \frac{K_F}{\rho}\right) + \frac{K_1 K_F K_{H_1^{\mathrm{av}}}}{\rho}.
$$

Proof The proof is exactly the same with the proof of Proposition 7. The only difference is the following: we should replace K_{H_1} in (16) with $K_{H_1^{\mathrm{av}}}$. ■

Now, we know that H^{av} is a contraction mapping under Assumption 1 and Assumption 3. Therefore, by Banach fixed point theorem, H^{av} has a unique fixed point μ^{av}_*. Let

$$
Q^{\mathrm{av},*}_{\mu^{\mathrm{av}}_*} = H_1^{\mathrm{av}}(\mu^{\mathrm{av}}_*) \quad \text{and} \quad \pi^{\mathrm{av}}_*(\,\cdot\,|x) = \delta_{f_{Q^{\mathrm{av},*}_{\mu^{\mathrm{av}}_*}}(x)}(\,\cdot\,).
$$

Then, the pair $(\pi^{\mathrm{av}}_*, \mu^{\mathrm{av}}_*)$ is a mean-field equilibrium since $(\mu^{\mathrm{av}}_*, Q^{\mathrm{av},*}_{\mu^{\mathrm{av}}_*})$ satisfy the following equations

$$
\mu^{\mathrm{av}}_*(\cdot) = \sum_{x \in \mathsf{X}} p(\cdot|x,a,\mu^{\mathrm{av}}_*)\, \pi^{\mathrm{av}}_*(a|x)\, \mu^{\mathrm{av}}_*(x), \tag{36}
$$

$$
Q^*_{\mu^{\mathrm{av}}_*}(x,a) = c(x,a,\mu^{\mathrm{av}}_*) + \sum_{y \in \mathsf{X}} Q^{\mathrm{av},*}_{\mu^{\mathrm{av}}_*,\min}(y)\, q(y|x,a,\mu^{\mathrm{av}}_*). \tag{37}
$$

Here, (37) implies that $\pi^{\mathrm{av}}_* \in \Psi(\mu^{\mathrm{av}}_*)$ since

$$
f_{Q^{\mathrm{av},*}_{\mu^{\mathrm{av}}_*}}(x) \coloneqq \arg\min Q^{\mathrm{av},*}_{\mu^{\mathrm{av}}_*}(x,a)
$$

for every $x \in \mathsf{X}$, and (36) implies $\mu^{\mathrm{av}}_* \in \Lambda(\pi^{\mathrm{av}}_*)$. Hence, $(\pi^{\mathrm{av}}_*, \mu^{\mathrm{av}}_*)$ is a mean-field equilibrium. Therefore, since H^{av} is a contraction, we can compute this mean-field equilibrium by applying H^{av} recursively starting from arbitrary state-measure.

Note that if the transition probability p, the one-stage cost function c, and the minorizing sub-probability measure λ are not available to the decision maker, we need to replace H^{av} 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^{av} is crucial, similar to the discounted-case.

7. Finite-Agent Game for Average-cost

The finite-agent game model for average-cost is exactly the same with the model introduced in Section 4 for the discounted-cost case. The only difference is the cost function. Here, under an N-tuple of policies $\boldsymbol{\pi}^{(N)} \coloneqq (\pi^1, \dots, \pi^N)$, for Agent i, the average-cost is given by

$$J_i^{\mathrm{av},(N)}(\boldsymbol{\pi}^{(N)}) = \limsup_{T\to\infty} \frac{1}{T} E^{\boldsymbol{\pi}^{(N)}}\left[\sum_{t=0}^{T-1} c(x_i^N(t), a_i^N(t), e_t^{(N)})\right].$$

Using this, we define Nash equilibrium and δ-Nash equilibrium similarly.

Definition 22 *An N-tuple of policies $\boldsymbol{\pi}^{(N*)} = (\pi^{1*}, \dots, \pi^{N*})$ constitutes a* Nash equilibrium *if*

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

for each $i = 1, \dots, N$. An N-tuple of policies $\boldsymbol{\pi}^{(N*)} = (\pi^{1*}, \dots, \pi^{N*})$ constitutes an δ-Nash equilibrium *if*

$$J_i^{\mathrm{av},(N)}(\boldsymbol{\pi}^{(N*)}) \leq \inf_{\pi^i \in \Pi_i} J_i^{\mathrm{av},(N)}(\boldsymbol{\pi}_{-i}^{(N*)}, \pi^i) + \delta$$

for each $i = 1, \dots, N$.

As in the discounted-cost case, if the number of agents is large enough in the finite-agent setting, one can obtain δ-Nash equilibrium by considering the infinite-population limit $N \to \infty$ of the game (i.e., mean-field game). Then, it is possible to prove that if each agent in the finite-agent N game problem adopts the policy in mean-field equilibrium, the resulting N-tuple of policies will be an approximate Nash equilibrium for all sufficiently large N. This was indeed proved in Wiecek (2019); Saldi (2020). In the below theorem, we prove that if each agent in the finite-agent game model adopts the ε-mean-field equilibrium policy (instead of exact mean-field equilibrium policy), the resulting policy will still be an approximate Nash equilibrium for all sufficiently large N-agent game models.

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

$$C_1^{\mathrm{av}} \coloneqq \left(\frac{3\,K_1}{2} + \frac{K_1\,K_F}{2\rho}\right),\ C_2^{\mathrm{av}} \coloneqq \frac{2c_{\mathbf{m}}(K_1)^2}{(1-C_1^{\mathrm{av}})\lambda(\mathsf{X})},\ C_3^{\mathrm{av}} \coloneqq \frac{2c_{\mathbf{m}}}{\lambda(\mathsf{X})}.$$

Note that by Assumption 3, the constant C_1^{av} is strictly less than 1.

Theorem 23 *Let π_ε be an ε-mean-field equilibrium policy for the mean-field equilibrium $(\pi_*, \mu_*) \in \Pi_d \times \mathcal{P}(\mathsf{X})$ given by the unique fixed point of the MFE operator H^{av}. Let $\eta_0 \in \Lambda(\pi_\varepsilon)$. 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 $\boldsymbol{\pi}^{(N)} = \{\pi_\varepsilon, \pi_\varepsilon, \dots, \pi_\varepsilon\}$ is a $(\delta + \tau^{\mathrm{av}}\varepsilon)$-Nash equilibrium for the game with N agents, where $\tau^{\mathrm{av}} \coloneqq 2C_2^{\mathrm{av}} + C_3^{\mathrm{av}}$.*

Proof By an abuse of notation, we denote the deterministic mappings from X to A that induce policies π_* and π_ε as π_* and π_ε as well, respectively. As in the proof of Theorem 11, one can prove that

$$\|\pi_*(x) - \pi_*(y)\| \le \frac{K_F}{\rho}\, d_{\mathsf{X}}(x,y) \quad \text{and} \quad \|\mu_\varepsilon - \mu_*\|_1 \le \frac{K_1\,\varepsilon}{1 - C_1^{\mathrm{av}}},$$

where $\mu_\varepsilon \in \Lambda(\pi_\varepsilon)$ and $C_1^{\mathrm{av}} := \left(\frac{3\,K_1}{2} + \frac{K_1\,K_F}{2\rho}\right)$. Note that by Assumption 3, $C_1^{\mathrm{av}} < 1$.

For any policy $\pi \in \Pi_d$ and state measure μ, Assumption 3-(a) (i.e., minorization condition) implies that there exists a unique invariant measure $\nu_{\pi,\mu} \in \mathcal{P}(\mathsf{X})$ of the transition probability $P_{\pi,\mu}(\,\cdot\,|x) := \sum_x p(\,\cdot\,|x,\pi(x),\mu)$ such that for any initial state $x \in \mathsf{X}$, we have

$$J_\mu^{\mathrm{av}}(\pi, x) = \sum_x c(x,\pi(x),\mu)\,\nu_{\pi,\mu}(x),$$

where the last identity follows from ergodic theorem (Hernández-Lerma, 1989, Lemma 3.3). Therefore, the value of any policy π under μ does not depend on the initial state. Let us define

$$J_\mu^{\mathrm{av}}(\pi, x) = J_\mu^{\mathrm{av}}(\pi, y) =: J_\mu^{\mathrm{av}}(\pi), \quad \text{for all } x, y \in \mathsf{X}.$$

Now, fix any policy $\pi \in \Pi_d$. Then, we have

$$\begin{aligned} |J_{\mu_*}^{\mathrm{av}}(\pi) - J_{\mu_\varepsilon}^{\mathrm{av}}(\pi)| &= \left| \sum_x c(x,\pi(x),\mu_*)\,\nu_{\pi,\mu_*}(x) - \sum_x c(x,\pi(x),\mu_\varepsilon)\,\nu_{\pi,\mu_\varepsilon}(x) \right| \\ &\le c_{\mathbf{m}}\,\|\nu_{\pi,\mu_*} - \nu_{\pi,\mu_\varepsilon}\|_1. \end{aligned}$$

Hence, to bound $|J_{\mu_*}^{\mathrm{av}}(\pi) - J_{\mu_\varepsilon}^{\mathrm{av}}(\pi)|$, it is sufficient to bound $\|\nu_{\pi,\mu_*} - \nu_{\pi,\mu_\varepsilon}\|_1$. Note that invariant measures ν_{π,μ_*} and $\nu_{\pi,\mu_\varepsilon}$ satisfy the following fixed point equations

$$\begin{aligned} \nu_{\pi,\mu_*}(\,\cdot\,) &= \sum_x p(\,\cdot\,|x,\pi(x),\mu_*)\,\nu_{\pi,\mu_*}(x) \\ \nu_{\pi,\mu_\varepsilon}(\,\cdot\,) &= \sum_x p(\,\cdot\,|x,\pi(x),\mu_\varepsilon)\,\nu_{\pi,\mu_\varepsilon}(x). \end{aligned}$$

