Reinforcement Learning for Mean Field Games with Strategic Complementarities

Kiyeob Lee* **Desik Rengarajan*** **Dileep Kalathil** **Srinivas Shakkottai**
Texas A&M University
{kiyeoblee,desik,dileep.kalathil,sshakkot}@tamu.edu

Abstract

Mean Field Games (MFG) are the class of games with a very large number of agents and the standard equilibrium concept is a Mean Field Equilibrium (MFE). Algorithms for learning MFE in dynamic MFGs are unknown in general. Our focus is on an important subclass that possess a monotonicity property called Strategic Complementarities (MFG-SC). We introduce a natural refinement to the equilibrium concept that we call Trembling-Hand-Perfect MFE (T-MFE), which allows agents to employ a measure of randomization while accounting for the impact of such randomization on their payoffs. We propose a simple algorithm for computing T-MFE under a known model. We also introduce a model-free and a model-based approach to learning T-MFE and provide sample complexities of both algorithms. We also develop a fully online learning scheme that obviates the need for a simulator. Finally, we empirically evaluate the performance of the proposed algorithms via examples motivated by real-world applications.

1 Introduction

Strategic complementarities refers to a well-established strategic game structure wherein the marginal returns to increasing one's strategy rise with increases in the competitors' strategies, and many qualitative results of the dynamic systems are based on these properties (Milgrom and Roberts, 1990; Vives, 2009; Adlakha and Johari, 2013). Many practical scenarios with this property, including pricing in oligopolistic markets, adoption of network technology and standards, deposits and withdrawals in banking, weapons purchasing in arms races etc., have been identified, and our running example is that of adopting actions to prevent the spread of a computer or human virus, wherein stronger actions towards maintaining health (eg., installing patches, or wearing masks) by members enhances the returns (eg., system reliability or economic value) to a particular individual following suit.

The above examples are characterized by a large number of agents following the dynamics of repeated action, reward, and state transition that is a characteristic of stochastic games. However, analytical complexity implies that most work has focused on the static scenario with a small number of agents (see Milgrom and Roberts (1990)). Recently, there have been attempts to utilize the information structure of a mean field game (MFG) (Lasry and Lions, 2007; Tembine et al., 2009; Adlakha and Johari, 2013; Iyer et al., 2014; Li et al., 2016, 2018) to design algorithms to compute equilibria under a known model in the large population setting (Adlakha and Johari, 2013). Here, each agent assumes that the states of all others are drawn in an i.i.d. manner from a common mean field belief distribution, and optimizes accordingly. However, the model is non-stationary due to the change in the mean field distribution at each time step, and provably convergent learning algorithms for identifying equilibrium strategies are currently unavailable.

Main Contributions: We study the problem of learning under an unknown model in stochastic games with strategic complementarities under the mean field setting. Our main contributions are:

(i) We introduce the notion of trembling-hand perfection to the context of mean field games, under which a known randomness is introduced into all strategies. Unlike an ϵ-greedy policy in which randomization is added as an afterthought to the optimal action, under

*Equal Contribution

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

trembling-hand perfection, optimal value is computed consistently while accounting for this randomness.

(ii) We describe TQ-value iteration, based on generalized value iteration, as a means of computing trembling-hand consistent TQ-values. We introduce a notion of equilibrium that we refer to as trembling-hand-perfect mean field equilibrium (T-MFE), and show existence of T-MFE and a globally convergent computational method in games with strategic complementarities.

(iii) We propose two learning algorithms—model-free and model-based—for learning T-MFE under an unknown model. The algorithms follow a structure of identifying TQ-values to find a candidate strategy, and then taking only one-step Mckean-Vlasov dynamics to update the mean field distribution. We show convergence and determine their sample complexity bounds.

(iv) Finally, to the best of our knowledge, ours is the first work that develops a fully online learning scheme that utilizes the large population of agents to concurrently sample the real world. Our algorithm only needs a one-step Mckean-Vlasov mean field update under each strategy, which automatically happens via agents applying the current strategy for one step. This obviates the need for a multi-step simulator required by typical RL approaches.

Related Work: There has recently been much interest in the intersection of machine learning and collective behavior, often under the large population regime. Much of this work focuses on specific classes of stochastic games that possess verifiable properties on information, payoffs and preferences that provide structure to the problem. In line with this approach is learning in structured MFGs, such as in linear-quadratic, oscillator or potential game settings (Kizilkale and Caines, 2012; Yin et al., 2013; Cardaliaguet and Hadikhanloo, 2017; Carmona et al., 2019). Other approaches provide structure to the problem by considering localized effects, such as local interactions (Yang et al., 2018), local convergence (Mguni et al., 2018), or a local version of Nash equilibrium (Subramanian and Mahajan, 2019). While we are not aware of any work that considers learning in games with strategic complementarities, a survey on the algorithmic aspects of multi-agent RL, including a review of existing work in the mean field domain is available in Zhang et al. (2019).

Many of the issues faced in simultaneous learning and decision making in games with large populations are contained in Yang et al. (2018); Guo et al. (2019); Subramanian and Mahajan (2019), which are the closest to our work. Yang et al. (2018) considers the scenario wherein interactions are local in that each agent is impacted only by the set of neighbors, and so sampling only among them is sufficient to obtain a mean field estimate. This, however, requires a specific structure in which the Q-function of agents can be decomposed into such local versions. Subramanian and Mahajan (2019) proposed a simulator-based (not fully online) policy gradient algorithm for learning the mean field equilibrium in stationary mean field games. However, the convergence guarantees are limited to only a local version of Nash equilibrium. Finally, Guo et al. (2019) presents an existence and simulator-based model-free learning algorithm for MFE without structural assumptions on the game. Instead, there are contraction assumptions imposed on the trajectories of state and action distributions over time, which, however, may not be verifiable for a given game in a straightforward manner.

Our work is distinguished from existing approaches in several ways. We introduce the concept of a *strategically consistent* approach to learning via the trembling-hand idea, rather than arbitrarily adding a modicum of exploration to best responses as do most existing works. We provide a structured application scenario to apply this idea in the form of games with strategic complementarities. In turn, this allows us to explore both model-free and model-based methods to compute and learn optimal trembling-hand perfect MFE, including showing global convergence and determining their sample complexities. Perhaps most importantly, we exploit the large population setting to obtain samples without the aid of a simulator, which in turn enables learning directly from the real system.

2 Mean Field Games and Trembling-Hand Perfection

Mean Field Games: An N-agent stochastic dynamic game is represented as $(\mathcal{S}, \mathcal{A}, P, (r^i)_i^N, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, respectively, both assumed to be finite. At time k, agent i has state $s_k^i \in \mathcal{S}$, takes action $a_k^i \in \mathcal{A}$, and receives a reward $r^i(s_k, a_k)$. Here, $s_k = (s_k^i)_{i=1}^N$ is the system state and $a_k = (a_k^i)_{i=1}^N$ is the joint action. The system state evolves according to transition kernel $s_{k+1} \sim P(\cdot|s_k, a_k)$. Each agent aims at maximizing the infinite horizon cumulative discounted reward $\mathbb{E}[\sum_{k=0}^{\infty} \gamma^t r^i(s_k, a_k)]$, with discount factor $\gamma \in (0, 1)$.

Identification of a best response is computationally hard under a Bayesian framework, and a more realistic approach, aligned with a typical agent's computational capabilities is to reduce the information state of each agent to the so-called *mean field* state distribution (Lasry and Lions, 2007; Tembine et al., 2009; Adlakha and Johari, 2013; Iyer et al., 2014; Li et al., 2016, 2018). The mean field z_k at time k is defined as $z_k(s) = \frac{1}{N} \sum_{i=1}^N \mathbf{1}\{s_k^i = s\}$, and is the empirical distribution of the states of all agents. It represents

the agent's belief that the states of all others will be drawn in an i.i.d. manner from z_k. Agent i at time k receives a reward $r^i(s_k^i, a_k^i, z_k)$, and its state evolves according to $s_{k+1}^i \sim P(\cdot|s_k^i, a_k^i, z_k)$. The mean field approximation is accurate under structural assumptions on correlation decay across agent states as the number of agents becomes asymptotically large (Graham and Méléard, 1994; Iyer et al., 2014).

We represent a MFG as $\Gamma = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$, and restrict our attention to stationary MFGs with a homogeneous reward function. Here, all agents follow the same stationary strategy $\mu : \mathcal{S} \to \mathcal{P}(\mathcal{A})$, where $\mathcal{P}(\mathcal{A})$ is the probability distribution over the action space, and reward function $r^i = r$ for all i. We assume that $|r(s, a, z)| \leq 1$. The mean field z_k evolves following the discrete time McKean-Vlasov equation

$$z_{k+1} = \Phi(z_k, \mu), \text{ where, } z_{k+1}(s') = \Phi(z_k, \mu)(s') \quad (1)$$

$$= \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} z_k(s)\mu(s, a)P(s'|s, a, z_k). \quad (2)$$

The value function $V_{\mu,z}$ corresponding to the strategy μ and the mean field z is defined as $V_{\mu,z}(s) = \mathbb{E}[\sum_{k=0}^{\infty} \gamma^t r(s_k, a_k, z)|s_0 = s]$, with, $a_k \sim \mu(s_k, \cdot), s_{k+1} \sim P(\cdot|s_k, a_k, z)$.

Mean Field Games with Strategic Complementarities (MFG-SC): Structural assumptions on the nature of the MFG are needed in order to show existence of an equilibrium, and to identify convergent dynamics. Our focus is on a particular structure called *strategic complementarities* that aligns the increase of an agent's strategy with increases the competitors' strategies (Nowak, 2007; Vives, 2009; Adlakha and Johari, 2013).

We introduce some concepts before defining MFG-SC. The partially ordered set $(X, \succeq)$ is called a *lattice* if for all $x, y \in X$, the elements $\sup\{x, y\}$ and $\inf\{x, y\}$ are in X. X is a *complete lattice* if for any $S \subset X$, both $\sup S$ and $\inf S$ are in X. A function $f : X \to \mathbb{R}$ is said to be *supermodular* if $f(\sup\{x, x'\}) + f(\inf\{x, x'\}) \geq f(x) + f(x')$ for any $x, x' \in X$. Given lattices X, Y, a function $f : X \times Y \to \mathbb{R}$ is said to have *increasing differences* in x and y if for all $x' \succeq x, y' \succeq y$, $f(x', y') - f(x', y) \geq f(x, y') - f(x, y)$. A correspondence $T : X \to Y$ is *non-decreasing* if $x' \succeq x, y \in T(x)$, and $y' \in T(x')$ implies that $\sup\{y, y'\} \in T(x')$ and $\inf\{y, y'\} \in T(x)$.

For probability distributions $p, p' \in \mathcal{P}(X)$, we say p *stochastically dominates* p', denoted as $p \succeq_{SD} p'$, if $\sum_{x\in X} f(x)p(x) \geq \sum_{x\in X} f(x)p'(x)$ for any bounded non-decreasing function f. The conditional distribution $p(\cdot|y)$ is stochastically non-decreasing in y if for all $y' \succeq y$, we have $p(\cdot|y') \succeq_{SD} p(\cdot|y)$. Finally, the conditional distribution $p(\cdot|y, z)$ has *stochastically increasing differences* in y and z if $\sum_x f(x)p(x|y, z)$ has increasing differences in y and z for any bounded non-decreasing function f.

We now give the formal definition of mean field games with strategic complementarities (Adlakha and Johari, 2013).

Definition 1 (MFG-SC)**.** *Let Γ be a stationary mean field game. We say that Γ is a mean field game with strategic complementaries if it has the following properties:*
(i) Reward function: $r(s, a, z)$ is non-decreasing in s, supermodular in (s, a), and has increasing differences in (s, a) and z. Also, $\max_a r(s, a, z)$ is non-decreasing in s for all fixed z.
(ii) Transition probability: $P(\cdot|s, a, z)$ is stochastically supermodular in (s, a), has stochastically increasing differences in (s, a) and z, and is stochastically non-decreasing in each of s, a, and z.