Hence, we have

$$\begin{aligned} \|\nu_{\pi,\mu_*} - \nu_{\pi,\mu_\varepsilon}\|_1 &= \sum_y \left| \sum_x p(y|x,\pi(x),\mu_*)\,\nu_{\pi,\mu_*}(x) - \sum_x p(y|x,\pi(x),\mu_\varepsilon)\,\nu_{\pi,\mu_\varepsilon}(x) \right| \\ &\le \sum_y \left| \sum_x p(y|x,\pi(x),\mu_*)\,\nu_{\pi,\mu_*}(x) - \sum_x p(y|x,\pi(x),\mu_*)\,\nu_{\pi,\mu_\varepsilon}(x) \right| \\ &\quad + \sum_y \left| \sum_x p(y|x,\pi(x),\mu_*)\,\nu_{\pi,\mu_\varepsilon}(x) - \sum_x p(y|x,\pi(x),\mu_\varepsilon)\,\nu_{\pi,\mu_\varepsilon}(x) \right| \\ &\le \sum_y \left| \sum_x p(y|x,\pi(x),\mu_*)\,\nu_{\pi,\mu_*}(x) - \sum_x p(y|x,\pi(x),\mu_*)\,\nu_{\pi,\mu_\varepsilon}(x) \right| \end{aligned}$$

$$+\sum_x \|p(\cdot|x,\pi(x),\mu_*)-p(\cdot|x,\pi(x),\mu_\varepsilon)\|_1\,\nu_{\pi,\mu_\varepsilon}(x)$$
$$\leq \sum_y \left|\sum_x p(y|x,\pi(x),\mu_*)\,\nu_{\pi,\mu_*}(x)-\sum_x p(y|x,\pi(x),\mu_*)\,\nu_{\pi,\mu_\varepsilon}(x)\right| + K_1\,\|\mu_\varepsilon-\mu_*\|_1.$$

Note that (Hernández-Lerma, 1989, Lemma 3.3) implies that for all $x,z\in\mathsf{X}$, we have

$$\|p(\cdot|x,\pi(x),\mu_*)-p(\cdot|z,\pi(z),\mu_*\|_1 \leq (2-\lambda(\mathsf{X}))\,d_{\mathsf{X}}(x,z).$$

Then, by Lemma 32, we have

$$\sum_y \left|\sum_x p(y|x,\pi(x),\mu_*)\,\nu_{\pi,\mu_*}(x)-\sum_x p(y|x,\pi(x),\mu_*)\,\nu_{\pi,\mu_\varepsilon}(x)\right| \leq \frac{2-\lambda(\mathsf{X})}{2}\,\|\nu_{\pi,\mu_*}-\nu_{\pi,\mu_\varepsilon}\|_1.$$

Since $1-\lambda(\mathsf{X})/2<1$, the last inequality gives the following

$$\|\nu_{\pi,\mu_*}-\nu_{\pi,\mu_\varepsilon}\|_1 \leq \frac{2K_1}{\lambda(\mathsf{X})}\,\|\mu_\varepsilon-\mu_*\|_1$$

Therefore, we obtain

$$|J^{\rm av}_{\mu_*}(\pi)-J^{\rm av}_{\mu_\varepsilon}(\pi)| \leq \frac{2c_{\mathbf{m}}(K_1)^2}{(1-C_1^{\rm av})\lambda(\mathsf{X})}\varepsilon =: C_2^{\rm av}\,\varepsilon. \tag{38}$$

By using a similar analysis as above, we can also obtain the following

$$\|\nu_{\pi_*,\mu_*}-\nu_{\pi_\varepsilon,\mu_*}\|_1 \leq \frac{2}{\lambda(\mathsf{X})}\sum_x \|\pi_*(x)-\pi_\varepsilon(x)\|\,\nu_{\pi_\varepsilon,\mu_*}(x).$$

Note that $\sup_x\|\pi_*(x)-\pi_\varepsilon(x)\|\leq\varepsilon$ as π_ε is ε-mean-field equilibrium policy. Therefore, we obtain

$$|J^{\rm av}_{\mu_*}(\pi_*)-J^{\rm av}_{\mu_*}(\pi_\varepsilon)|_\infty \leq \frac{2c_{\mathbf{m}}}{\lambda(\mathsf{X})}\varepsilon =: C_3^{\rm av}\,\varepsilon. \tag{39}$$

Note that we must prove that

$$J_i^{{\rm av},(N)}(\boldsymbol{\pi}^{(N)}) \leq \inf_{\pi^i\in\Pi_i} J_i^{{\rm av},(N)}(\boldsymbol{\pi}^{(N)}_{-i},\pi^i)+\tau^{\rm av}\,\varepsilon+\delta \tag{40}$$

for each $i=1,\ldots,N$, when N is sufficiently large. As the transition probabilities and the one-stage cost functions are the same for all agents, it is sufficient to prove (40) for Agent 1 only. Given $\delta>0$, for each $N\geq 1$, let $\tilde\pi^{(N)}\in\Pi_1$ be a deterministic policy such that

$$J_1^{{\rm av},(N)}(\tilde\pi^{(N)},\pi_\varepsilon,\ldots,\pi_\varepsilon) < \inf_{\pi'\in\Pi_1} J_1^{{\rm av},(N)}(\pi',\pi_\varepsilon,\ldots,\pi_\varepsilon)+\frac{\delta}{3}.$$

On the other hand, by (Wiecek, 2019, Lemma 8) and (Saldi et al., 2018, Theorem 4.10) we get

$$\lim_{N\to\infty} J_1^{{\rm av},(N)}(\tilde\pi^{(N)},\pi_\varepsilon,\ldots,\pi_\varepsilon) = \lim_{N\to\infty} J^{\rm av}_{\mu_\varepsilon}(\tilde\pi^{(N)})$$

$$
\begin{aligned}
&\geq \lim_{N\to\infty} J^{\mathrm{av}}_{\mu_*}(\tilde{\pi}^{(N)}) - C^{\mathrm{av}}_2\varepsilon \quad \text{(by (38))}\\
&\geq \inf_{\pi'\in\Pi_d} J^{\mathrm{av}}_{\mu_*}(\pi') - C^{\mathrm{av}}_2\varepsilon\\
&= J^{\mathrm{av}}_{\mu_*}(\pi_*) - C^{\mathrm{av}}_2\varepsilon\\
&\geq J^{\mathrm{av}}_{\mu_*}(\pi_\varepsilon) - C^{\mathrm{av}}_2\varepsilon - C^{\mathrm{av}}_3\varepsilon \quad \text{(by (39))}\\
&\geq J^{\mathrm{av}}_{\mu_\varepsilon}(\pi_\varepsilon) - 2C^{\mathrm{av}}_2\varepsilon - C^{\mathrm{av}}_3\varepsilon \quad \text{(by (38))}\\
&=: J^{\mathrm{av}}_{\mu_\varepsilon}(\pi_\varepsilon) - \tau^{\mathrm{av}}\,\varepsilon.
\end{aligned}
$$

Note that by (Saldi et al., 2018, Theorem 4.10), we also have

$$\lim_{N\to\infty} J^{\mathrm{av},(N)}_1(\pi_\varepsilon, \pi_\varepsilon, \ldots, \pi_\varepsilon) = J^{\mathrm{av}}_{\mu_\varepsilon}(\pi_\varepsilon).$$

Hence, there exists $N(\delta)$ such that for all $N \geq N(\delta)$, we have

$$
\begin{aligned}
J^{\mathrm{av},(N)}_1(\tilde{\pi}^{(N)}, \pi_\varepsilon, \ldots, \pi_\varepsilon) + \frac{\delta}{3} &\geq J^{\mathrm{av}}_{\mu_\varepsilon}(\pi_\varepsilon) - \tau^{\mathrm{av}}\,\varepsilon\\
J^{\mathrm{av}}_{\mu_\varepsilon}(\pi_\varepsilon) + \frac{\delta}{3} &\geq J^{\mathrm{av},(N)}_1(\pi_\varepsilon, \pi_\varepsilon, \ldots, \pi_\varepsilon).
\end{aligned}
$$

Therefore, for all $N \geq N(\delta)$, we obtain

$$
\begin{aligned}
\inf_{\pi'\in\Pi_1} J^{\mathrm{av},(N)}_1(\pi', \pi_\varepsilon, \ldots, \pi_\varepsilon) + \delta + \tau^{\mathrm{av}}\,\varepsilon &\geq J^{\mathrm{av},(N)}_1(\tilde{\pi}^{(N)}, \pi_\varepsilon, \ldots, \pi_\varepsilon) + \frac{2\delta}{3} + \tau^{\mathrm{av}}\,\varepsilon\\
&\geq J^{\mathrm{av}}_{\mu_\varepsilon}(\pi_\varepsilon) + \frac{\delta}{3}\\
&\geq J^{\mathrm{av},(N)}_1(\pi_\varepsilon, \pi_\varepsilon, \ldots, \pi_\varepsilon).
\end{aligned}
$$

■

Theorem 23 implies that, by learning ε-mean-field equilibrium policy in the infinite-population limit, one can obtain an approximate Nash equilibrium for the finite-agent game problem for which computing or learning the exact Nash equilibrium is in general prohibitive. In the next section, we approximate the MFE operator H^{av} introduced in Section 6 via random operator $\hat{H}^{\mathrm{av}}$ to develop an algorithm for learning ε-mean-field equilibrium policy in the model-free setting.

8. Learning Algorithm for Average-cost

In this section, we develop an offline learning algorithm to learn approximate mean-field equilibrium policy. Similar to the discounted-cost case, we assume that a generic agent has access to a simulator, which generates a new state $y \sim p(\cdot\,|x, a, \mu)$ and gives a cost $c(x, a, \mu)$ for any given state x, action a, and state-measure μ.

In this learning algorithm, there are two stages in each iteration. In the first stage, we learn the Q-function $Q^{\mathrm{av},*}_\mu$ upto a constant additive factor for a given μ using fitted Q-iteration algorithm. This stage replaces the operator H^{av}_1 with a random operator $\hat{H}^{\mathrm{av}}_1$

that will be described below. Note that as opposed to the discounted-cost case, here, to construct the random operator $\hat{H}_1^{\mathrm{av}}$ that replaces the operator H_1^{av}, we normally need an additional simulator that generates realizations of the minorizing sub-probability measure λ in addition to the simulator for the transition probability $p(\,\cdot\,|x,a,\mu)$. However, this simulator is in general not available to the decision maker, since a generic agent does not know this minorizing sub-probability measure in the absence of the transition probability. Therefore, we need to modify the approach used in the discounted-cost case appropriately for the average-cost setup. Indeed, this is achieved by performing convergence analysis of the random operator $\hat{H}_1^{\mathrm{av}}$ using span-seminorm instead of sup-norm on Q-functions. Luckily, convergence analysis of the learning algorithm established using sup-norm in discounted-cost case can easily be adapted to the span-seminorm.

We select Q-functions from a fixed function class $\mathcal{F}^{\mathrm{av}}$ such as the set of neural networks with some fixed architecture or linear span of some finite number of basis functions. Depending on this choice, there will be an additional representation error in the learning algorithm, which is in general negligible. Let $\mathcal{F}_{\min}^{\mathrm{av}} \coloneqq \{Q_{\min} : Q \in \mathcal{F}^{\mathrm{av}}\}$.

In the second stage, we update the state-measure by approximating the transition probability via its empirical estimate. This stage replaces the operator H_2^{av} in the model-based algorithm with a random operator $\hat{H}_2^{\mathrm{av}}$. Indeed, since $H_2^{\mathrm{av}} = H_2$, we also have $\hat{H}_2^{\mathrm{av}} = \hat{H}_2$, and so, the error analysis of $\hat{H}_2^{\mathrm{av}}$ is exactly the same with the error analysis of $\hat{H}_2$.

We proceed with the definition of the random operator $\hat{H}_1^{\mathrm{av}}$. To describe $\hat{H}_1^{\mathrm{av}}$, we need to pick a probability measure ν on X and a policy $\pi_b \in \Pi$. Indeed, we can choose ν and π_b as in discounted-cost case. Recall the constants $\zeta_0 \coloneqq 1/\sqrt{\min_x \nu(x)}$ and $\pi_0 \coloneqq \inf_{(x,a)\in\mathsf{X}\times\mathsf{A}} \pi_b(a|x) > 0$. Now, we can give the definition of the random operator $\hat{H}_1^{\mathrm{av}}$.

Algorithm 4 Algorithm $\hat{H}_1^{\mathrm{av}}$

Input μ, Data size N, Number of iterations L
Generate i.i.d. samples $\{(x_t, a_t, c_t, y_{t+1})_{t=1}^N\}$ using

$$x_t \sim \nu,\ a_t \sim \pi_b(\cdot|x_t),\ c_t = c(x_t,a_t,\mu),\ y_{t+1} \sim p(\cdot|x_t,a_t,\mu)$$

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

$$Q_{l+1} = \operatorname*{arg\,min}_{f\in\mathcal{F}} \frac{1}{N}\sum_{t=1}^{N} \frac{1}{m(\mathsf{A})\,\pi_b(a_t|x_t)} \left| f(x_t,a_t) - \left[c_t + \min_{a'\in\mathsf{A}} Q_l(y_{t+1},a') \right] \right|^2$$

end for
return Q_L

Note that if we used the same method as in the discounted-cost case, we should have generated y_{t+1} using $q(\,\cdot\,|x,a,\mu) \coloneqq p(\,\cdot\,|x,a,\mu) - \lambda(\,\cdot\,)$ instead of $p(\,\cdot\,|x,a,\mu)$ since $H_1^{\mathrm{av}}(\mu)$ gives the unique fixed point of the contraction operator H_μ^{av} on Q-functions given by

$$H_\mu^{\mathrm{av}} Q(x,a) = c(x,a,\mu) + \sum_y Q_{\min}(y)\, q(y|x,a,\mu).$$

However, in general, a generic agent does not have access to a simulator for λ, and so, we must construct the algorithm as above using the simulator for $p(\cdot\,|x,a,\mu)$. As a consequence of this, we perform the error analysis of the above learning algorithm in terms of span-seminorm instead of sup-norm. To this end, for any μ, define the following operator on Q-functions:

$$R^{\rm av}_\mu Q(x,a) \coloneqq c(x,a,\mu) + \sum_{y\in\mathsf{X}} Q_{\min}(y)\, p(dy|x,a,\mu).$$

The operator $R^{\rm av}_\mu$ is different from $H^{\rm av}_\mu$ in this case, and it is used in the proof of the error analysis of $\hat{H}^{\rm av}_1$ in place of $H^{\rm av}_\mu$.

To do error analysis of $\hat{H}^{\rm av}_1$, we need to define the following constants:

$$E(\mathcal{F})^{\rm av} \coloneqq \sup_{\mu\in\mathcal{P}(\mathsf{X})}\sup_{Q\in\mathcal{F}^{\rm av}}\inf_{Q'\in\mathcal{F}^{\rm av}} \|Q'-R^{\rm av}_\mu Q\|_\nu,\ L^{\rm av}_{\mathbf{m}} \coloneqq 2Q^{\rm av}_{\mathbf{m}} + c_{\mathbf{m}},\ C^{\rm av} \coloneqq \frac{(L^{\rm av}_{\mathbf{m}})^2}{m(\mathsf{A})\,\pi_0}$$

$$\Upsilon^{\rm av} = 8\,e^2\,(V_{\mathcal{F}^{\rm av}}+1)\,(V_{\mathcal{F}^{\rm av}_{\min}}+1)\left(\frac{128eQ^{\rm av}_{\mathbf{m}}L^{\rm av}_{\mathbf{m}}}{m(\mathsf{A})\pi_0}\right)^{V_{\mathcal{F}^{\rm av}}+V_{\mathcal{F}^{\rm av}_{\min}}},\ V^{\rm av} = V_{\mathcal{F}^{\rm av}} + V_{\mathcal{F}^{\rm av}_{\min}},\ \gamma^{\rm av} = 512(C^{\rm av})^2$$

$$\Delta^{\rm av} \coloneqq \frac{2}{1-\tilde{\beta}}\left[\frac{m(\mathsf{A})(\dim_{\mathsf{A}}+1)!\zeta_0}{\alpha(2/Q^{\rm av}_{\rm Lip})^{\dim_{\mathsf{A}}}}\,E(\mathcal{F})^{\rm av}\right]^{\frac{1}{\dim_{\mathsf{A}}+1}},\ \Lambda^{\rm av} \coloneqq \frac{2}{1-\tilde{\beta}}\left[\frac{m(\mathsf{A})(\dim_{\mathsf{A}}+1)!\zeta_0}{\alpha(2/Q^{\rm av}_{\rm Lip})^{\dim_{\mathsf{A}}}}\right]^{\frac{1}{\dim_{\mathsf{A}}+1}},$$

where

$$(0,1)\ni\tilde{\beta}\coloneqq 1-\lambda(\mathsf{X})/2 \geq \beta^{\rm av}\coloneqq 1-\lambda(\mathsf{X}).$$

The below theorem gives the error analysis of the algorithm $\hat{H}_1$. Before stating it, let us recall the definition of span-seminorm of any function $g:\mathsf{E}\to\mathbb{R}$ defined on some set E:

$$\operatorname{span}(g) \coloneqq \sup_{e\in\mathsf{E}} g(e) - \inf_{e\in\mathsf{E}} g(e).$$

It is a seminorm because $\operatorname{span}(g)=0$ if and only if g is a constant function. Moreover,

$$\begin{aligned}\operatorname{span}(g) &\coloneqq \sup_{e\in\mathsf{E}} g(e) - \inf_{e\in\mathsf{E}} g(e)\\ &= \sup_{e\in\mathsf{E}} g(e) + \sup_{e\in\mathsf{E}} -g(e)\\ &\leq 2\,\|g\|_\infty.\end{aligned}$$

Hence, we can upper bound span-seminorm via sup-norm.

Theorem 24 *For any $(\varepsilon,\delta)\in(0,1)^2$, with probability at least $1-\delta$, we have*

$$\operatorname{span}\left(\hat{H}^{\rm av}_1[N,L](\mu) - H^{\rm av}_1(\mu)\right) \leq \varepsilon + \Delta^{\rm av}$$

if $\frac{4\tilde{\beta}^L}{1-\tilde{\beta}}\,Q^{\rm av}_{\mathbf{m}} < \frac{\varepsilon}{2}$ and $N \geq m^{\rm av}_1(\varepsilon,\delta,L)$, where

$$m^{\rm av}_1(\varepsilon,\delta,L) \coloneqq \frac{\gamma^{\rm av}(2\Lambda^{\rm av})^{4(\dim_{\mathsf{A}}+1)}}{\varepsilon^{4(\dim_{\mathsf{A}}+1)}}\,\ln\left(\frac{\Upsilon^{\rm av}(2\Lambda^{\rm av})^{2V^{\rm av}(\dim_{\mathsf{A}}+1)}L}{\delta\varepsilon^{2V^{\rm av}(\dim_{\mathsf{A}}+1)}}\right).$$

Here, the constant error $\Delta^{\rm av}$ is as a result of the representation error $E(\mathcal{F})^{\rm av}$ in the algorithm, which is in general negligible.

Proof Recall the definition of $\|\cdot\|_\nu$-norm on Q-functions:

$$\|Q\|_\nu^2 \coloneqq \sum_{x\in\mathsf{X}} \int_{\mathsf{A}} Q(x,a)^2 \, m_{\mathsf{A}}(da)\, \nu(x).$$

For any μ, recall also the definition of the operator on Q-functions:

$$R_\mu^{\mathrm{av}} Q(x,a) \coloneqq c(x,a,\mu) + \sum_{y\in\mathsf{X}} Q_{\min}(y)\, p(dy|x,a,\mu).$$

This is very similar to the operator H_μ^{av}, but it is not β^{av}-contraction with respect to sup-norm. Indeed, the operator R_μ^{av} is $\tilde{\beta}$-contraction with respect to span-seminorm:

$$\operatorname{span}(R_\mu^{\mathrm{av}} Q_1 - R_\mu^{\mathrm{av}} Q_2) \le \tilde{\beta}\, \operatorname{span}(Q_1 - Q_2),$$

where

$$\tilde{\beta} \coloneqq 1 - \lambda(\mathsf{X})/2 \ge \beta^{\mathrm{av}} \coloneqq 1 - \lambda(\mathsf{X}).$$

Indeed, for any function $v : \mathsf{X} \to \mathbb{R}$, one can prove that (see (Hernández-Lerma, 1989, proof of Lemma 3.3 and proof of Lemma 3.5))

$$\sum_y v(y)\, p(y|x,a,\mu) - \sum_y v(y)\, p(y|x',a',\mu) \le \tilde{\beta}\, \operatorname{span}(v). \tag{41}$$