MFG-SC Example - Infection Spread: We assume a large but fixed population of agents (computers or humans). At any time step, an agent may leave the system (network or town) with probability ζ, and is immediately replaced with a new agent. The state of an agent $s \in \mathbb{Z}^+$ is its health level, and the agent can take action $a \in \{1, 2, 3, ...|\mathcal{A}|\}$ (installing security patches, wearing a mask etc.) to stay healthy. The susceptibility of each agent, $p(s)$, is a decreasing function in state (higher health implies lower susceptibility). At each time step, the agent interacts with the ensemble of agents who have a mean field state distribution z. We define infection intensity $i_z = c_f p(\sum_{s\in\mathcal{S}} sz(s))$, which can be interpreted as the probability of getting infected via interaction with the population and c_f is the infection intensity constant.

Given the current state-action pair (s, a), the next state s' is given by

$$s' = (s + a - w_1)_+ \mathbb{1}\{E_1\} + (s + a)\mathbb{1}\{E_2\} + w_2\mathbb{1}\{E_3\},$$

where $(x)_+ = \max\{0, x\}$ (state is non-negative), $w_1, w_2 \in \mathbb{Z}^+$ are realizations of non-negative random variables, and E_is are mutually exclusive events with probabilities $i_z(1-\zeta), (1-i_z)(1-\zeta)$, and ζ, respectively. Events E_1 and E_2 correspond to the agent remaining in the system, and being infected (health may deteriorate) or not infected, respectively. E_3 is the event that the agent leaves and is replaced with an agent with random state (regeneration). An agent receives a reward that depends on its own immunity $1 - p(s)$ as well as that of the population (i.e., system value increases with immunity), but pays for its action. Hence,

$$r(s, a, z) = \delta_1(1 - p(s)) + \delta_2 \sum_{s\in\mathcal{S}} z(s)(1 - p(s)) - \delta_3 a,$$

where δ_is are positive constants.

It is easy to verify that the model has non-decreasing differences in the transition matrix. Also, it encourages crowd seeking behavior, where the mean field positively affects rewards obtained. Showing that it has strategic complementarities is straightforward. We validate our algorithms via simulations on this model in Section 6.

MFG-SC Example - Amazon Mechanical Turk (MTurk) and other Gig Economy Marketplaces: We consider Amazon Mechanical Turk (MTurk), a crowd sourcing market wherein human workers (called Turkers) are recruited to perform so-called Human Intelligence Tasks (HITs). There is a natural alignment of effort employed by Turkers in MTurk, since higher efforts translate into more HITs done right, which results in a higher quality of work distribution, which results in firms willing to spend more on HITs. Note that free riding is difficult, since poor quality work results in payments being withheld and reputation loss. This notion of incentive alignment applies to essentially all Gig economy marketplaces such as Uber and Airbnb—the reputation of the agent directly enhances its reward, while the reputation of the marketplace as a whole (i.e., its mean field) draws customers willing to pay into the system, and so enhances the reward of the agent. A more detailed description and numerical simulations are provided in the supplementary material.

Trembling-Hand-Perfect Mean Field Equilibrium: The notion of trembling-hand-perfection is a means of refining the Nash equilibrium concept to account for the fact that equilibria that naturally occur are often those that are optimal when a known amount of randomness is introduced into the strategies employed to ensure that they are totally mixed, i.e., all actions will be played with some (however small) probability (Bielefeld, 1988). Thus, the agent is restricted to only playing such mixed (randomized) strategies, but maintains strategic consistency in that it accounts for the probability of playing each action while calculating the expected payoff of such a totally mixed strategy.

Formalizing the above thoughts, we denote the set of trembling-hand strategies as Π^ϵ. A trembling-hand strategy $\mu \in \Pi^\epsilon$ is a mapping $\mu : \mathcal{S} \to \mathcal{P}^\epsilon(\mathcal{A})$, where $\mathcal{P}^\epsilon(\mathcal{A})$ is the set of ϵ-randomized probability vectors over $\mathcal{A}$. Any probability vector in $\mathcal{P}^\epsilon(\mathcal{A})$ has the value $(1-\epsilon)$ for one element and the value $\epsilon/(|\mathcal{A}|-1)$ for all the other elements. Thus, any $\mu \in \Pi^\epsilon$ has the following form: $\mu(s,a) = (1-\epsilon)$ for $a = a_s$ for some a_s and $\mu(s,a) = \epsilon/(|\mathcal{A}|-1)$ for all other $a \in \mathcal{A}$. In the standard reinforcement learning parlance, Π^ϵ is essentially the set of all ϵ-greedy policies.

The main difference between a trembling-hand strategy and an ϵ-greedy policy lies in the value function. Recall that under the ϵ-greedy idea, the agent computes the pure (deterministic) best response policy, and then arbitrarily adds randomization. However, under the strategic game setting, choosing a strategy that could result in an arbitrary loss of value is impermissible. Rather, the agent must compute the best trembling-hand strategy, i.e., it must account for the impact on value of the ϵ randomness. Formally, we first define the optimal trembling-hand value function V_z^* and the optimal trembling hand strategy μ_z^* corresponding to the mean field z as

$$V_z^* = \max_{\mu \in \Pi^\epsilon} V_{\mu,z}, \tag{3}$$

$$\mu_z^* \in \Psi(z), \quad \text{where} \quad \Psi(z) = \arg\max_{\mu \in \Pi^\epsilon} V_{\mu,z}. \tag{4}$$

We define trembling-hand-prefect mean field equilibrium (T-MFE) in terms of a trembling-hand strategy μ^* and a mean field distribution z^* that must jointly satisfy, (i) *optimality*—the strategy μ^* must be superior to all other strategies, given the belief z^*, and (ii) *consistency*—given a candidate mean field distribution z^*, the strategy μ^* must regenerate z^* under the McKean-Vlasov dynamics (1).

Definition 2 (T-MFE)**.** *Let Γ be a stationary mean field game. A trembling-hand-perfect mean field equilibrium (μ^*, z^*) of Γ is a strategy $\mu^* \in \Pi^\epsilon$ and a mean field z^* such that, (optimality condition) $\mu^* \in \Psi(z^*)$, and (consistency condition) $z^* = \Phi(z^*, \mu^*)$*

3 Existence and Computation of T-MFE

Existence of T-MFE: We first introduce a method to compute the optimal trembling-hand value function V_z^* for a given mean field z. Note that classical value iteration for any given finite MDP will converge under deterministic policies (pure strategies)—something that is not possible under our restriction to trembling hand strategies Π^ϵ, which only allows totally mixed strategies. We overcome this issue by using a generalized value iteration approach (Szepesvári and Littman, 1996). Here, rather than the value function, we compute the Q-value function, which for a strategy μ and a given mean field z is defined as $Q_{\mu,z}(s,a) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, z) | s_0 = s, a_0 = a]$, with, $a_t \sim \mu(s_t, \cdot), s_{t+1} \sim P(\cdot|s_t, a_t, z)$. The optimal trembling-hand Q-value function (TQ-value function) for a given mean field z is then defined as $Q_z^* = \max_{\mu \in \Pi^\epsilon} Q_{\mu,z}$. The optimal trembling-hand strategy μ_z^* for a given mean field z can then be computed as $\mu_z^* = \pi^\epsilon_{Q_z^*}$.

For any given Q-value function, define the trembling-

hand strategy π^{ϵ}_{Q} and the function $G(Q)$ as

$$\pi^{\epsilon}_{Q}(s,a) = \begin{cases} (1-\epsilon) & \text{for } a = \arg\max_b Q(s,b) \\ \epsilon/(|\mathcal{A}|-1) & \text{for } a \neq \arg\max_b Q(s,b) \end{cases},$$

$$G(Q)(s) = \sum_{a\in\mathcal{A}} \pi^{\epsilon}_{Q}(s,a)\ Q(s,a).$$

Note that π^{ϵ}_{Q} is the usual ϵ-greedy policy with respect to Q. Using the above notation, we define the TQ-value operator F_z for a given mean field z as

$$F_z(Q)(s,a) = r(s,a,z) + \gamma \sum_{s'} P(s'|s,a,z)\ G(Q)(s'). \tag{5}$$

The TQ-value operator F_z has properties similar to the standard Bellman operator. In particular, we show below that F_z is a contraction with Q^*_z as its unique fixed point.

Proposition 1. *(i) F_z is a contraction mapping in sup norm for all $z \in \mathcal{P}(\mathcal{S})$. More precisely, $\|F_z(Q_1) - F_z(Q_2)\|_\infty \leq \gamma\|Q_1 - Q_2\|_\infty$ for any Q_1, Q_2, and for all $z \in \mathcal{P}(\mathcal{S})$.*
(ii) The optimal trembling-hand Q-value function Q^*_z for a given mean field z is the unique fixed point of F_z, i.e., $F_z(Q^*_z) = Q^*_z$.

The contraction property of F_z implies that the iteration $Q_{m+1,z} = F_z(Q_{m,z})$ will converge to the unique fixed point of F_z, i.e., $Q_{m,z} \to Q^*_z$. We call this procedure as TQ-value iteration.

From the above result, we can compute the optimal trembling-hand strategy for a given mean field z. However, it is not clear if there exists a T-MFE (z^*, μ^*) that simultaneously satisfies the optimality condition and consistency condition. We answer this question affirmatively below.

Theorem 1. *Let Γ be a stationary mean field game with strategic complementarities. Then, there exists a trembling-hand-perfect mean field equilibrium for Γ.*

The proof follows from the monotonicity properties of MFG-SC. Thus, given this game structure, no additional conditions are needed to show existence.

Computing T-MFE: Given that a T-MFE exists, the next goal is to devise an algorithm to compute a T-MFE. A natural approach is to use a form of best-response dynamics as follows. Given the candidate mean field z_k, the trembling best-response strategy μ_k can be computed as $\mu_k \in \Psi(z_k)$. While the typical approach would be to then compute the stationary distribution under μ_k, we simply update the next mean field z_{k+1} by using just one-step Mckean-Vlasov dynamics as $z_{k+1} = \Phi(z_k, \mu_k)$, and the cycle continues. While the approach is intuitive and reminiscent of the best-response dynamics proposed in Adlakha and Johari (2013), it is not clear that it will converge to any equilibrium. We show that in mean field games with strategic complementarities, such a trembling best-response (T-BR) process converges to a T-MFE. Our computation algorithm, which we call the T-BR algorithm, is presented in Algorithm 1.

Algorithm 1: T-BR Algorithm

1: Initialization: Initial mean field z_0
2: **for** $k = 0, 1, 2, ..$ **do**
3: For the mean field z_k, compute the optimal TQ-Value function $Q^*_{z_k}$ using the TQ-value iteration $Q_{m+1,z_k} = F(Q_{m,z_k})$
4: Compute the strategy $\mu_k = \Psi(z_k)$ as the trembling-hand strategy w.r.t $Q^*_{z_k}$, i.e, $\mu_k = \pi^{\epsilon}_{Q^*_{z_k}}$
5: Compute the next mean field $z_{k+1} = \Phi(z_k, \mu_k)$
6: **end for**

Theorem 2. *Let Γ be a stationary mean field game with strategic complementarities. Let $\{z_k\}$ and $\{\mu_k\}$ be the sequences of mean fields and strategies generated according to Algorithm 1. Then $\mu_k \to \mu^*$ and $z_k \to z^*$ as $k \to \infty$ where (μ^*, z^*) constitutes a T-MFE of Γ.*

Remark 1. *Consider the sequence of mean fields $\{z_k\}$ generated by the T-BR algorithm. According to Theorem 2, there exists a finite $k_0 = k_0(\bar{\epsilon})$ such that $\|z_k - z^*\| \leq \bar{\epsilon}$ for all $k \geq k_0$, where z^* is the T-MFE. A precise characterization of k_0 is difficult because the convergence of the T-BR algorithm is based on monotonicity properties of MFG-SC, rather than on contraction arguments.*

4 Learning T-MFE

We address the problem of learning T-MFE when the model is unknown. In this section, we assume the availability of a simulator, which, given the current state s, current action a and current mean field z, can generate the next state $s' \sim P(\cdot|s,a,z)$. We discuss how to learn directly from real-world samples without a simulator in the next section. We also assume that the reward function is known, as is common in the literature. We now introduce two reinforcement learning algorithms—a model-free algorithm and a model-based algorithm—for learning T-MFE.

4.1 Model-Free TMFQ-Learning Algorithm for learning T-MFE

We first describe the TMFQ-learning algorithm, which builds on the T-BR algorithm. Recall that the T-BR algorithm uses knowledge of the model in two locations.

The first is that at each step k, for a given mean field z_k, the optimal TQ-value function $Q^*_{z_k}$ is computed using TQ-value iteration. In the learning approach, we use the generalized Q-learning framework (Szepesvári and Littman, 1996) as the basis for the model-free TQ-learning algorithm as follows:

$$\begin{aligned} Q_{t+1,z_k}(s_t,a_t) &= (1-\alpha_t)Q_{t,z_k}(s_t,a_t) \\ &\quad + \alpha_t(r(s_t,a_t,z_k) + \gamma G(Q_{t,z_k})(s_{t+1})) \end{aligned} \tag{6}$$

where α_t is the appropriate learning rate. Here, the state sequence $\{s_t\}$ is generated using the simulator by fixing the mean field z_k, i.e., $s_{t+1} \sim P(\cdot|s_t,a_t,z_k), \forall t$. Using the properties of the generalized Q-learning formulation (Szepesvári and Littman, 1996), it can be shown that $Q_{t,z_k} \to Q^*_{z_k}$ as $t \to \infty$.

The second location where the model is needed in T-BR is for the one-step McKean-Vlasov update. In the learning approach, given the mean field z_k and the strategy μ_k, the next mean field z_{k+1} can be estimated to a desired accuracy using the simulator. The precise numerical approach is presented as the Next-MF scheme described in Algorithm 4 (supplementary material). We can now combine these steps to obtain the TMFQ-learning algorithm presented in Algorithm 2.

Algorithm 2: TMFQ-Learning Algorithm for T-MFE

```
Initialization: Initial mean field z_0
for k = 0, 1, 2, ... do
    Initialize time step t ← 0. Initialize s_0, Q_{0,z_k}
    repeat
        Take action a_t ~ π^ε_{Q_{t,z_k}}(s_t), observe reward
        and the next state s_{t+1} ~ P(·|s_t, a_t, z_k)
        Update Q_{t,z_k} according to TQ-learning (6)
        t ← t + 1
    until ||Q_{t,z_k} − Q_{t−1,z_k}||_∞ < ε_1
    Let Q_{z_k} = Q_{t,z_k} and the strategy μ_k = π^ε_{Q_{z_k}}
    z_{k+1} =Next-MF(z_k, μ_k)
end for
```

We first present an asymptotic convergence result of TMFQ-learning based on a perfect accuracy assumption on the TQ-learning and Next-MF steps, i.e., they are run to convergence. A complex, but more accurate analysis using two-timescale stochastic approximation is also possible. Instead, we will remove this assumption, and provide a PAC-type result further below.

Theorem 3. *Let Γ be a stationary mean field game with strategic complementarities. Let $\{z_k\}$ and $\{\mu_k\}$ be the sequences of the mean fields and policies generated by Algorithm 2. Then $\mu_k \to \mu^*$ and $z_k \to z^*$ as $k \to \infty$ where (μ^*, z^*) is a T-MFE of Γ.*

In a practical implementation, we may only run the TQ-learning step and the Next-MF step for a finite number of iterations. Hence, we develop a sample complexity bound under which TQ-learning and Next-MF provide an appropriate accuracy. We desire to compare TMFQ-learning with T-BR after k_0 iterations, where k_0 is the number of iterations of T-BR that yields a mean field that is $\bar{\epsilon}$-close to the T-MFE z^*. We make some necessary assumptions that are required for such a characterization.

Assumption 1. *(i) There exists $C_1 > 0$ such that $\|r(\cdot,\cdot,z) - r(\cdot,\cdot,\tilde{z})\|_1 \leq C_1 \|z - \tilde{z}\|_1$, for all $z, \tilde{z} \in \mathcal{P}(\mathcal{S})$.*
(ii) There exists a $C_2 > 0$ such that $\|P(\cdot|\cdot,\cdot,z) - P(\cdot|\cdot,\cdot,\tilde{z})\|_1 \leq C_2 \|z - \tilde{z}\|_1$, for all $z, \tilde{z} \in \mathcal{P}(\mathcal{S})$.
(iii) Let $P_{\mu,z}(s'|s) = \sum_a \mu(s,a)P(s'|s,a,z)$. Let μ_1, μ_2 be the trembling-hand policies corresponding to Q_1, Q_2, i.e., $\mu_1 = \pi^\epsilon_{Q_1}, \mu_2 = \pi^\epsilon_{Q_2}$. Then there exists a $C_3 > 0$ such that $\|P_{z,\mu_1} - P_{z,\mu_2}\|_1 \leq C_3 \|Q_1 - Q_2\|_\infty$ for all $z \in \mathcal{P}(\mathcal{S})$ and for any given Q_1, Q_2.

Assumption 1.(i) and 1.(ii) indicate that the reward function and transition kernel are Lipschitz with respect to the mean field, while Assumption 1.(iii) indicates that the distance between the Markov chains induced by two policies on the same transition kernel are upper bounded by a constant times the distance between their respective Q-functions. We then have the following.

Theorem 4. *Let Assumption 1 hold. For any $0 \leq \bar{\epsilon}, \bar{\delta} < 1$, let $k_0 = k_0(\bar{\epsilon})$. In Algorithm 2, for each $k \leq k_0$, assume that TQ-learning (according to (6)) update is performed T_0 number of times where T_0 is given as*

$$\begin{aligned} T_0 = O\Bigg(\Bigg(\frac{B^2 L^{1+3w} V_{\max}^2}{\beta^2 \bar{\epsilon}^2} \ln\Big(\frac{2Bk_0|\mathcal{S}||\mathcal{A}|V_{\max}}{\bar{\delta}\beta\bar{\epsilon}}\Big) \Bigg)^{\frac{1}{w}} \\ + \left(\frac{L}{\beta} \ln\Big(\frac{BV_{\max}}{\bar{\epsilon}}\Big) \right)^{\frac{1}{1-w}} \Bigg), \end{aligned} \tag{7}$$

where $B = (1 + C_2 + C_3 D)^{k_0+1}(C_3 + 1)$, $D = (C_1 + \gamma C_2)/(1-\gamma)$, $V_{\max} = 1/(1-\gamma), \beta = (1-\gamma)/2$, L is an upper bound on the covering time [1], and $w \in (1/2, 1)$. Then,

$$\mathbb{P}(\|z_{k_0} - z^*\| \leq 2\bar{\epsilon}) \geq (1 - \bar{\delta}).$$

We may also eliminate the dependence of the constant term B on k_0 under a contraction assumption on the McKean-Vlasov dynamics Φ (for instance, following conditions similar to Borkar and Sundaresan (2013)).

[1] Covering time of a state-action pair sequence is the number of steps needed to visit all state-action pairs starting from any arbitrary state-action pair Even-Dar and Mansour (2003).

Assumption 2. *Let Q_1, Q_2 be two arbitrary Q-value functions and let $\mu_1 = \pi^{\epsilon}_{Q_1}, \mu_2 = \pi^{\epsilon}_{Q_2}$. Let z_1, z_2 be two arbitrary mean fields. Then there exists positive constants C_4 and C_5 such that $\|\Phi(z_1, \mu_1) - \Phi(z_2, \mu_2)\|_1 \leq C_4\|z_1 - z_2\|_1 + C_5\|Q_1 - Q_2\|_\infty$. Also assume that $(C_4 + C_5 D) < 1$, where $D = (C_1 + \gamma C_2)/(1-\gamma)$.*

Corollary 1. *Let Assumption 1 and Assumption 2 hold. Then, the we obtain the bound on T_0 as in (7) with $B = (C_5 + 1)/(1 - (C_4 + C_5 D))$, which does not depend on k_0.*

4.2 Generative Model-Based Reinforcement Learning for T-MFE

A model-based variant of the T-BR algorithm is also straightforward to construct. We note that non-stationarity in our system (the model changes at each step) implies that there is no single model. Hence, our approach follows the generation of a new model each time that the mean field evolves under a T-BR-like procedure. Full details are presented in Appendix B.

5 Online Learning of T-MFE

The availability of a large number of agents that explore via trembling hand strategies suggests that we can do away with a system simulator via an online algorithm that simply aggregates these concurrently generated samples and computes a new strategy that is then pushed to all agents. Typically, the assumption in such large population scenarios is that the system size is fixed, but any agent may leave the system at time step t with probability $\zeta \in (0, 1)$, and be immediately replaced by a new agent with random state referred to as a *regeneration event.*

A generic RL approach would require that given the current strategy μ_k and mean field z_k, we would need to compute the stationary distribution of the model P_{z_k, μ_k}, and set it as the next mean field. This would preclude learning without a simulator, since running one step in the real world would immediately cause a mean field update to z_{k+1}, and induce a new model $P_{z_{k+1},.}$, making the system non-stationary.

Since our RL algorithms only need a one-step McKean-Vlasov update under each strategy, an online learning approach at time k when the underlying state distribution is z_k, would be to apply μ_{k-1} to the system, with the resultant state distribution being z_{k+1}. However, we face the issue that the samples obtained pertain to $P_{z_k, \mu_{k-1}}$, whereas the system model is now $P_{z_{k+1},.}$, and so an online sample-based trembling-hand strategy μ_k will lag the current system model by one step. Fortunately, convergence of the TMFQ-learning approach is robust to this lag, and we present its online version in Algorithm 3.

Algorithm 3: Online TMFQ-learning Algorithm for T-MFE

```
Initialize mean field z_0 and strategy μ_0
for k = 1, 2, ... do
    Reset memory buffer
    for agents i = 1, 2, ..., I do
        Given current state s_k^i, take action
        a_k^i ~ μ_{k-1}(s_k^i), get reward r_k^i, observe the
        next state s_{k+1}^i
        Add the sample (s_k^i, a_k^i, r_k^i, s_{k+1}^i) to the
        memory buffer
    end for
    Perform TQ-learning on the memory buffer to
    obtain Q_{z_k} and μ_k = π^ε_{Q_{z_k}}
end for
```

Algorithm 3 aggregates the samples generated by executing strategy μ_{k-1} to estimate the TQ-value function, and then passes back μ_k to the agents. We note that regeneration of a fraction of agents and execution of trembling hand strategies ensures that we have sufficient samples of each state-action pair in the large population regime to ensure that off-policy learning such as TQ-learning on the *memory buffer* converges to the optimal TQ-value function. In order to characterize the sample complexity of Algorithm 3, we employ bounds pertaining to synchronous Q-learning (Even-Dar and Mansour, 2003) using this guarantee that all state-action pairs are sampled a desired number of times. The PAC result is similar to TMFQ-learning, with accuracy increasing in the number of agents, rather than with the number of samples as in Theorem 4.

Theorem 5. *Let Assumption 1 hold. For any $0 \leq \bar{\epsilon}, \bar{\delta} < 1$, let $k_0 = k_0(\bar{\epsilon})$. In Algorithm 3, for each $k \leq k_0$, assume that there are a total of $I = I_0$ number of agents where I_0 is given as*

$$I_0 = O\left(\frac{|S||A|}{\epsilon\zeta}\left(\frac{B^2 V_{\max}^2}{\beta^2\bar{\epsilon}^2}\ln\left(\frac{2Bk_0|\mathcal{S}||\mathcal{A}|V_{\max}}{\bar{\delta}\beta\bar{\epsilon}}\right)\right)^{\frac{1}{w}} + \left(\frac{1}{\beta}\ln\left(\frac{BV_{\max}}{\bar{\epsilon}}\right)\right)^{\frac{1}{1-w}}\right),$$

where $B = \frac{((1-\zeta)2A)^{k_0}(C_3\epsilon)}{(1-\zeta)(A-1)}$, $A = \max\{1 + C_2, C_3 D\}$, $D = (C_1 + \gamma C_2)/(1-\gamma)$, $V_{\max} = 1/(1-\gamma)$, $\beta = (1-\gamma)/2$, $w \in (1/2, 1)$, ζ is a regeneration probability and ϵ is the trembling-hand strategy randomization. Then,

$$\mathbb{P}(\|z_{k_0} - z^*\| \leq 2\bar{\epsilon}) \geq (1 - \bar{\delta}).$$

Note that a result similar to that of Corollary 1 can easily be shown here as well under the same assumption. We omit details due to page limitations.