Now let Q_1 and Q_2 be two Q-functions. Then, for any (x,a) and (x',a'), we have

$$\begin{aligned}
&(R_\mu^{\mathrm{av}} Q_1 - R_\mu^{\mathrm{av}} Q_2)(x,a) - (R_\mu^{\mathrm{av}} Q_1 - R_\mu^{\mathrm{av}} Q_2)(x',a') \\
&\quad = \sum_y \{Q_{1,\min}(y) - Q_{2,\min}(y)\}\; p(y|x,a,\mu) - \sum_y \{Q_{1,\min}(y) - Q_{2,\min}(y)\}\; p(y|x',a',\mu) \\
&\quad \le \tilde{\beta}\, \operatorname{span}(Q_{1,\min} - Q_{2,\min}) \quad \text{(by (41))} \\
&\quad \le \tilde{\beta}\, \operatorname{span}(Q_1 - Q_2).
\end{aligned}$$

This implies that

$$\operatorname{span}(R_\mu^{\mathrm{av}} Q_1 - R_\mu^{\mathrm{av}} Q_2) \le \tilde{\beta}\, \operatorname{span}(Q_1 - Q_2),$$

which means that R_μ^{av} is span-seminorm $\tilde{\beta}$-contraction. Moreover, $H_1^{\mathrm{av}}(\mu) \coloneqq Q_\mu^{\mathrm{av},*}$ is a fixed point of R_μ^{av} with respect to span-seminorm; that is,

$$\operatorname{span}(R_\mu^{\mathrm{av}} H_1^{\mathrm{av}}(\mu) - H_1^{\mathrm{av}}(\mu)) = 0.$$

Indeed, for all x, a, we have

$$\begin{aligned}
&R_\mu^{\mathrm{av}} H_1^{\mathrm{av}}(\mu)(x,a) - H_1^{\mathrm{av}}(\mu)(x,a) \\
&\qquad = c(x,a,\mu) + \sum_y Q_{\mu,\min}^{\mathrm{av},*}\, p(y|x,a,\mu) - c(x,a,\mu) - \sum_y Q_{\mu,\min}^{\mathrm{av},*}\, q(y|x,a,\mu) \\
&\qquad = \sum_y Q_{\mu,\min}^{\mathrm{av},*}\, \lambda(y) \;\; \text{(i.e., constant)}
\end{aligned}$$

and so, $\operatorname{span}(R_\mu^{\mathrm{av}} H_1^{\mathrm{av}}(\mu) - H_1^{\mathrm{av}}(\mu)) = 0$. Hence, $H_1^{\mathrm{av}}(\mu)$ is a fixed point of R_μ^{av} with respect to span-seminorm. Since R_μ^{av} is also $\tilde{\beta}$-contraction with respect to span-seminorm, one can also prove that for all $L > 1$, we have

$$\begin{aligned}\operatorname{span}((R_\mu^{\mathrm{av}})^L Q - H_1^{\mathrm{av}}(\mu)) &\leq \frac{\tilde{\beta}^L}{1-\tilde{\beta}} \operatorname{span}(Q - R_\mu^{\mathrm{av}} Q) \\ &\leq \frac{2\tilde{\beta}^L}{1-\tilde{\beta}} \|Q - R_\mu^{\mathrm{av}} Q\|_\infty \\ &\leq \frac{4\tilde{\beta}^L}{1-\tilde{\beta}} Q_{\mathbf{m}}^{\mathrm{av}}\end{aligned}$$

for any $Q \in \mathcal{C}^{\mathrm{av}}$ as $\|Q\|_\infty \leq Q_{\mathbf{m}}^{\mathrm{av}}$.

Now, using above results, we easily complete the proof by using the same techniques as in the proof of Theorem 13. Let Q_l be the random Q-function at the l^{th}-step of the algorithm. First, we find an upper bound to the following probability

$$P_0 \coloneqq \mathbb{P}\left(\|Q_{l+1} - R_\mu^{\mathrm{av}} Q_l\|_\nu^2 > (E(\mathcal{F})^{\mathrm{av}})^2 + \varepsilon'\right),$$

for a given $\varepsilon' > 0$. To that end, we define

$$\hat{L}_N(f; Q) \coloneqq \frac{1}{N} \sum_{t=1}^{N} \frac{1}{m(\mathsf{A})\, \pi_b(a_t|x_t)} \left| f(x_t, a_t) - \left[c_t + \min_{a' \in \mathsf{A}} Q(y_{t+1}, a') \right] \right|^2.$$

As in the proof of Theorem 13, one can show that

$$\mathbb{E}\left[\hat{L}_N(f; Q)\right] = \|f - R_\mu^{\mathrm{av}} Q\|_\nu^2 + L^{\mathrm{av},*}(Q) =: L^{\mathrm{av}}(f; Q),$$

where $L^{\mathrm{av},*}(Q)$ is some quantity independent of f.

Now, using exactly the same steps as in the proof of Theorem 13, we can obtain the following bound on the probability P_0:

$$P_0 \leq \Upsilon^{\mathrm{av}}\, \varepsilon'^{-V^{\mathrm{av}}}\, e^{\frac{-N\varepsilon'^2}{\gamma^{\mathrm{av}}}} =: \frac{\delta'}{L}. \tag{42}$$

The only difference is that in this case, we take $\beta = 1$. Hence, for each $l = 0, \ldots, L-1$, with probability at most $\frac{\delta'}{L}$

$$\|Q_{l+1} - R_\mu^{\mathrm{av}} Q_l\|_\nu^2 > \varepsilon' + (E(\mathcal{F})^{\mathrm{av}})^2.$$

This implies that with probability at most $\frac{\delta'}{L}$

$$\|Q_{l+1} - R_\mu^{\mathrm{av}} Q_l\|_\nu > \sqrt{\varepsilon'} + E(\mathcal{F})^{\mathrm{av}}.$$

Using this and the fact that R_μ^{av} is $\tilde{\beta}$-contraction with respect to span-seminorm, we can conclude that with probability at least $1 - \delta'$, we have

$$\operatorname{span}(Q_L - H_1^{\mathrm{av}}(\mu)) \leq \sum_{l=0}^{L-1} \tilde{\beta}^{L-(l+1)} \operatorname{span}(Q_{l+1} - R_\mu^{\mathrm{av}} Q_l) + \operatorname{span}((R_\mu^{\mathrm{av}})^L Q_0 - H_1^{\mathrm{av}}(\mu))$$

$$
\begin{aligned}
&\leq 2\left(\sum_{l=0}^{L-1}\tilde{\beta}^{L-(l+1)}\left\|Q_{l+1}-R_{\mu}^{\mathrm{av}}Q_l\right\|_{\infty}+\frac{2\tilde{\beta}^{L}}{1-\tilde{\beta}}\,Q_{\mathbf{m}}^{\mathrm{av}}\right)\\
&\overset{(I)}{\leq} 2\sum_{l=0}^{L-1}\tilde{\beta}^{L-(l+1)}\left[\frac{m(\mathsf{A})(\dim_{\mathsf{A}}+1)!\zeta_0}{\alpha(2/Q_{\mathrm{Lip}}^{\mathrm{av}})^{\dim_{\mathsf{A}}}}\left\|Q_{l+1}-R_{\mu}^{\mathrm{av}}Q_l\right\|_{\nu}\right]^{\frac{1}{\dim_{\mathsf{A}}+1}}+\frac{4\tilde{\beta}^{L}}{1-\tilde{\beta}}\,Q_{\mathbf{m}}^{\mathrm{av}}\\
&\leq 2\sum_{l=0}^{L-1}\tilde{\beta}^{L-(l+1)}\left[\frac{m(\mathsf{A})(\dim_{\mathsf{A}}+1)!\zeta_0}{\alpha(2/Q_{\mathrm{Lip}}^{\mathrm{av}})^{\dim_{\mathsf{A}}}}\left(\sqrt{\varepsilon'}+E(\mathcal{F})^{\mathrm{av}}\right)\right]^{\frac{1}{\dim_{\mathsf{A}}+1}}+\frac{4\tilde{\beta}^{L}}{1-\tilde{\beta}}\,Q_{\mathbf{m}}^{\mathrm{av}}\\
&\leq \frac{2}{1-\tilde{\beta}}\Bigg(\left[\frac{m(\mathsf{A})(\dim_{\mathsf{A}}+1)!\zeta_0}{\alpha(2/Q_{\mathrm{Lip}}^{\mathrm{av}})^{\dim_{\mathsf{A}}}}\,E(\mathcal{F})^{\mathrm{av}}\right]^{\frac{1}{\dim_{\mathsf{A}}+1}}\\
&\qquad\qquad+\left[\frac{m(\mathsf{A})(\dim_{\mathsf{A}}+1)!\zeta_0}{\alpha(2/Q_{\mathrm{Lip}}^{\mathrm{av}})^{\dim_{\mathsf{A}}}}\right]^{\frac{1}{\dim_{\mathsf{A}}+1}}{\varepsilon'}^{\frac{1}{2(\dim_{\mathsf{A}}+1)}}\Bigg)+\frac{4\tilde{\beta}^{L}}{1-\tilde{\beta}}\,Q_{\mathbf{m}}^{\mathrm{av}},
\end{aligned}
$$

where (I) follows from Lemma 29. Therefore, with probability at least $1-\delta'$, we have

$$
\operatorname{span}(Q_L-H_1^{\mathrm{av}}(\mu))\leq \Lambda^{\mathrm{av}}{\varepsilon'}^{\frac{1}{2(\dim_{\mathsf{A}}+1)}}+\Delta^{\mathrm{av}}+\frac{4\tilde{\beta}^{L}}{1-\tilde{\beta}}\,Q_{\mathbf{m}}^{\mathrm{av}}. \tag{43}
$$

Now, the result follows by picking $\delta=\delta'\coloneqq L\,\Upsilon^{\mathrm{av}}\,{\varepsilon'}^{-V^{\mathrm{av}}}\,e^{\frac{-N{\varepsilon'}^2}{\gamma^{\mathrm{av}}}}$, $\Lambda^{\mathrm{av}}{\varepsilon'}^{\frac{1}{2(\dim_{\mathsf{A}}+1)}}=\varepsilon/2$, and $\frac{4\tilde{\beta}^{L}}{1-\tilde{\beta}}\,Q_{\mathbf{m}}^{\mathrm{av}}=\varepsilon/2$. ■

We now give the description of the random operator $\hat{H}_2^{\mathrm{av}}$. In this algorithm, the goal is to replace the operator H_2^{av}, which gives the next state-measure, with $\hat{H}_2^{\mathrm{av}}$. Since $H_2^{\mathrm{av}}=H_2$, we also have $\hat{H}_2^{\mathrm{av}}=\hat{H}_2$. Therefore, the error analysis of $\hat{H}_2^{\mathrm{av}}$ is exactly the same with Theorem 15.