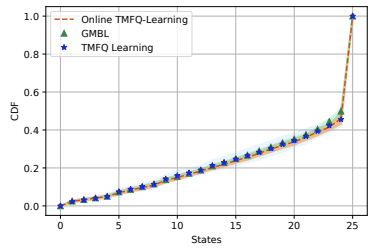

Figure 1: TMFE with $c_f = 0.1$

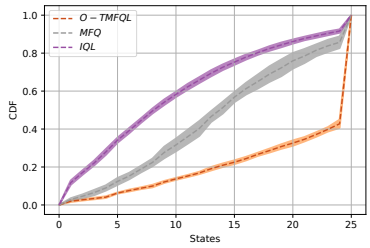

Figure 2: CDF of states for different algorithms: $c_f = 0.05$

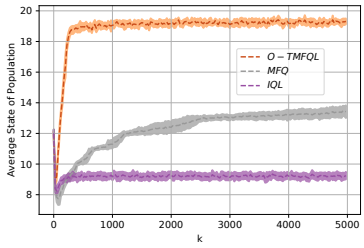

Figure 3: Mean states for different algorithms: $c_f = 0.05$

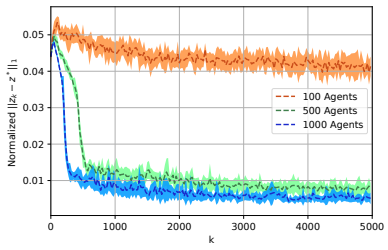

Figure 4: Convergence of O-TMFQ-Learning: $c_f = 0.05$.

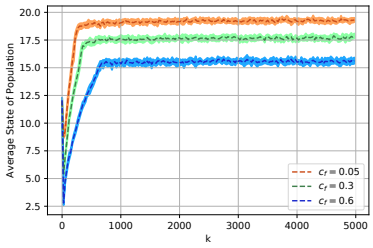

Figure 5: Mean state of O-TMFQ-Learning over time

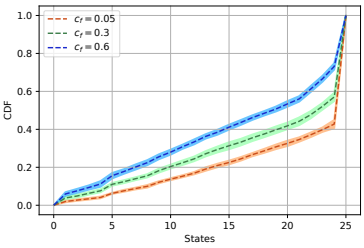

Figure 6: CDF of O-TMFQ-Learning

6 Experiments

We consider *Infection Spread* described in Section 2. State space is $\mathcal{S} = \{0, \ldots, 24\}$ with 0 being the lowest health level. Action space is $\mathcal{A} = \{0, \ldots, 4\}$ with 4 being the strongest preventive action, and c_f is the infection intensity constant. Details of the parameters and more experiments are presented in the supplementary materials. We evaluate the performance of our algorithms, TMFQ-Learning (TMFQ), GMBL, and Online TMFQ-Learning (O-TMFQ) on this model, along with comparisons with Independent Q-Learning (IQL) and Mean Field Q-learning (MFQ) (Yang et al., 2018). In IQL, each agent ignores other agents, maintains an individual Q-function and performs TQ-learning independently. We implement a variant of MFQ, where each agent maintains a Q-function parameterized by the average states of a subset of the population and performs TQ-Learning. We average over 20 runs in each experiment, and the dashed line and band in figures show the average and standard deviation, respectively.

Figure 1 shows the final mean field distribution obtained by each of our algorithms, simulated with 1000 agents. Note that all of them converge to the same T-MFE, indicating the accuracy of O-TMFQ. We next compare the performance O-TMFQ with IQL and MFQ. Figure 2 shows that the final equilibrium distributions of IQL and MFQ are inaccurate (not the true T-MFE), and that O-TMFQ results in higher states (health levels). Figure 3 shows the evolution of the mean health of the population, $\sum_s s z_k(s)$ with iteration number k. The mean health of the population quickly converges to a higher value under O-TMFQ while the corresponding value is lower under MFQ and IQL.

Figure 4 shows rate of convergence of the mean field under different numbers of agents. Here, z^* is the final mean field obtained by O-TMFQ. We see that the asymptotically accurate mean field approximation becomes increasingly correct even with a relatively small number of agents of 500 or 1000. We show the evolution of average health level of the population for O-TMFQ in Figure 5 for different values of c_f. This plot indicates that the convergence of the algorithm is fairly fast. Finally, Figure 6 explores the impact of the mean field on the model and its associated equilibrium via c_f. As expected, more agents are in lower health states for larger c_f.

7 Conclusions

We introduced the notion of trembling-hand perfection to MFG as a means of providing strategically consistent exploration. We showed existence of T-MFE in MFG with strategic complementarities, and developed an algorithm for computation. Based on this algorithm, we developed model-free, model-based and fully online learning algorithms, and provided PAC bounds on their performance. Experiments illustrated the accuracy and good convergence properties of our algorithms.

Acknowledgement

This work was supported in part by NSF grants CRII-CPS-1850206, ECCS-EPCN-1839616, CNS-1955696, ECCS-1839816, CNS-1719384, CPS-2038963 and ARO grant W911NF-19-1-0367.

References

S. Adlakha and R. Johari. Mean field equilibrium in dynamic games with strategic complementarities. *Operations Research*, 61(4):971–989, 2013.

K. Asadi and M. L. Littman. An alternative softmax operator for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 243–252. JMLR. org, 2017.

R. S. Bielefeld. Reexamination of the perfectness concept for equilibrium points in extensive games. In *Models of Strategic Rationality*, pages 1–31. Springer, 1988.

V. S. Borkar and R. Sundaresan. Asymptotics of the invariant measure in mean field models with jumps. *Stochastic Systems*, 2(2):322–380, 2013.

P. Cardaliaguet and S. Hadikhanloo. Learning in mean field games: the fictitious play. *ESAIM: Control, Optimisation and Calculus of Variations*, 23(2):569–591, 2017.

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

E. Even-Dar and Y. Mansour. Learning rates for Q-learning. *Journal of machine learning Research*, 5 (Dec):1–25, 2003.

C. Graham and S. Méléard. Chaos hypothesis for a system interacting through shared resources. *Probability Theory and Related Fields*, 100(2):157–174, 1994.

X. Guo, A. Hu, R. Xu, and J. Zhang. Learning mean-field games. In *Advances in Neural Information Processing Systems*, pages 4967–4977, 2019.

K. Iyer, R. Johari, and M. Sundararajan. Mean field equilibria of dynamic auctions with learning. *Management Science*, 60(12):2949–2970, 2014.

A. C. Kizilkale and P. E. Caines. Mean field stochastic adaptive control. *IEEE Transactions on Automatic Control*, 58(4):905–920, 2012.

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

J. Li, R. Bhattacharyya, S. Paul, S. Shakkottai, and V. Subramanian. Incentivizing sharing in realtime D2D streaming networks: A mean field game perspective. *IEEE/ACM Transactions on Networking*, 25(1):3–17, 2016.

J. Li, B. Xia, X. Geng, H. Ming, S. Shakkottai, V. Subramanian, and L. Xie. Mean field games in nudge systems for societal networks. *ACM Transactions on Modeling and Performance Evaluation of Computing Systems (TOMPECS)*, 3(4):15, 2018.

D. Mguni, J. Jennings, and E. M. de Cote. Decentralised learning in systems with many, many strategic agents. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

P. Milgrom and J. Roberts. Rationalizability, learning, and equilibrium in games with strategic complementarities. *Econometrica: Journal of the Econometric Society*, pages 1255–1277, 1990.

A. S. Nowak. On stochastic games in economics. *Mathematical Methods of Operations Research*, 66(3):513–530, 2007.

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

C. Szepesvári and M. L. Littman. Generalized Markov decision processes: Dynamic-programming and reinforcement-learning algorithms. In *Proceedings of International Conference of Machine Learning*, volume 96, 1996.

A. Tarski et al. A lattice-theoretical fixpoint theorem and its applications. *Pacific journal of Mathematics*, 5(2):285–309, 1955.

H. Tembine, J.-Y. Le Boudec, R. El-Azouzi, and E. Altman. Mean field asymptotics of Markov decision evolutionary games and teams. In *2009 International Conference on Game Theory for Networks*, pages 140–150. IEEE, 2009.

X. Vives. Strategic complementarity in multi-stage games. *Economic Theory*, 40(1):151–171, 2009.

T. Weissman, E. Ordentlich, G. Seroussi, S. Verdu, and M. J. Weinberger. Inequalities for the l1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, 2003.

Y. Yang, R. Luo, M. Li, M. Zhou, W. Zhang, and J. Wang. Mean field multi-agent reinforcement learning. In *35th International Conference on Machine Learning, ICML 2018*, volume 80, pages 5571–5580. PMLR, 2018.

H. Yin, P. G. Mehta, S. P. Meyn, and U. V. Shanbhag. Learning in mean-field games. *IEEE Transactions on Automatic Control*, 59(3):629–644, 2013.

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

A Algorithms Description

A.1 Next-MF Function for One-Step McKean-Vlasov Update

Algorithm 4: Next-MF

1: **Input:** Mean field z and strategy μ
2: Initialize $\hat{z}$, sample number $j \leftarrow 0$
3: **repeat**
4: Sample the state $s \sim z(\cdot)$, action $a \in \mu(s,\cdot)$, next state $s' \sim P(\cdot|s,a,z)$
5: $\hat{z}(s') \leftarrow \hat{z}(s') + 1$
6: $\bar{z}_j = \texttt{normalize}(\hat{z})$
7: $j \leftarrow j + 1$
8: **until** $\|\bar{z}_j - \bar{z}_{j-1}\|_1 \leq \epsilon_2$
9: $\bar{z} = \bar{z}_j$
10: **return** $\bar{z}$

Algorithm 4 estimates the next mean field according to equation (1). We maintain a frequency estimate of s' in line 5 and normalize the frequency estimate to obtain a density function in line 6. Note that s' is sampled according to line 4.

B Generative Model-Based Reinforcement Learning for T-MFE

In this section we present a model-based variant of the T-BR algorithm. Non-stationarity in our system (model changes at each step) implies that there is no single model that can be learned. Our approach follows the generation of a new model each time that the mean field evolves under a T-BR like approach.

At each iteration, we first estimate the model $P(\cdot|\cdot,\cdot,z_k)$ for the given z_k using n_0 simulator samples for each (s,a). We next define the approximate TQ-value operator $\widehat{F}_{z_k}$ as in (5) by replacing the actual model P with the estimated model $\widehat{P}$. It is straightforward to show that $\widehat{F}_{z_k}$ is also a contraction. This ensures that approximate TQ-value iteration will converge to an approximate TQ-value function $\widehat{Q}^*_{z_k}$, which will, of course, have an error with respect to the true TQ-value function $Q^*_{z_k}$. We determine the trembling-hand best response strategy with respect to $\widehat{Q}^*_{z_k}$. Finally, the next mean field z_{k+1} is obtained using this strategy and the estimated model in the McKean-Vlasov update equation, denoted by $\hat{\Phi}(\cdot,\cdot)$. The GMBL algorithm is summarized in Algorithm 5.

Note that we are not estimating the model for all the possible mean fields, but only for the sequence $\{z_k\}$. So, if the process converges in a finite number of steps, then we need only a finite number of simulation samples. We show that this is indeed true in the theorem below.