Algorithm 5 Algorithm $\hat{H}_2^{\mathrm{av}}$

Inputs (μ,Q), Data size M, Number of iterations $|\mathsf{X}|$
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 $\sum_{x\in\mathsf{X}}p_M(\cdot|x,f_Q(x),\mu)\,\mu(x)$

This is the error analysis of the random operator $\hat{H}_2^{\mathrm{av}}$.

Theorem 25 *For any $(\varepsilon,\delta)\in(0,1)^2$, with probability at least $1-\delta$*

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

if $M \geq m_2^{\mathrm{av}}(\varepsilon, \delta)$*, where*

$$m_2^{\mathrm{av}}(\varepsilon, \delta) \coloneqq \frac{|\mathsf{X}|^2}{\varepsilon^2} \ln\left(\frac{2\,|\mathsf{X}|^2}{\delta}\right).$$

Proof See the proof of Theorem 15. ■

Below, we give the overall description of the learning algorithm for the average-cost. In this algorithm, we successively apply the random operator $\hat{H}^{\mathrm{av}}$, which replaces MFE operator H^{av}, to obtain approximate mean-field equilibrium policy.

Algorithm 6 Learning Algorithm

Input μ_0, Number of iterations K, Parameters of $\hat{H}_1^{\mathrm{av}}$ and $\hat{H}_2^{\mathrm{av}}$ $\left(\{[N_k, L_k]\}_{k=0}^{K-1}, \{M_k\}_{k=0}^{K-1}\right)$
Start with μ_0
for $k = 0, \ldots, K-1$ **do**
$\mu_{k+1} = \hat{H}^{\mathrm{av}}\left([N_k, L_k], M_k\right)(\mu_k) \coloneqq \hat{H}_2^{\mathrm{av}}[M_k]\left(\mu_k, \hat{H}_1^{\mathrm{av}}[N_k, L_k](\mu_k)\right)$
end for
return μ_K

Above, we have completed the error analyses of the operators $\hat{H}_1^{\mathrm{av}}$ and $\hat{H}_2^{\mathrm{av}}$ in Theorem 24 and Theorem 25, respectively. Since the random operator $\hat{H}^{\mathrm{av}}$ is a composition of $\hat{H}_1^{\mathrm{av}}$ with $\hat{H}_2^{\mathrm{av}}$, we can obtain the following error analysis for the operator $\hat{H}^{\mathrm{av}}$.

Theorem 26 *Fix any* $(\varepsilon, \delta) \in (0,1)^2$*. Define*

$$\varepsilon_1 \coloneqq \frac{\rho\,(1 - K_{H^{\mathrm{av}}})^2\,\varepsilon^2}{32(K_1)^2},\ \varepsilon_2 \coloneqq \frac{(1 - K_{H^{\mathrm{av}}})\,\varepsilon}{4}.$$

Let K, L *be such that*

$$\frac{(K_{H^{\mathrm{av}}})^K}{1 - K_{H^{\mathrm{av}}}} \leq \frac{\varepsilon}{2},\ \frac{4\tilde{\beta}^L}{1 - \tilde{\beta}} Q_{\mathbf{m}}^{\mathrm{av}} \leq \frac{\varepsilon_1}{2}.$$

Then, pick N, M *such that*

$$N \geq m_1^{\mathrm{av}}\left(\varepsilon_1, \frac{\delta}{2K}, L\right),\ M \geq m_2^{\mathrm{av}}\left(\varepsilon_2, \frac{\delta}{2K}\right). \tag{44}$$

Let μ_K *be the output of the learning algorithm* $\hat{H}^{\mathrm{av}}$ *with inputs*

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

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

$$\|\mu_K - \mu_*^{\mathrm{av}}\|_1 \leq \frac{K_1\sqrt{2\Delta^{\mathrm{av}}}}{\sqrt{\rho}(1 - K_{H^{\mathrm{av}}})} + \varepsilon,$$

where μ_*^{av} *is the state-measure in mean-field equilibrium given by the MFE operator* H^{av}*.*

Proof The proof is similar to the proof of Theorem 16. The key difference is that we perform the analysis in terms of span-seminorm in place of sup-norm.

For any $\mu \in \mathcal{P}(\mathsf{X})$, $Q \in \mathcal{F}^{\mathrm{av}}$, $\hat{Q} \in \mathcal{C}^{\mathrm{av}}$, we have

$$\begin{aligned}\|H_2^{\mathrm{av}}(\mu, Q) - H_2^{\mathrm{av}}(\mu, \hat{Q})\|_1 &= \sum_{y \in \mathsf{X}} \left| \sum_{x \in \mathsf{X}} p(y|x, f_Q(x), \mu)\, \mu(x) - \sum_{x \in \mathsf{X}} p(y|x, f_{\hat{Q}}(x), \mu)\, \mu(x) \right| \\ &\leq \sum_{x \in \mathsf{X}} K_1 \, \|f_Q(x) - f_{\hat{Q}}(x)\|\, \mu(x). \end{aligned} \tag{45}$$

Now, using exactly the same steps as in the proof of Theorem 16, for any $x \in \mathsf{X}$, we have

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

For any $x \in \mathsf{X}$, this leads to

$$\begin{aligned}\|f_{\hat{Q}}(x) - f_Q(x)\|^2 &\leq \frac{2}{\rho} \left(\hat{Q}(x, f_Q(x)) - Q(x, f_Q(x)) + Q(x, f_Q(x)) - \hat{Q}(x, f_{\hat{Q}}(x)) \right) \\ &= \frac{2}{\rho} \left(\hat{Q}(x, f_Q(x)) - Q(x, f_Q(x)) + \min_{a \in \mathsf{A}} Q(x, a) - \min_{a \in \mathsf{A}} \hat{Q}(x, a) \right) \\ &\leq \frac{2}{\rho} \left(\sup_{(z,a) \in \mathsf{X} \times \mathsf{A}} (\hat{Q}(z, a) - Q(z, a)) + \sup_{(z,a) \in \mathsf{X} \times \mathsf{A}} (Q(z, a) - \hat{Q}(z, a)) \right) \\ &= \frac{2}{\rho}\, \mathrm{span}(Q - \hat{Q}). \end{aligned} \tag{46}$$

Therefore, here, we can also perform a similar analysis as in the proof of Theorem 16 using span-seminorm in place of sup-norm.

Now, combining (45) and (46) yields

$$\|H_2^{\mathrm{av}}(\mu, Q) - H_2^{\mathrm{av}}(\mu, \hat{Q})\|_1 \leq \frac{\sqrt{2} K_1}{\sqrt{\rho}} \sqrt{\mathrm{span}(Q - \hat{Q})}. \tag{47}$$

Using (47) and the fact that $H_1^{\mathrm{av}}(\mu_k) \in \mathcal{C}^{\mathrm{av}}$ and $\hat{H}_1^{\mathrm{av}}[N, L](\mu_k)) \in \mathcal{F}^{\mathrm{av}}$, for any $k = 0, \ldots, K-1$, we have

$$\begin{aligned}\|H^{\mathrm{av}}(\mu_k) - \hat{H}^{\mathrm{av}}([N, L], M)(\mu_k)\|_1 &\leq \|H_2^{\mathrm{av}}(\mu_k, H_1^{\mathrm{av}}(\mu_k)) - H_2^{\mathrm{av}}(\mu_k, \hat{H}_1^{\mathrm{av}}[N, L](\mu_k))\|_1 \\ &\quad + \|H_2^{\mathrm{av}}(\mu_k, \hat{H}_1^{\mathrm{av}}[N, L](\mu_k)) - \hat{H}_2^{\mathrm{av}}[M](\mu_k, \hat{H}_1^{\mathrm{av}}[N, L](\mu_k))\|_1 \\ &\leq \frac{\sqrt{2} K_1}{\sqrt{\rho}} \sqrt{\mathrm{span}(H_1^{\mathrm{av}}(\mu_k) - \hat{H}_1^{\mathrm{av}}[N, L](\mu_k))} \\ &\quad + \|H_2^{\mathrm{av}}(\mu_k, \hat{H}_1^{\mathrm{av}}[N, L](\mu_k)) - \hat{H}_2^{\mathrm{av}}[M](\mu_k, \hat{H}_1^{\mathrm{av}}[N, L](\mu_k))\|_1. \end{aligned}$$

The last term is upper bounded by

$$\frac{K_1 \sqrt{2(\varepsilon_1 + \Delta^{\mathrm{av}})}}{\sqrt{\rho}} + \varepsilon_2$$

with probability at least $1-\frac{\delta}{K}$ by Theorem 24 and Theorem 25. Therefore, with probability at least $1-\delta$

$$\begin{aligned}\|\mu_K-\mu_*^{\mathrm{av}}\|_1 &\le \sum_{k=0}^{K-1} K_{H^{\mathrm{av}}}^{K-(k+1)}\,\|\hat{H}^{\mathrm{av}}([N,L],M)(\mu_k)-H^{\mathrm{av}}(\mu_k)\|_1+\|(H^{\mathrm{av}})^K(\mu_0)-\mu_*^{\mathrm{av}}\|_1\\ &\le \sum_{k=0}^{K-1} K_{H^{\mathrm{av}}}^{K-(k+1)}\left(\frac{K_1\sqrt{2(\varepsilon_1+\Delta^{\mathrm{av}})}}{\sqrt{\rho}}+\varepsilon_2\right)+\frac{(K_{H^{\mathrm{av}}})^K}{1-K_{H^{\mathrm{av}}}}\\ &\le \frac{K_1\sqrt{2\,\Delta^{\mathrm{av}}}}{\sqrt{\rho}(1-K_{H^{\mathrm{av}}})}+\varepsilon.\end{aligned}$$

This completes the proof. ■

Now, we state the main result of this section. It states that, by using a learning algorithm, one can learn approximate mean-field equilibrium policy. By Theorem 23, this gives an approximate Nash-equilibrium for the finite-agent game.

Corollary 27 *Fix any $(\varepsilon,\delta)\in(0,1)^2$. Suppose that K,L,N,M satisfy the conditions in Theorem 26. Let μ_K be the output of the earning algorithm with inputs*

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

Define $\pi_K(x)\coloneqq\arg\min_{a\in\mathsf{A}}Q_K(x,a)$, where $Q_K\coloneqq\hat{H}_1^{\mathrm{av}}([N,L])(\mu_K)$. Then, with probability at least $1-\delta(1+\frac{1}{2K})$, the policy π_K is a $\kappa^{\mathrm{av}}(\varepsilon,\Delta)$-mean-field equilibrium policy, where

$$\kappa^{\mathrm{av}}(\varepsilon,\Delta)=\sqrt{\frac{2}{\rho}\left(\frac{\rho\,(1-K_{H^{\mathrm{av}}})^2\,\varepsilon^2}{32(K_1)^2}+\Delta+2K_{H_1^{\mathrm{av}}}\left(\frac{K_1\sqrt{2\,\Delta}}{\sqrt{\rho}(1-K_{H^{\mathrm{av}}})}+\varepsilon\right)\right)}.$$

Therefore, by Theorem 23, an N-tuple of policies $\pi^{(N)}=\{\pi_K,\pi_K,\dots,\pi_K\}$ is a $\tau^{\mathrm{av}}\kappa^{\mathrm{av}}(\varepsilon,\Delta)+\sigma$-Nash equilibrium for the game with $N\ge N(\sigma)$ agents.

Proof By Theorem 26, with probability at least $1-\delta(1+\frac{1}{2K})$, we have

$$\begin{aligned}\operatorname{span}(Q_K-H_1^{\mathrm{av}}(\mu_*^{\mathrm{av}})) &\le \operatorname{span}(Q_K-H_1^{\mathrm{av}}(\mu_K))+2\|H_1^{\mathrm{av}}(\mu_K)-H_1^{\mathrm{av}}(\mu_*^{\mathrm{av}})\|_\infty\\ &\le \varepsilon_1+\Delta^{\mathrm{av}}+2K_{H_1^{\mathrm{av}}}\|\mu_K-\mu_*^{\mathrm{av}}\|_1\\ &\le \varepsilon_1+\Delta^{\mathrm{av}}+2K_{H_1^{\mathrm{av}}}\left(\frac{K_1\sqrt{2\,\Delta^{\mathrm{av}}}}{\sqrt{\rho}(1-K_{H^{\mathrm{av}}})}+\varepsilon\right)\\ &= \frac{\rho\,(1-K_{H^{\mathrm{av}}})^2\,\varepsilon^2}{32(K_1)^2}+\Delta^{\mathrm{av}}+2K_{H_1^{\mathrm{av}}}\left(\frac{K_1\sqrt{2,\Delta^{\mathrm{av}}}}{\sqrt{\rho}(1-K_{H^{\mathrm{av}}})}+\varepsilon\right).\end{aligned}$$

Let $\pi_K(x)\coloneqq\arg\min_{a\in\mathsf{A}}Q_K(x,a)$. Using the same analysis that leads to (46), we can obtain the following bound:

$$\sup_{x\in\mathsf{X}}\|\pi_K(x)-\pi_*(x)\|^2\le\frac{2}{\rho}\operatorname{span}(Q_K-H_1^{\mathrm{av}}(\mu_*^{\mathrm{av}})).$$

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

Remark 28 *Note that, in Corollary 27, there is a constant Δ^{av}, which depends on the representation error $E(\mathcal{F})^{\mathrm{av}}$. In general, $E(\mathcal{F})^{\mathrm{av}}$ is negligible. Hence, in this case, we have the following error bound:*

$$\kappa^{\mathrm{av}}(\varepsilon,0)=\sqrt{\frac{2}{\rho}\left(\frac{\rho\,(1-K_{H^{\mathrm{av}}})^2\,\varepsilon^2}{32(K_1)^2}+2K_{H_1^{\mathrm{av}}}\varepsilon\right)}.$$

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

9. Numerical Examples

In this section, we present two numerical examples in the case of discounted cost and average cost, respectively, to demonstrate the applicability of our learning algorithm.

9.1 Discounted Cost

We consider the mean-field game that was introduced in Example 1, where we take $\mathsf{X}=[0,0.1,0.,2,\dots,1]$, $\mathsf{A}=[0,1]$, $c_2(a)=\rho\,a^2$, and $c_1(x,\mu)=\eta\,x\,(1-\xi\,\langle\mu\rangle)$ with $\langle\mu\rangle$ denoting the mean of μ. We also take $\mathbb{P}[h(a,\mu,w)=0.1]=\kappa\,a\,(1-\gamma\,\langle\mu\rangle)$, $\mathbb{P}[h(a,\mu,w)=0]=\kappa\,a\,\gamma\,\langle\mu\rangle$, and $\mathbb{P}[h(a,\mu,w)=-0.1]=\kappa\,a$; that is, the state can only go one unit up, go one unit down, or remain the same. Here, the constants $\rho,\eta,\xi,\kappa,\gamma$ are all non-negative. As the conditions (i) and (ii) are clearly satisfied in Example 1 for these particular choices of system components, Assumption 1 holds true in this case. In the numerical experiments, we use the following values for the parameters: $\eta=2,\quad \xi=0.4,\quad \rho=1\quad \kappa=1,\quad \gamma=0.4,\quad \beta=0.9$. We run the learning algorithm using the following parameters: $N=10000, L=50, M=1000, K=50$. The output of the learning algorithm contains the average of the state-measure (i.e., mean-field distribution) and mean-field equilibrium policies for states $x=0.1$ and $x=0.6$. In the fitted Q-iteration algorithm, we pick the function class $\mathcal{F}$ as two-layer neural networks with 10 hidden units. We use neural network fitting tool of MATLAB. In particular, we use 'fittnet', '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 set to default values. We also run the value iteration algorithm using MFE operator H to find the true average of state-measure and mean-field equilibrium policies for states $x=0.1$ and $x=0.6$. Then, we compare the learned outputs with outputs of the value iteration algorithm. Figures 1, 2, and 3 show this comparison. It can be seen that learned outputs converge to the outputs of the value iteration algorithm.

9.2 Average Cost

We consider a mean-field game with state space $\mathsf{X}=\{0,1\}$ and 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