Theorem 6. *Let Assumption 1 hold. For any $0 \leq \bar{\epsilon}, \bar{\delta} < 1$, let $k_0 = k_0(\bar{\epsilon})$. In Algorithm 5, for each $k \leq k_0$, assume that the estimate $\widehat{P}(\cdot|\cdot,\cdot,z_k)$ is obtained by a total of $N_0 = n_0|S||A|$ simulator samples where n_0 is given as*

$$n_0 = O\bigg(\max\bigg(\frac{2V_{\max}^4 B^2}{\bar{\epsilon}^2}\log\bigg(\frac{2|\mathcal{S}||\mathcal{A}|k_0}{\bar{\delta}}\bigg), \frac{2B^2}{\bar{\epsilon}^2}\log(\frac{2^{|\mathcal{S}|}|\mathcal{S}||\mathcal{A}|k_0}{\bar{\delta}}\bigg)\bigg)\bigg)$$

where $B = (1 + C_2 + C_3 D)^{k_0+1}$, $D = (C_1 + \gamma C_2)/(1-\gamma)$, $V_{\max} = 1/(1-\gamma)$. Then,

$$\mathbb{P}(\|z_{k_0} - z^*\| \leq 2\bar{\epsilon}) \geq (1 - \bar{\delta}).$$

A result similar to that of Corollary 1 can easily be shown here as well under the same assumption. The proof of

theorem 6 is given in section D

Algorithm 5: GMBL Algorithm for T-MFE

1: Initialization: Initial mean field z_0
2: **for** $k = 0, 1, 2, ..$ **do**
3: For the mean field z_k estimate the model to get $\widehat{P}(\cdot|\cdot,\cdot,z_k)$ by taking n_0 next-state samples for each state-action pair (s,a)
4: Compute $\widehat{Q}^*_{z_k}$ using the approximate TQ-value iteration $\widehat{Q}_{m+1,z_k} = \widehat{F}_{z_k}(\widehat{Q}_{m,z_k})$
5: Compute the strategy $\mu_k = \pi^{\epsilon}_{\widehat{Q}^*_{z_k}}$. Then compute the next mean field $z_{k+1} = \widehat{\Phi}(z_k, \mu_k)$
6: **end for**

C Proofs of Results in Section 3

We use the following result from Asadi and Littman (2017)

Lemma 1 (Asadi and Littman (2017))**.** *For any Q_1, Q_2, and for any $s \in \mathcal{S}$,*

$$|G(Q_1)(s) - G(Q_2)(s)| \leq \max_a |Q_1(s,a) - Q_2(s,a)|.$$

C.1 Proof of Proposition 1

Proof of Proposition 1. For any given $(s,a) \in \mathcal{S} \times \mathcal{A}$ and mean field z,

$$|F_z(Q_1)(s,a) - F_z(Q_2)(s,a)| \leq \gamma |\sum_{s'} P(s'|s,a,z)(G(Q_1)(s') - G(Q_2)(s'))|$$

$$\overset{(a)}{\leq} \gamma \sum_{s'} P(s'|s,a,z) \max_b |Q_1(s',b) - Q_2(s',b)| \leq \gamma \|Q_1 - Q_2\|_\infty$$

where (a) follows from Lemma 1. Since $(s,a) \in \mathcal{S} \times \mathcal{A}$ was arbitrary, we have $\|F_z(Q_1) - F_z(Q_2)\|_\infty \leq \gamma \|Q_1 - Q_2\|_\infty$. Existence of a unique fixed point for F_z follows directly from the Banach's fixed point theorem since F_z is a contraction. The claim that this unique fixed point is equal to Q^*_z follows from the Bellman optimality principle. □

C.2 Proof of Theorem 1 and Theorem 2

We follow a proof approach that uses the strategic complementarity conditions to establish some monotone properties, and exploit that to show the existence of T-MFE and the convergence of T-BR algorithm. We first state some useful results from Adlakha and Johari (2013). Note that, however, since we are considering trembling-hand polices, proofs in Adlakha and Johari (2013) are not directly applicable to our setting.

Lemma 2 (Lemma 4 in Adlakha and Johari (2013))**.** *Suppose that $V_z(s)$ is a non-decreasing bounded function in s and has increasing differences in s and z. Then, $\sum_{s' \in S} P(s'|s,a,z)V_z(s')$ is non-decreasing in s and a and has increasing differences in (s,a) and z. Moreover, the function $V'_z(s)$ defined as, $V'_z(s) = \max_a(r(s,a,z) + \gamma \sum_{s'} P(s'|s,a,z)V_z(s))$, is non-decreasing in s and has increasing differences in s and z.*

Lemma 3 (Lemma 6 in Adlakha and Johari (2013))**.** *Suppose that $V_z(s)$ is non-decreasing in s and has increasing differences in s and z. Define a correspondence*

$$\Omega(s,z) = \arg\max_{a \in \mathcal{A}} \ (r(s,a,z) + \gamma \sum_{s' \in S} P(s'|s,a,z)V_z(s')).$$

Then, Ω is a non-decreasing correspondence in (s,z).

Proposition 2. *Let V^*_z and μ^*_z be the optimal trembling-hand value function and optimal trembling-hand strategy corresponding to mean field z. Then, $V^*_z(s)$ is non-decreasing in s and has increasing differences in s and z. Moreover, μ^*_z is stochastically non-decreasing in s and z.*

Proof. Let $\{Q_{m,z}, m \geq 0\}$ be the TQ-value iterates corresponding to a mean field z. If $\arg\max_b Q_{m,z}(s,b)$ is not unique, then $\pi^{\epsilon}_{Q_{m,z}}$ can be defined in more than one way. We define it as

$$\pi^{\epsilon}_{Q_{m,z}}(s,a) = \begin{cases} 1-\epsilon & \text{for } a = \sup\{\arg\max_b Q_{m,z}(s,b)\} \\ \epsilon/(|\mathcal{A}|-1) & \text{for } a \neq \sup\{\arg\max_b Q_{m,z}(s,b)\} \end{cases} \tag{8}$$

Note that the sup of a set is well defined with respect to a lattice according to the definition of MFG-SC. In the following we denote $\pi^{\epsilon}_{Q_{m,z}}$ simply as $\mu_{m,z}$.

By definition, $V^*_z = G(Q^*_z)$ and define $V_{m,z} = G(Q_{m,z})$. To show that $V^*_z(s)$ is non-decreasing in s and has increasing differences in s and z, it suffices to show $V_{m,z}(s)$ has the same properties for all m. This is because, since G is continuous, monotonicity and increasing differences are preserved under limits.

Letting $Q_{0,z}(s,a) = 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, we have $Q_{1,z}(s,a) = r(s,a,z)$. Then, $Q_{1,z}(s,a)$ is non-decreasing in s and has increasing differences in s and z by Definition 1. Define the correspondence Ω_m as

$$\Omega_m(s,z) = \arg\max_{a \in \mathcal{A}} \ (r(s,a,z) + \gamma \sum_{s' \in S} P(s'|s,a,z) V_{m-1,z}(s')).$$

By Lemma 3, $\Omega_1(s,z) = \arg\max Q_{1,z}(s,a)$ is non-decreasing in s and z. Using this, we can conclude that $\mu_{1,z} = \pi^{\epsilon}_{Q_{1,z}}$ is stochastically non-decreasing in s and z for ϵ such that $(1-\epsilon) > \epsilon/(|\mathcal{A}|-1)$. To see this, first note that for $\epsilon = 0$, the deterministic strategy $\pi^{\epsilon}_{Q_{1,z}}$ is non-decreasing in s and z by Lemma 3. Now, for ϵ with $(1-\epsilon) > \epsilon/(|\mathcal{A}|-1)$, $\pi^{\epsilon}_{Q_{1,z}}$ is stochastically non-decreasing in s and z because the probability of the maximizing action is greater than all other actions. Thus, $\mu_{m,z}$ is stochastically non-decreasing in s and z for all $m \in \mathbb{N}$ if each $Q_{m,z}$ is non-decreasing in s and has increasing differences in s and z.

Now, recall that $V_{1,z}(s) = \sum_a \mu_{1,z}(s,a) Q_{1,z}(s,a) = (1-\epsilon) Q_{1,z}(s,\bar{a}) + \epsilon \sum_a Q_{1,z}(s,a)$ where $\bar{a} = \sup\{\arg\max Q_{1,z}(s,a)\}$. Since $Q_{1,z}(s,a)$ is non-decreasing in s and has increasing differences in s and z for all a, we can conclude that $V_{1,z}(s)$ is is non-decreasing in s and has increasing differences in s and z.

As the induction hypothesis, suppose that both $Q_{m,z}(s,a)$ and $V_{m,z}(s)$ are non-decreasing in s and z and have increasing differences in s and z and $\mu_{m,z}$ is stochastically non-decreasing in s and z.

Induction step is as follows: By Definition 1 and Lemma 2, both $r(s,a,z)$ and $\gamma \sum_{s'} P(s'|s,a,z) V_{m,z}(s')$ are non-decreasing in s and have increasing differences in (s,a) and z. Therefore, $Q_{m+1,z}(s,a) = r(s,a,z) + \gamma \sum_{s'} P(s'|s,a,z) V_{m,z}(s')$ also satisfies the same properties. By Lemma 3, notice that $\Omega_{m+1}(s,z) := \arg\max_{a \in \mathcal{A}}\{Q_{m+1,z}(s,a)\}$ is non-decreasing in (s,z) and therefore $\mu_{m+1,z} = \pi^{\epsilon}_{Q_{m+1,z}}$ defined as in (8) is stochastically non-decreasing in s and z as argued before. Finally, $V_{m+1,z}(s) = \sum_a \mu_{m+1,z}(s,a) Q_{m+1,z}(s,a) = (1-\epsilon) Q_{m,z}(s,\bar{a}) + \epsilon \sum_a Q_{m,z}(s,a)$ where $\bar{a} = \sup\{\arg\max Q_{m+1,z}(s,a)\}$ is non-decreasing in s and has increasing differences in s and z because $Q_{m+1,z}(s,a)$ is non-decreasing in s and has increasing differences in s and z for all a.

Since TQ-value iteration converges, $Q^*_z(s,a)$ and $V^*_z(s)$ are non-decreasing in s and have increasing differences in s and z. By repeating the same argument again, $\Omega(s,z) = \arg\max_{a \in \mathcal{A}}\{Q^*_z(s,a)\}$ is non-decreasing in (s,z). So, $\mu^*_z = \pi^{\epsilon}_{Q^*_z}$ is stochastically non-decreasing in s and z. □

Tarski's fixed-point theorem Tarski et al. (1955) ensures that monotone functions on a lattice have a fixed point. We use that result to prove the existence of T-MFE. We first state Tarski's fixed-point theorem for completeness.

Theorem 7 (Tarski's Fixed-point Theorem Tarski et al. (1955))**.** *Suppose that $\mathscr{L}$ is a nonempty complete lattice, and $T : \mathscr{L} \to \mathscr{L}$ is a non-decreasing function. Then the set of fixed points of T is a nonempty complete lattice.*

We now give the proof of Theorem 1.

Proof of Theorem 1. For any strategy μ such that μ is stochastically non-decreasing in s, and for any given mean fields z and z', define the function $K_{\mu,z}(z')$ as

$$K_{\mu,z}(z')(s') = \sum_{s \in S} \sum_{a \in \mathcal{A}} z'(s) \mu(s,a) P(s'|s,a,z).$$

From (Adlakha and Johari, 2013, Lemma 7), for $z_2 \succeq_{SD} z_1$ and $z_2' \succeq_{SD} z_1'$ and $\mu_2(s,\cdot) \succeq_{SD} \mu_1(s,\cdot)$ with μ_2 and μ_1 are stochastically non-decreasing in s, we have $K_{\mu_2,z_2}(z_2') \succeq_{SD} K_{\mu_1,z_1}(z_1')$.

Observe that McKean-Vlasov update function $\Phi(z,\mu)$ is a special case of $K_{\mu,z}(z')$ by setting $z' = z$. By the above argument, $\Phi(z,\mu) = K_{\mu,z}(z)$ is stochastically non-decreasing in (z,μ) provided that μ is stochastically non-decreasing in s.