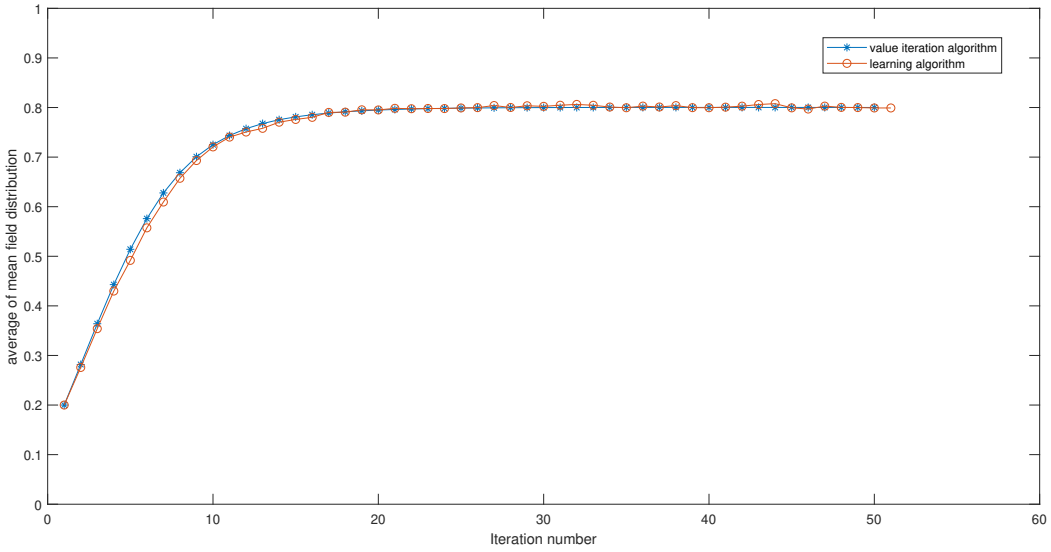

Figure 1: Comparison of state-measures: discounted-cost

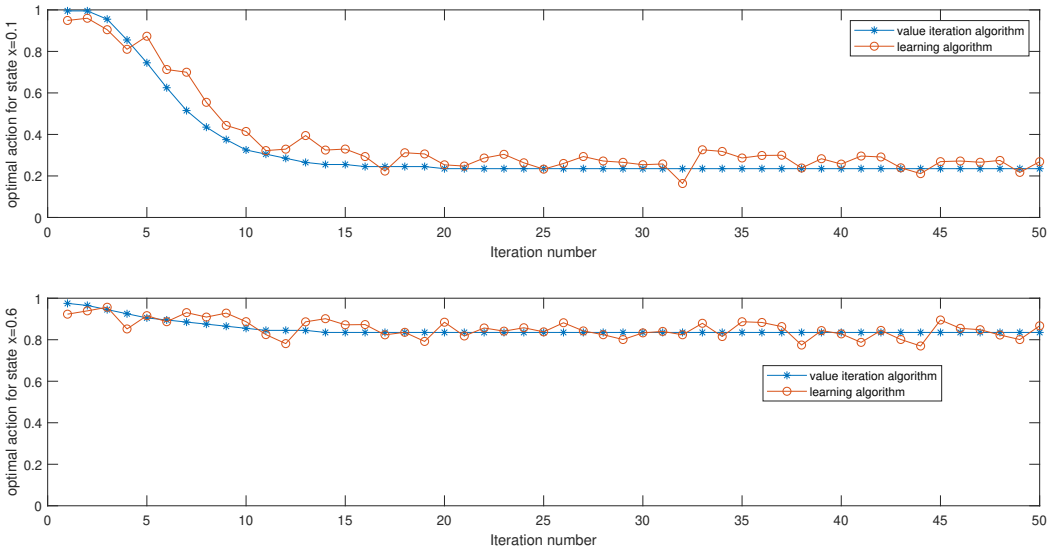

Figure 2: Comparison of policies: discounted-cost

by

$$p(\cdot\,|x,a) = l_0(\cdot\,|x,)\cdot a + l_1(\cdot\,|x)\cdot(1-a),$$

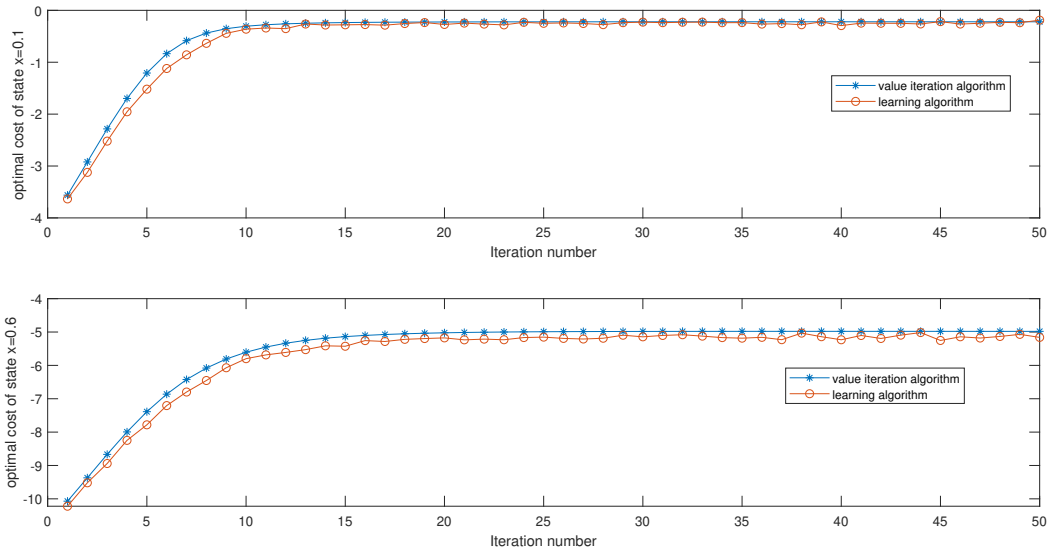

Figure 3: Comparison of costs: discounted-cost

where

$$l_0(1|0) = \eta, \quad l_0(1|1) = 1 - \alpha,$$
$$l_1(1|0) = \kappa, \quad l_1(1|1) = 1 - \xi.$$

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

$$c(x, a, \mu) = \tau \langle \mu \rangle x + \lambda (1 - \langle \mu \rangle)(1 - a) + \gamma a^2,$$

where $\langle \mu \rangle$ is the mean of the distribution μ on X. It can be verified that Assumption 1 holds in this example. We use the following values of the parameters:

$$\eta = 0.7, \quad \alpha = 0.1, \quad \kappa = 0.1, \quad \xi = 0.8$$
$$\tau = 0.1, \quad \lambda = 0.4, \quad \gamma = 0.2.$$

With these parameters, it is also straightforward to check that Assumption 3-(a) holds. We run the learning algorithm using the following parameters: $N = 1000$, $L = 50$, $M = 1000$, $K = 50$. Output of the learning algorithms contain the average of the state-measure (i.e., mean-field distribution) and mean-field equilibrium policies. In the fitted Q-iteration algorithm, we pick the function class $\mathcal{F}$ as two-layer neural networks with 20 hidden units. We use the neural network fitting tool of MATLAB. In particular, we use 'fittnet', '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 set to default values. We also run the value iteration algorithm using MFE operator H^{av} to find the correct average of

state-measure and mean-field equilibrium policies. Then, we compare the learned outputs with the outputs of the value iteration algorithm. Figures 4 and 5 show this comparison for the average-cost. It can be seen that learned outputs converge to the outputs of the value iteration algorithm.

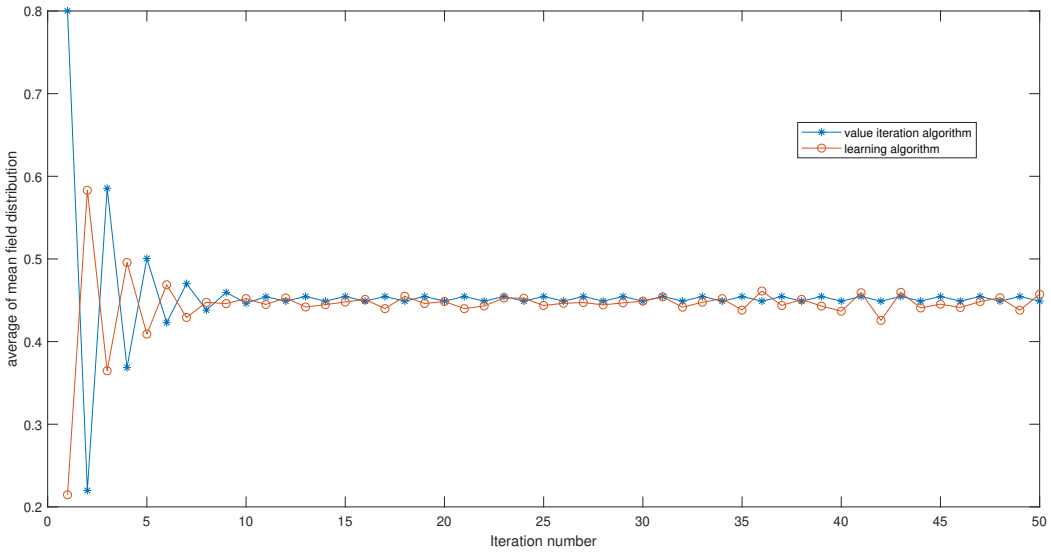

Figure 4: Comparison of state-measures: average-cost

10. Conclusion

This paper has established a learning algorithm for discrete-time mean-field games subject to discounted-cost and average-cost criteria. Under certain regularity conditions on system components, we have proved that the policy obtained from the learning algorithm converges to the policy in mean-field equilibrium with some probabilistic convergence rate. We have then used the learned policy to construct an approximate Nash equilibrium for the finite-agent game problem.

Appendix

In this appendix, we state some auxiliary results that will be frequently used in the paper. The first result gives a bound on l_∞-norm of uniformly Lipschitz continuous function $g(x,a)$ with respect to the action a in terms of its l_2-norm.

Lemma 29 *Let $g : \mathsf{X} \times \mathsf{A} \to \mathbb{R}$ be a uniformly Lipschitz continuous function of the action a with Lipschitz constant L. Then, under Assumption 1-(c), we have*

$$\|g\|_\infty \leq \max\left(\left[\frac{m(\mathsf{A})\,(\dim_{\mathsf{A}}+1)!\,\zeta_0}{\alpha\,(2/L)^{\dim_{\mathsf{A}}}}\,\|g\|_\nu\right]^{1/(\dim_{\mathsf{A}}+1)}, (\dim_{\mathsf{A}}+1)\,\zeta_0\,\|g\|_\nu\right),$$

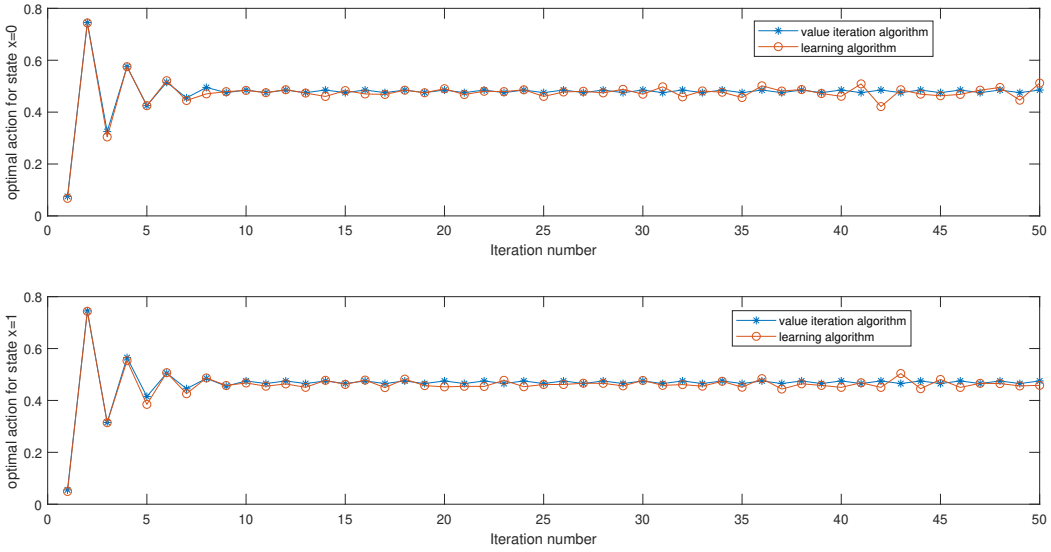

Figure 5: Comparison of policies: average-cost

where $\alpha > 0$ *is the constant in Assumption 1-(c) and* $\zeta_0 := \dfrac{1}{\sqrt{\min_x \nu(x)}}$.

Proof Under Assumption 1-(c), by following the same steps as in the proof of (Antos et al., 2007b, Lemma D.2), for all $(x,a) \in \mathsf{X} \times \mathsf{A}$, we obtain the following

$$\left(\int_{\mathsf{A}} |g(x,\hat{a})|^2 \, m_{\mathsf{A}}(d\hat{a}) \right)^{1/2} \geq \min \left(\left[\frac{\alpha \, (2/L)^{\dim_{\mathsf{A}}}}{m(\mathsf{A}) \, (\dim_{\mathsf{A}}+1)!} \, |g(x,a)| \right]^{(\dim_{\mathsf{A}}+1)} , \frac{|g(x,a)|}{(\dim_{\mathsf{A}}+1)} \right)$$

Note that we also have

$$\left(\sum_x \int_{\mathsf{A}} |g(x,a)|^2 \, m_{\mathsf{A}}(da) \, \nu(x) \right)^{1/2} \geq \sqrt{\min_x \nu(x)} \, \sup_x \left(\int_{\mathsf{A}} |g(x,a)|^2 \, m_{\mathsf{A}}(da) \right)^{1/2}.$$

Therefore, above inequalities lead to

$$\|g\|_\infty \leq \max \left(\left[\frac{m(\mathsf{A}) \, (\dim_{\mathsf{A}}+1)! \, \zeta_0}{\alpha \, (2/L)^{\dim_{\mathsf{A}}}} \, \|g\|_\nu \right]^{1/(\dim_{\mathsf{A}}+1)} , (\dim_{\mathsf{A}}+1) \, \zeta_0 \, \|g\|_\nu \right).$$

■

Remark 30 *In the paper, to simplify the notation, we will always assume that*

$$\left[\frac{m(\mathsf{A}) \ (\dim_{\mathsf{A}}+1)! \, \zeta_0}{\alpha \, (2/L)^{\dim_{\mathsf{A}}}} \, \|g\|_\nu \right]^{1/(\dim_{\mathsf{A}}+1)} \geq (\dim_{\mathsf{A}}+1) \, \zeta_0 \, \|g\|_\nu.$$

Therefore, the bound in Lemma 29 will always be in the following form:

$$\|g\|_\infty \leq \left[\frac{m(\mathsf{A})\ (\dim_{\mathsf{A}}+1)!\,\zeta_0}{\alpha\,(2/L)^{\dim_{\mathsf{A}}}}\,\|g\|_\nu\right]^{1/(\dim_{\mathsf{A}}+1)}.$$

Before we state the next result, we need to give some definitions. Let E be some set. Let $\mathcal{G}$ be a set of real-valued functions on E taking values in $[0,K]$. For any $e^{1:N} \coloneqq \{e_i\}_{i=1}^N \in \mathsf{E}^N$, define the following semi-metric on $\mathcal{G}$:

$$d_{e^{1:N}}(g,h) \coloneqq \frac{1}{N}\sum_{i=1}^N |g(e_i)-h(e_i)|.$$

Then, for any $\varepsilon>0$, let $N_1\left(\varepsilon,\{e_i\}_{i=1}^N,\mathcal{G}\right)$ denote the ε-covering number of $\mathcal{G}$ in terms of semi-metric $d_{e^{1:N}}$ (Vidyasagar, 2010, pp. 14). Moreover, let $V_{\mathcal{G}}$ denote the pseudo-dimension of the function class $\mathcal{G}$ (Vidyasagar, 2010, Definition 4.2, pp. 120).

Lemma 31 *(Antos et al., 2007b, Proposition E.3) For any $e^{1:N}$, we have*

$$N_1\left(\varepsilon,\{e_i\}_{i=1}^N,\mathcal{G}\right) \leq e\,(V_{\mathcal{G}}+1)\left(\frac{2eK}{\varepsilon}\right)^{V_{\mathcal{G}}}.$$

Let $P(\,\cdot\,|x)$ be a transition probability on X with the following contraction coefficient

$$\theta_P \coloneqq \frac{1}{2}\sup_{x,z}\|P(\,\cdot\,|x)-P(\,\cdot\,|z)\|_1.$$

Then, the following result holds.

Lemma 32 *(Kontorovich and Ramanan, 2008, Lemma A.2) Let $\mu,\nu\in\mathcal{P}(\mathsf{X})$. Then,*

$$\sum_y\left|\sum_x P(y|x)\,\mu(x)-\sum_x P(y|x)\,\nu(x)\right| \leq \theta_P\,\|\mu-\nu\|_1.$$

In other words, if we define $\mu P(\,\cdot\,)\coloneqq\sum_x P(\,\cdot\,|x)\,\mu(x)\in\mathcal{P}(\mathsf{X})$ and $\nu P(\,\cdot\,)\coloneqq\sum_x P(\,\cdot\,|x)\,\nu(x)\in\mathcal{P}(\mathsf{X})$, then we have $\|\mu P-\nu P\|_1\leq\theta_P\,\|\mu-\nu\|_1$. Indeed, the last inequality explains why θ_P is called contraction coefficient.

Acknowledgments

This research was supported by The Scientific and Technological Research Council of Turkey (TÜBİTAK) BİDEB 2232 Research Grant.

References

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

A. Agarwal, N. Jiang, and S. Kakade. Reinforcement learning: Theory and algorithms. 2019.

C.D. Aliprantis and K.C. Border. *Infinite Dimensional Analysis*. Berlin, Springer, 3rd ed., 2006.

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

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

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*, pages 9–16, 2007a.

A. Antos, R. Munos, and C. Szepesvári. Fitted Q-iteration in continuous action-space MDPs. Technical report, 2007b.

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

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

J.F. Bonnans and A. Shapiro. *Perturbation Analysis of Optimization Problems*. Springer, New York, 2000.

P. Cardaliaguet. *Notes on Mean-field Games*. 2011.

R. Carmona and F. Delarue. Probabilistic analysis of mean-field games. *SIAM J. Control Optim.*, 51(4):2705–2734, 2013.

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

R. Carmona, M. Lauriere, and Z. Tan. Model-free mean-field reinforcement learning: Mean-field MDP and mean-field Q-learning. arXiv:1910.12802, 2019b.

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

R. Elliot, X. Li, and Y. Ni. Discrete time mean-field stochastic linear-quadratic optimal control problems. *Automatica*, 49:3222–3233, 2013.

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.

H.O. Georgii. *Gibbs Measures and Phase Transitions*. De Gruyter studies in mathematics. De Gruyter, 2011.

D.A. Gomes and J. Saúde. Mean field games models - a brief survey. *Dyn. Games Appl.*, 4 (2):110–154, 2014.

D.A. Gomes, J. Mohr, and R.R. Souza. Discrete time, finite state space mean field games. *J. Math. Pures Appl.*, 93:308–328, 2010.

X. Guo, A. Hu, R. Xu, and J. Zhang. Learning mean-field games. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019.

M. Hairer. Ergodic properties of Markov processes. *Lecture Notes*, 2006.

B. Hajek and M. Raginsky. Statistical learning theory. *Lecture Notes*, 2019.

O. Hernández-Lerma. *Adaptive Markov Control Processes*. Springer-Verlag, 1989.

O. Hernández-Lerma and J.B. Lasserre. *Discrete-Time Markov Control Processes: Basic Optimality Criteria*. Springer, 1996.

O. Hernández-Lerma and J.B. Lasserre. *Further Topics on Discrete-Time Markov Control Processes*. Springer, 1999.

O. Hernández-Lerma, R. Montes-De-Oca, and R. Cavazos-Cadena. Recurrence conditions for Markov decision processes with Borel state space: a survey. *Ann. Oper. Res.*, 28(1): 29–46, 1991.

M. Huang. Large-population LQG games involving major player: The Nash certainty equivalence principle. *SIAM J. Control Optim.*, 48(5):3318–3353, 2010.

M. Huang and Y. Ma. Binary mean field stochastic games: Stationary equilibria and comparative statics. *Modeling, Stochastic Control, Optimization, and Applications*, 2019.

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

M. Huang, P.E. Caines, and R.P. Malhamé. 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, 2007.

I. Kash, E. Friedman, and J. Halpern. Multiagent learning in large anonymous games. *Journal of Artificial Intelligence Research*, 40:571–598, 2011.

L. Kontorovich and K. Ramanan. Concentration inequalities for dependent random variables via the martingale method. *The Annals of Probability*, 36(6):2126–2158, 2008.

H.J. Langen. Convergence of dynamic programming models. *Math. Oper. Res.*, 6(4):493–512, Nov. 1981.

J. Lasry and P.Lions. Mean field games. *Japan. J. Math.*, 2:229–260, 2007.

B. Light and G.Y. Weintraub. Mean field equilibrium: Uniqueness, existence, and comparative statics. *Operations Research*, 70(1):585–605, 2022.

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.

J. Moon and T. Başar. Robust mean field games for coupled Markov jump linear systems. *International Journal of Control*, 89(7):1367–1381, 2016a.

J. Moon and T. Başar. Discrete-time mean field Stackelberg games with a large number of followers. In *CDC 2016*, Las Vegas, Dec. 2016b.

R. Munos and C. Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(27):815–857, 2008.

M. Nourian and G.N. Nair. Linear-quadratic-Gaussian mean field games under high rate quantization. In *CDC 2013*, Florence, Dec. 2013.

D. Pollard. *Convergence of Stochastic Processes*. Springer, New York, 1984.

N. Saldi. Discrete-time average-cost mean-field games on Polish spaces. *Turkish Journal of Mathematics*, 44:463 – 480, 2020.

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

N. Saldi, T. Başar, and M. Raginsky. Approximate Nash equilibria in partially observed stochastic games with mean-field interactions. *Mathematics of Operations Research*, 44 (3):1006–1033, 2019.

N. Saldi, T. Başar, and M. Raginsky. Approximate Markov-Nash equilibria for discrete-time risk-sensitive mean-field games. *Mathematics of Operations Research*, 45(4):1596–1620, 2020.

J. Subramanian and A. Mahajan. A policy gradient algorithm to compute boundedly rational stationary mean field equilibria. ICML/IJCAI/AAMAS Workshop on Planning and Learning, 2018.

J. Subramanian and A. Mahajan. Reinforcement learning in stationary mean-field games. page 251–259. International Foundation for Autonomous Agents and Multiagent Systems, 2019.

H. Tembine, Q. Zhu, and T. Başar. Risk-sensitive mean field games. *IEEE. Trans. Autom. Control*, 59(4):835–850, 2014.

M. Vidyasagar. *Learning and Generalization: With Applications to Neural Networks*. Springer, 2nd edition, 2010.

G.Y. Weintraub, C.L. Benkard, and B. Van Roy. Oblivious equilibrium: A mean field approximation for large-scale dynamic games. volume 18, 01 2005.

G.Y. Weintraub, C.L. Benkard, and B. Van Roy. Markov perfect industry dynamics with many firms. *Econometrica*, 76(6):1375–1411, 2008.

G.Y. Weintraub, C.L. Benkard, and B. Van Roy. Computational methods for oblivious equilibrium. *Operations Research*, 58(4-part-2):1247–1265, 2010.

P. Wiecek. Discrete-time ergodic mean-field games with average reward on compact spaces. *Dynamic Games and Applications*, pages 1–35, 2019.

P. Wiecek and E. Altman. Stationary anonymous sequential games with undiscounted rewards. *Journal of Optimization Theory and Applications*, 166(2):686–710, 2015.

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

Y. Yang, R. Luo, M. Li, M. Zhou, W. Zhang, and J. Wang. Mean field multi-agent reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5571–5580, 10–15 Jul 2018b.

H. Yin, P. Mehta, S. Meyn, and U. Shanbhag. Learning in mean-field games. *Automatic Control, IEEE Transactions on*, 59:629–644, 03 2014.