Define $\Upsilon : \mathcal{P}(\mathcal{S}) \to \mathcal{P}(\mathcal{S})$ as $\Upsilon(z) = \Phi(z,\mu_z^*)$ where μ_z^* is the optimal trembling-hand strategy corresponding to the mean field z. Recall that, by Proposition 2, μ_z^* is stochastically non-decreasing in s and z. Then, for $z_2 \succeq_{SD} z_1$, $\Upsilon(z_2) = \Phi(z_2,\mu_{z_2}^*) \succeq_{SD} \Phi(z_1,\mu_{z_1}^*) = \Upsilon(z_1)$. From this, we can conclude that Υ is a stochastically non-decreasing function in z and hence $\Upsilon(\cdot)$ has a fixed point by Tarski's theorem (Theorem 7). In other words, there exists a z^* such that $z^* = \Upsilon(z^*) = \Phi(z^*,\mu_{z^*}^*)$. This implies that there exists a mean field z^* and strategy μ^* such that they satisfy the optimality condition (i.e., $\mu^* = \mu_{z^*}^*$) and the consistency condition (i.e. $z^* = \Phi(z^*,\mu_{z^*}^*)$). Thus, there exists a trembling-hand-perfect mean field equilibrium for mean field games with strategic complementarities. □

Proof of Theorem 2. We exploit two key monotonicity properties established before. First, from proof of Theorem 1, $\Upsilon(z)$ is stochastically non-decreasing in z. Second, from Proposition 2, an optimal trembling-hand strategy μ_z^* corresponding to a mean field z is stochastically non-decreasing in s and z.

Let z_0 be the smallest distribution by initialization in the $\preceq_{SD}$ ordering and let $\{Q_{z_k}, z_k, k \geq 0\}$ be the TQ-value functions and mean fields generated corresponding to Algorithm 1. In the following, we denote an optimal trembling-hand strategy $\mu_{z_k}^*$ corresponding to a mean field z_k simply as μ_k. By Proposition 2, $\mu_0 = \pi^{\epsilon}_{Q^*_{z_0}}$ is stochastically non-decreasing in s and z. Hence we take the following as our induction base: $z_0 \preceq_{SD} \Upsilon(z_0) = z_1$ and $\mu_0(s,\cdot) \preceq_{SD} \mu_1(s,\cdot)$ for all s where μ_1 is stochastically non-decreasing in s and z.

Now as the induction hypothesis, suppose that $z_0 \preceq_{SD} \Upsilon(z_0) = z_1 \preceq_{SD} \Upsilon(z_1) = z_2 \preceq_{SD} \cdots \preceq_{SD} \Upsilon(z_{k-1}) = z_k$ and that $\mu_0(s,\cdot) \preceq_{SD} \mu_1(s,\cdot) \preceq_{SD} \cdots \preceq_{SD} \mu_k(s,\cdot)$ for all s where μ_i are stochastically non-decreasing in s and z.

Then as an induction step, we have $\Upsilon(z_{k-1}) = z_k \preceq_{SD} z_{k+1} = \Upsilon(z_k)$ because $\Upsilon(z)$ is stochastically non-decreasing in z. Now, since $z_k \preceq_{SD} z_{k+1}$ and both μ_k and μ_{k+1} are stochastically non-decreasing in s and z, it follows that $\mu_k(s,\cdot) \preceq_{SD} \mu_{k+1}(s,\cdot)$ for all s.

Observe that Π^{ϵ} is compact for a fixed $\epsilon > 0$ and since $(\mu_k)_{k\in\mathbb{N}} \subset \Pi^{\epsilon}$ is a stochastically non-decreasing (monotone) sequence, there must be a pointwise limit μ^* such that $\mu_k \to \mu^*$ as $k \to \infty$. Moreover, $\mathcal{P}(\mathcal{S})$ is also compact since we assume that $|\mathcal{S}|$ is finite. Since $(z_k)_{k\in\mathbb{N}} \subset \mathcal{P}(\mathcal{S})$ is a stochastically non-decreasing (monotone) sequence, there must be a limit z^* such that $z_k \to z^*$ as $k \to \infty$.

It is straight forward to show that the optimal trembling-hand strategy μ_z^* and $\Upsilon(z)$ are continuous in z. So, since $\mu_k \to \mu_{z^*}^*$ and $z_k \to z^*$, we can conclude that $\mu^* = \mu_{z^*}^*$ and $z^* = \Upsilon(z^*) = \Phi(z^*,\mu^*)$.

This concludes the proof that T-BR converges to a T-MFE. □

D Proof of the Results in Section 4

D.1 Proof of Theorem 3

Proof. We only sketch the proof since it is almost the same to the proof of Theorem 2. At each time k with a fixed z_k, the value of generalized Q-learning converges, i.e., $Q_{t,z_k} \to Q^*_{z_k}$ as $t \to \infty$. Then, under assumptions of the model, $Q^*_{z_k}$ satisfies the same complementarity properties for each $k \in \mathbb{N}$. Thus, we can conclude that, for all $k \in \mathbb{N}$, $\mu_k(s,\cdot) \preceq_{SD} \mu_{k+1}(s,\cdot)$ for all $s \in \mathcal{S}$ where each μ_k is stochastically non-decreasing in s and z and that $z_k \preceq_{SD} \Upsilon(z_k) = z_{k+1}$. The rest of the proof that the limits exist follows the same as in the proof of Theorem 2. □

D.2 Proof of Theorem 4

We first prove some useful lemmas.

Lemma 4. *Let z_1 and z_2 be two arbitrary mean fields and $Q^*_{z_1}$ and $Q^*_{z_2}$ be the optimal TQ-value functions*

corresponding to them. Then, under Assumption 1,

$$||Q^*_{z_1} - Q^*_{z_2}||_\infty \leq D||z_1 - z_2||_1 \tag{9}$$

where $D = (C_1 + \gamma C_2)/(1-\gamma)^2$.

Proof. For any $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\begin{aligned}
&|Q_{z_1}(s,a) - Q_{z_2}(s,a)| \\
&\leq |r(s,a,z_1) - r(s,a,z_2)| + \gamma|\sum_{s'} P(s'|s,a,z_1)G(Q_{z_1})(s') - \sum_{s'} P(s'|s,a,z_2)G(Q_{z_2})(s')| \\
&\overset{(a)}{\leq} C_1\|z_1 - z_2\|_1 + \gamma|\sum_{s'} P(s'|s,a,z_1)|(G(Q_{z_1})(s') - G(Q_{z_2})(s'))| \\
&\qquad\qquad + \gamma\sum_{s'}|P(s'|s,a,z_1) - P(s'|s,a,z_2)||G(Q_{z_2})(s')| \\
&\overset{(b)}{\leq} C_1\|z_1 - z_2\|_1 + \gamma|\sum_{s'} P(s'|s,a,z_1)\ \max_a|Q_{z_1}(s',a) - Q_{z_2}(s',a)| \\
&\qquad\qquad + \gamma\sum_{s'}|P(s'|s,a,z_1) - P(s'|s,a,z_2)||G(Q_{z_2})\|_\infty \\
&\overset{(c)}{\leq} C_1\|z_1 - z_2\|_1 + \gamma\|Q_{z_1} - Q_{z_2}\|_\infty + \frac{\gamma}{(1-\gamma)}\|P(\cdot|\cdot,\cdot,z_1) - P(\cdot|\cdot,\cdot,z_2)\|_1 \\
&\overset{(d)}{\leq} C_1\|z_1 - z_2\|_1 + \gamma\|Q_{z_1} - Q_{z_2}\|_\infty + \frac{\gamma}{(1-\gamma)}C_2\|z_1 - z_2\|_1.
\end{aligned} \tag{10}$$

Here (a) follows from Assumption 1.(i), (b) from Lemma 1, (c) from the fact that the maximum Q-value for a finite MDP is $1/(1-\gamma)$, and (d) from Assumption 1.(ii). Since (10) is true for any (s,a), by taking the maximum on the left hand side and re arranging, we get $\|Q_{z_1} - Q_{z_2}\|_\infty \leq D\|z_1 - z_2\|_1$, where $D = \frac{C_1}{1-\gamma} + \frac{\gamma C_2}{(1-\gamma)^2}$. □

Lemma 5. *Let Q_1, Q_2 be two arbitrary TQ-value functions and let μ_1 and μ_2 be the trembling-hand strategies corresponding to them, i.e., $\mu_1 = \pi^\epsilon_{Q_1}, \mu_2 = \pi^\epsilon_{Q_2}$. Let z_1, z_2 be two arbitrary mean fields. Then, under Assumption 1,*

$$\|\Phi(z_1,\mu_1) - \Phi(z_2,\mu_2)\|_1 \leq (1 + C_2)\|z_1 - z_2\|_1 + C_3\|Q_1 - Q_2\|_\infty. \tag{11}$$

Proof. We have,

$$\begin{aligned}
&\|\Phi(z_1,\mu_1) - \Phi(z_2,\mu_2)\|_1 = \sum_{s'}|\Phi(z_1,\mu_1)(s') - \Phi(z_2,\mu_2)(s')| \\
&= \sum_{s'}|\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} z_1(s)\mu_1(s,a)P(s'|s,a,z_1) - \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} z_2(s)\mu_2(s,a)P(s'|s,a,z_2)| \\
&= \sum_{s'}|\sum_{s\in\mathcal{S}} z_1(s)P_{z_1,\mu_1}(s'|s) - \sum_{s\in\mathcal{S}} z_2(s)P_{z_2,\mu_2}(s'|s)| \\
&\leq \sum_{s'}|\sum_{s\in\mathcal{S}}(z_1(s) - z_2(s))P_{z_1,\mu_1}(s'|s)| + \sum_{s'}|\sum_{s\in\mathcal{S}} z_2(s)(P_{z_1,\mu_1}(s'|s) - P_{z_2,\mu_2}(s'|s))| \\
&\leq \|z_1 - z_2\|_1 + \|P_{z_1,\mu_1} - P_{z_2,\mu_2}\|_1 \\
&\leq \|z_1 - z_2\|_1 + \|P_{z_1,\mu_1} - P_{z_2,\mu_1}\|_1 + \|P_{z_2,\mu_1} - P_{z_2,\mu_2}\|_1 \\
&\overset{(a)}{\leq} \|z_1 - z_2\|_1 + C_2\|z_1 - z_2\|_1 + C_3\|Q_1 - Q_2\|_\infty
\end{aligned}$$

where (a) follows from Assumption 1. □

We use the following Q-learning sample complexity results from Even-Dar and Mansour (2003).

Theorem 8 (Theorem 4 in Even-Dar and Mansour (2003)). *Let Q_t be the t-th update in Q-learning algorithm using a polynomial time learning rate given as $\alpha_t(s,a) = 1/(n_t(s,a)+1)^w$, where $n_t(s,a)$ is the number of times the state-action pair (s,a) is visited until time t and $w \in (1/2, 1)$. Let L be the upper bound on the covering time. Then, $\mathbb{P}(\|Q_{T_0} - Q^*\|_\infty \leq \epsilon_3) \geq (1-\delta_3)$, for any $0 < \epsilon_3, \delta_3 < 1$, given that*

$$T_0 = O\left(\left(\frac{L^{1+3w} V_{\max}^2 \ln\left(\frac{|\mathcal{S}||\mathcal{A}|V_{\max}}{\delta_3 \beta \epsilon_3}\right)}{\beta^2 \epsilon_3^2}\right)^{\frac{1}{w}} + \left(\frac{L}{\beta}\ln(\frac{V_{\max}}{\epsilon_3})\right)^{\frac{1}{1-w}}\right). \tag{12}$$

where $V_{\max} = 1/(1-\gamma), \beta = (1-\gamma)/2$.

Here the covering time of a state-action pair sequence is the number of steps needed to visit all state-action pairs starting from any arbitrary state-action pair.

We note that the Q-learning update used in TQ-learning algorithm satisfies all the conditions necessary for the above theorem. So, we will use the above result. We refer the reader to Even-Dar and Mansour (2003) for the details.

We now give the proof Theorem 4.

Proof of Theorem 4 . Let $\{z_k\}$ and $\{\mu_k\}$ be the sequences of mean fields and strategies generated by the TMFQ-learning algorithm. Let $\{Q_{t,k}, t \geq 0\}$ be TQ-learning iterates corresponding to the mean field z_k and let $Q_k = Q_{T_0,k}$ where T_0 is as given in (12). We assume that the number of samples used in the `Next-MF` function is such that, for any given mean field z and strategy μ, `Next-MF` function returns a mean field z' such that $\mathbb{P}(\|z' - \Phi(z,\mu)\|_1 \leq \epsilon_3) \geq (1-\delta_3)$.

Define the event $E_k = \{\|Q_k - Q^*_{z_k}\|_\infty \leq \epsilon_3 \text{ and } \|z_{k+1} - \Phi(z_k,\mu_k)\|_1 \leq \epsilon_3\}$. Then, according to Theorem 8 and the assumption on the `Next-MF` function, $\mathbb{P}(E_k) \geq (1-2\delta_3)$. Define the event $E = \cap_{k=1}^{k_0} E_k$. So, $\mathbb{P}(E) \geq (1-2k_0\delta_3)$. We will now analyze the TMFQ-learning algorithm conditioned on the event E.

Let $\{\bar{z}_k\}$ and $\{\bar{\mu}_k\}$ be the sequences of mean fields and strategies generated by the T-BR algorithm. We assume that T-BR algorithm and TMFQ-learning algorithm have the same initialization, i.e., $\bar{z}_0 = z_0$. Now, conditioned on the event E,

$$\begin{aligned}
\|\bar{z}_{k+1} - z_{k+1}\|_1 &\leq \|\Phi(\bar{z}_k, \bar{\mu}_k) - \Phi(z_k,\mu_k)\|_1 + \|\Phi(z_k,\mu_k) - z_{k+1}\|_1 \\
&\overset{(a)}{\leq} (1+C_2)\|\bar{z}_k - z_k\|_1 + C_3\|Q^*_{\bar{z}_k} - Q^*_k\|_\infty + \epsilon_3 \\
&\leq (1+C_2)\|\bar{z}_k - z_k\|_1 + C_3\|Q^*_{\bar{z}_k} - Q^*_{z_k}\|_\infty + C_3\|Q^*_{z_k} - Q^*_k\|_\infty + \epsilon_3 \\
&\overset{(b)}{\leq} (1+C_2)\|\bar{z}_k - z_k\|_1 + C_3 D\|\bar{z}_k - z_k\|_1 + (C_3+1)\epsilon_3 \\
&\leq (1+C_2+C_3D)\|\bar{z}_k - z_k\|_1 + (C_3+1)\epsilon_3
\end{aligned}$$

Here (a) follows from Lemma 5 and the assumption on the `Next-MF` function and (b) follows from Lemma 4.

Iteratively applying the above inequality, we get $\|\bar{z}_{k_0} - z_{k_0}\|_1 \leq B\epsilon_3$, where $B = (1+C_2+C_3D)^{k_0+1}(C_3+1)$. Now, $\|\bar{z}^* - z_{k_0}\|_1 \leq \|\bar{z}^* - \bar{z}_{k_0}\|_1 + \|\bar{z}_{k_0} - z_{k_0}\|_1 \leq \bar{\epsilon} + B\epsilon_3$ because $\|\bar{z}^* - \bar{z}_{k_0}\|_1 \leq \bar{\epsilon}$ by the definition of k_0.

So, $\mathbb{P}(\|\bar{z}^* - z_{k_0}\|_1 \leq \bar{\epsilon} + B\epsilon_3) \geq \mathbb{P}(E) = 1 - 2k_0\delta$.

Setting $\epsilon_3 = \bar{\epsilon}/B$ and $\delta_3 = \delta/2k_0$, and using the corresponding T_0 from (12), we get the desired result. □

Proof of Corollary 1 . Conditioned on the event E as defined in the proof Theorem 4, we get,

$$\begin{aligned}
\|\bar{z}_{k+1} - z_{k+1}\|_1 &\leq \|\Phi(\bar{z}_k, \bar{\mu}_k) - \Phi(z_k,\mu_k)\|_1 + \|\Phi(z_k,\mu_k) - z_{k+1}\|_1 \\
&\overset{(a)}{\leq} C_4\|\bar{z}_k - z_k\|_1 + C_5\|Q^*_{\bar{z}_k} - Q^*_k\|_\infty + \epsilon_3 \\
&\leq C_4\|\bar{z}_k - z_k\|_1 + C_5\|Q^*_{\bar{z}_k} - Q^*_{z_k}\|_\infty + C_5\|Q^*_{z_k} - Q^*_k\|_\infty + \epsilon_3 \\
&\overset{(b)}{\leq} C_4\|\bar{z}_k - z_k\|_1 + C_5 D\|\bar{z}_k - z_k\|_1 + (C_5+1)\epsilon_3 \\
&\leq (C_4+C_5D)\|\bar{z}_k - z_k\|_1 + (C_5+1)\epsilon_3
\end{aligned}$$

Here (a) follows from Assumption 2 and and (b) follows from Lemma 4.

Iteratively applying the above inequality, we get $\|\bar{z}_{k_0} - z_{k_0}\|_1 \leq B\epsilon_3$, where $B = (C_5+1)/(1-((C_4+C_5D)))$. Rest of the proof is similar to that of Theorem 4. □

D.3 Proof of Theorem 6

For a given mean field z, let $\widehat{P}(\cdot|\cdot,\cdot,z)$ be the estimate of the model obtained by taking n_0 next-state samples for each (s,a). Let $\widehat{F}_z$ be the approximate TQ-value operator obtained by replacing P by $\widehat{P}$ in (5). Similar to the proof of Proposition 1, it is straight forward to show that $\widehat{F}_z$ is a contraction. Let $\widehat{Q}_z^*$ be its unique fixed point. Since $\widehat{P}$ is different from P, $\widehat{Q}_z^*$ and Q_z^* will also be different. However, for sufficiently large n_0, we can give the following bound.

Lemma 6. *For any* $0 < \epsilon_4, \delta_4 < 1$,

$$\mathbb{P}(\|\widehat{Q}_z^* - Q_z^*\|_\infty \leq \epsilon_4) \geq (1-\delta_4), \quad \textit{for} \;\; n_0 \geq \frac{2V_{\max}^4}{\epsilon_4^2}\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta_4}\right) \tag{13}$$

where $V_{\max} = 1/(1-\gamma)$.

Proof.

$$\begin{aligned}
&|\widehat{Q}_z^*(s,a) - Q_z^*(s,a)| = |\widehat{F}_z(\widehat{Q}_z^*)(s,a) - F_z(Q_z^*)(s,a)| \\
&= \gamma|\sum_{s'}\widehat{P}(s'|s,a,z)G(\widehat{Q}_z^*)(s') - \sum_{s'}P(s'|s,a,z)G(Q_z^*)(s')| \\
&\leq \gamma|\sum_{s'}\widehat{P}(s'|s,a,z)(G(\widehat{Q}_z^*)(s') - G(Q_z^*)(s'))| + \gamma|\sum_{s'}(\widehat{P}(s'|s,a,z) - P(s'|s,a,z))G(Q_z^*)(s')| \\
&\overset{(a)}{\leq} \gamma\|\widehat{Q}^* - Q_z^*\|_\infty + \gamma|\sum_{s'}(\widehat{P}(s'|s,a,z)) - P(s'|s,a,z))G(Q_z^*(s')|,
\end{aligned} \tag{14}$$

where (a) follows from Lemma 1.

For bounding $|\sum_{s'}(\widehat{P}(s'|s,a,z) - P(s'|s,a,z))G(Q_z^*)(s')|$, note that $\sum_{s'}\widehat{P}(s'|s,a,z))G(Q_z^*)(s')$ is an unbiased estimated of $\sum_{s'}P(s'|s,a,z)G(Q_z^*)(s')$. Also note that
$\max_{s'}|G(Q_z^*)(s')| \leq 1/(1-\gamma) = V_{\max}$. So, by applying Hoeffding's inequality, for a given (s,a), we get

$$\mathbb{P}(|\sum_{s'}(\widehat{P}(s'|s,a,z)) - P(s'|s,a,z))G(Q_z^*)(s')| \geq \epsilon) \leq 2\exp\left(\frac{-n_0\epsilon^2}{2V_{\max}^2}\right).$$

Using the union bound argument, for all (s,a), we get

$$\mathbb{P}(|\sum_{s'}(\widehat{P}(s'|s,a,z)) - P(s'|s,a,z))G(Q_z^*)(s')| \leq \epsilon) \geq 1 - 2|\mathcal{S}||\mathcal{A}|\exp\left(\frac{-n_0\epsilon^2}{2V_{\max}^2}\right).$$

So, with $n_0 \geq \frac{2V_{\max}^2}{\epsilon^2}\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)$,

$$\mathbb{P}(|\sum_{s'}(\widehat{P}(s'|s,a,z)) - P(s'|s,a,z))G(Q_z^*)(s')| \leq \epsilon) \leq 1-\delta, \;\forall (s,a) \in \mathcal{S}\times\mathcal{A}. \tag{15}$$

Now, for the above n_0, from (14) and (15), $|\widehat{Q}_z^*(s,a) - Q_z^*(s,a)| \leq \gamma\|\widehat{Q}^* - Q_z^*\|_\infty + \epsilon$, with a probability greater than $(1-\delta)$ for all (s,a). This implies that $\|\widehat{Q}^* - Q_z^*\|_\infty \leq \epsilon/(1-\gamma) = \epsilon V_{\max}$. Now, using $\epsilon = \epsilon_4/V_{\max}$ and $\delta = \delta_4$ in the expression for n_0 above, we get the desired result. □

We now bound the error in the Mckean-Vlasov update due to of replacing P by $\widehat{P}$.

Lemma 7. *Let $\widehat{\Phi}$ be the approximate Mckean-Vlasov update function as defined in* (1) *but by replacing P by $\widehat{P}$. Then,*

$$\mathbb{P}(\|\widehat{\Phi}(z,\mu)-\Phi(z,\mu)\|_1 \leq \epsilon_4) \geq (1-\delta_4), \text{ for } n_0 \geq \frac{2}{\epsilon_4^2}\log\left(\frac{2^{|\mathcal{S}|}|\mathcal{S}||\mathcal{A}|}{\delta_4}\right) \tag{16}$$

Proof. From Weissman et al. (2003), for a given (s,a),

$$\mathbb{P}(\|\widehat{P}(\cdot|s,a,z)-P(\cdot|s,a,z)\|_1 \geq \epsilon_4) \leq 2^{|\mathcal{S}|}\exp\left(\frac{-n\epsilon_4^2}{2}\right).$$

By the union bound argument, for all (s,a), we get

$$\mathbb{P}(\|\widehat{P}(\cdot|s,a,z)-P(\cdot|s,a,z)\|_1 \leq \epsilon_4) \geq 1-|\mathcal{S}||\mathcal{A}|2^{|\mathcal{S}|}\exp\left(\frac{-n\epsilon_4^2}{2}\right).$$

So, with $n_0 \geq \frac{2}{\epsilon_4^2}\log\left(\frac{2^{|\mathcal{S}|}|\mathcal{S}||\mathcal{A}|}{\delta_4}\right)$

$$\mathbb{P}(\|\widehat{P}(\cdot|s,a,z)-P(\cdot|s,a,z)\|_1 \leq \epsilon_4) \geq 1-\delta_4, \ \forall (s,a)\in\mathcal{S}\times\mathcal{A}. \tag{17}$$

Now, with the above n_0, with a probability greater than $(1-\delta_4)$, we get

$$\begin{aligned}
\|\widehat{\Phi}(z,\mu)-\Phi(z,\mu)\|_1 &= \sum_{s'}|\widehat{\Phi}(z,\mu)(s')-\Phi(z,\mu)(s')| \\
&\leq \sum_{s'}\sum_{s}\sum_{a} z(s)\mu(s,a)|\widehat{P}(s'|s,a,z)-P(s'|s,a,z)| \\
&= \sum_{s}\sum_{a} z(s)\mu(s,a)\|\widehat{P}(\cdot|s,a,z)-P(\cdot|s,a,z)\|_1 \leq \epsilon_4
\end{aligned}$$

where the last inequality follows from (17) and the fact that $\sum_s\sum_a z(s)\mu(s,a)=1$. □

We now give the proof Theorem 6.

Proof of Theorem 6. Let $\{z_k\}$ and $\{\mu_k\}$ be the sequences of mean fields and strategies generated by the GMBL algorithm. Let n_0 be the maximum of the two values given by Lemma 6 and Lemma 7, i.e,

$$n_0 = \max\left(\frac{2V_{\max}^4}{\epsilon_4^2}\log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta_4}\right), \frac{2}{\epsilon_4^2}\log\left(\frac{2^{|\mathcal{S}|}|\mathcal{S}||\mathcal{A}|}{\delta_4}\right)\right)$$

Define the event $E_k = \{\|\widehat{Q}^*_{z_k}-Q^*_{z_k}\|_\infty \leq \epsilon_4 \text{ and } \|\widehat{\Phi}(z_k,\mu_k)-\Phi(z_k,\mu_k)\|_1 \leq \epsilon_4\}$. Then, from Lemma 6 and Lemma 7, we get $\mathbb{P}(E_k) \geq (1-2\delta_4)$. Define the event $E = \cap_{k=1}^{k_0} E_k$. So, $\mathbb{P}(E) \geq (1-2k_0\delta_4)$. We will now analyze the GMBL algorithm conditioned on the event E.

Let $\{\bar{z}_k\}$ and $\{\bar{\mu}_k\}$ be the sequences of mean fields and strategies generated by the T-BR algorithm. We assume that T-BR algorithm and GMBL algorithm have the same initialization, i.e., $\bar{z}_0 = z_0$. Now, conditioned on the event E

$$\begin{aligned}
\|\bar{z}_{k+1}-z_{k+1}\|_1 &= \|\Phi(\bar{z}_k,\bar{\mu}_k)-\widehat{\Phi}(z_k,\mu_k)\|_1 \\
&\leq \|\Phi(\bar{z}_k,\bar{\mu}_k)-\Phi(z_k,\mu_k)\|_1 + \|\Phi(z_k,\mu_k)-\widehat{\Phi}(z_k,\mu_k)\|_1 \\
&\overset{(a)}{\leq} (1+C_2)\|\bar{z}_k-z_k\|_1 + C_3\|Q^*_{\bar{z}_k}-\widehat{Q}^*_{z_k}\|_\infty + \epsilon_4 \\
&\leq (1+C_2)\|\bar{z}_k-z_k\|_1 + C_3\|Q^*_{\bar{z}_k}-Q^*_{z_k}\|_\infty + C_3\|Q^*_{z_k}-\widehat{Q}^*_{z_k}\|_\infty + \epsilon_4 \\
&\overset{(b)}{\leq} (1+C_2)\|\bar{z}_k-z_k\|_1 + C_3D\|\bar{z}_k-z_k\|_1 + (C_3+1)\epsilon_4 \\
&\leq (1+C_2+C_3D)\|\bar{z}_k-z_k\|_1 + (C_3+1)\epsilon_4
\end{aligned}$$

Here (a) follows from Lemma 5 and (b) follows from Lemma 4.

Iteratively applying the above inequality, we get $\|\bar{z}_{k_0} - z_{k_0}\|_1 \leq B\epsilon_4$, where $B = (1 + C_2 + C_3 D_1)^{k_0+1}(C_3 + 1)$. Now, conditioned on the event E, we have

$$\|\bar{z}^* - z_{k_0}\|_1 \leq \|\bar{z}^* - \bar{z}_{k_0}\|_1 + \|\bar{z}_{k_0} - z_{k_0}\|_1 \leq \bar{\epsilon} + B\epsilon_4$$

because $\|\bar{z}^* - \bar{z}_{k_0}\|_1 \leq \bar{\epsilon}$ by the definition of k_0.

Setting $\epsilon_4 = \bar{\epsilon}/B$ and $\delta_4 = \delta/2k_0$ in the expression for n_0, we get the desired result. □

E Proof of the Results in Section 5

E.1 Proof of Theorem 5

While Q-learning algorithm is an asynchronous process since a particular state-action pair is updated at a time, if all state-action pairs are updated at each time, it is called synchronous Q-learning algorithm Even-Dar and Mansour (2003) and we define synchronous TQ-learning algorithm as follows: For a fixed mean field z,

$$\begin{aligned} &Q_{0,z}(s,a) = 0 \ \text{ for all } \ (s,a) \in S \times A \\ &Q_{t+1,z}(s,a) = (1-\alpha)Q_{t,z}(s,a) + \alpha(r(s,a,z) + \gamma G(Q)(s)) \ \ \forall \ (s,a) \in S \times A \end{aligned} \tag{18}$$

where α_t is the appropriate learning rate. It can be shown that $Q_{t+1,z} \to Q_z^*$ as $t \to \infty$ and observe that the synchronous TQ-value function update (18) requires that all state-action pairs are sampled at each t. Due to the availability of a large population of agents that regenerate to occupy all states, and explore all actions via trembling hand strategies, this is a mild condition.

We use the following synchronous Q-learning sample complexity results from Even-Dar and Mansour (2003).

Theorem 9 (Theorem 2 in Even-Dar and Mansour (2003))**.** *Let Q_t be the t-th update in synchronous Q-learning algorithm using a polynomial time learning rate given as $\alpha_t(s,a) = 1/(n_t(s,a)+1)^w$, where $n_t(s,a)$ is the number of times the state-action pair (s,a) is visited until time t and $w \in (1/2, 1)$. Then, $\mathbb{P}(\|Q_{I_0} - Q^*\|_\infty \leq \epsilon_3) \geq (1-\delta_3)$, for any $0 < \epsilon_3, \delta_3 < 1$, given that*

$$I_0 = O\left(\left(\frac{V_{\max}^2 \ln\left(\frac{|\mathcal{S}||\mathcal{A}|V_{\max}}{\delta_3 \beta \epsilon_3}\right)}{\beta^2 \epsilon_3^2}\right)^{\frac{1}{w}} + \left(\frac{1}{\beta}\ln(\frac{V_{\max}}{\epsilon_3})\right)^{\frac{1}{1-w}}\right). \tag{19}$$

where $V_{\max} = 1/(1-\gamma), \beta = (1-\gamma)/2$.

Unlike asynchronous Q-learning sample complexity, there is no dependence of covering time since all state-action pairs are updated at each time. We note that the synchronous Q-learning update used in synchronous TQ-learning algorithm satisfies all the conditions necessary for the above theorem. So, we will use the above result. We refer the reader to Even-Dar and Mansour (2003) for the details.

We provide a slight modification of T-BR algorithm which intentionally include a mismatch observed in online TMFQ-learning algorithm. As done in the proof of Theorem 4 where we compare trajectories of T-BR and TMFQ-learning algorithms, we compare modified T-BR and Online TMFQ-learning in a similar way because there is no loss of generality, i.e., modified T-BR also converges to a T-MFE. We can compare with T-BR and Online TMFQ-learning but there is an additional term. To elaborate about the Modified T-BR Algorithm 6, at each k with z_k, Algorithm 6 computes $Q_{z_k}^*$ and μ_k but, in McKean-Vlasov equation, μ_{k-1} is employed that is computed in previous step $k-1$, i.e., $\Phi(z_k, \mu_{k-1})$, and uses μ_k in next time step $k+1$. This is intentional and almost sure convergence to a T-MFE can be proved similarly.

Proof of Theorem 5. Let $\{z_k\}$ and $\{\mu_k\}$ be the sequences of mean fields and strategies generated by the Online TMFQ-learning algorithm. Let $\{Q_{t,k}, t \geq 0\}$ be Offline (or Batch) TQ-learning iterates corresponding to the mean field z_k and let $Q_k = Q_{I_0,k}$ where I_0 is as given in (19).

Define the event $E_k = \{\|Q_k - Q_{z_k}^*\|_\infty \leq \epsilon_3\}$. Then, according to Theorem 9, $\mathbb{P}(E_k) \geq (1 - 2\delta_3)$. Define the event $E = \cap_{k=1}^{k_0} E_k$. So, $\mathbb{P}(E) \geq (1 - 2k_0\delta_3)$. We will now analyze the Online TMFQ-learning algorithm conditioned on the event E.

Algorithm 6: Modified T-BR Algorithm

1: Initialization: Initial mean field z_0 and strategy μ_0
2: **for** $k = 1, 2, 3, \dots$ **do**
3: For the mean field z_k, compute the optimal TQ-Value function $Q^*_{z_k}$ using the TQ-value iteration $Q_{m+1,z_k} = F(Q_{m,z_k})$
4: Compute the strategy $\mu_k = \Psi(z_k)$ as the trembling-hand strategy w.r.t $Q^*_{z_k}$, i.e, $\mu_k = \pi^{\epsilon}_{Q^*_{z_k}}$
5: Compute the next mean field $z_{k+1} = \Phi(z_k, \mu_{k-1})$
6: **end for**

Let $\{\bar{z}_k\}$ and $\{\bar{\mu}_k\}$ be the sequences of mean fields and strategies generated by Modified T-BR algorithm. We assume that Modified T-BR algorithm and TMFQ-learning algorithm have the same initialization, i.e., $\bar{z}_0 = z_0$. Now, conditioned on the event E,

$$\begin{aligned}
\|\bar{z}_{k+1} - z_{k+1}\|_1 &= \|\Phi(\bar{z}_k, \bar{\mu}_{k-1}) - \Phi(z_k, \mu_{k-1})\|_1 \\
&\overset{(a)}{=} (1-\zeta) \sum_{s'} |\sum_{s \in \mathcal{S}} \bar{z}_k(s) P_{\bar{z}_k, \bar{\mu}_{k-1}}(s'|s) - \sum_{s \in \mathcal{S}} z_k(s) P_{z_k, \mu_{k-1}}(s'|s)| + \zeta \sum_{s'} |\Psi(s') - \Psi(s')| \\
&\overset{(b)}{\leq} (1-\zeta)(1+C_2)\|\bar{z}_k - z_k\|_1 + (1-\zeta)C_3 \|Q^*_{\bar{z}_{k-1}} - Q^*_{k-1}\|_\infty \\
&\leq (1-\zeta)(1+C_2)\|\bar{z}_k - z_k\|_1 + (1-\zeta)C_3 \|Q^*_{\bar{z}_{k-1}} - Q^*_{z_{k-1}}\|_\infty + (1-\zeta)C_3 \|Q^*_{z_{k-1}} - Q^*_{k-1}\|_\infty \\
&\overset{(c)}{\leq} (1-\zeta)(1+C_2)\|\bar{z}_k - z_k\|_1 + (1-\zeta)C_3 D \|\bar{z}_{k-1} - z_{k-1}\|_1 + (1-\zeta)C_3 \epsilon_3
\end{aligned}$$

Here (a) follows from ζ regeneration event where Ψ is the probability measure of the agent regeneration process, (b) follows from Lemma 5 and (c) follows from Lemma 4.

Iteratively applying the above inequality, we get $\|\bar{z}_{k_0} - z_{k_0}\|_1 \leq B\epsilon_3$, where $B = \frac{(2(1-\zeta)A)^{k_0}(C_3\epsilon)}{(1-\zeta)(A-1)}$ where $A = \max\{1 + C_2, C_3 D\}$. Now, $\|\bar{z}^* - z_{k_0}\|_1 \leq \|\bar{z}^* - \bar{z}_{k_0}\|_1 + \|\bar{z}_{k_0} - z_{k_0}\|_1 \leq \bar{\epsilon} + B\epsilon_3$ because $\|\bar{z}^* - \bar{z}_{k_0}\|_1 \leq \bar{\epsilon}$ by the definition of k_0.

So, $\mathbb{P}(\|\bar{z}^* - z_{k_0}\|_1 \leq \bar{\epsilon} + B\epsilon_3) \geq \mathbb{P}(E) = 1 - 2k_0\delta$.

Setting $\epsilon_3 = \bar{\epsilon}/B$ and $\delta_3 = \delta/2k_0$, and using the corresponding I_0 from (19), we get the desired result. □

F Experiments

F.1 Parameters for Infection Spread Model

We use the following parameters for simulations

$$|\mathcal{S}| = 25 \quad |\mathcal{A}| = 5 \quad k = 0.05 \quad \delta_1 = 1 \quad \delta_2 = 0.2 \qquad \delta_3 = 0.01 \quad \zeta = 0.1 \quad \epsilon = 0.3 \quad \gamma = 0.75$$

$$w_1 \sim \begin{cases} \text{uniform}\{1,2,3\} & \text{w.p} \quad 0.9 \\ 0 & \text{w.p} \quad 0.1 \end{cases} \qquad w_2 \sim \begin{cases} \text{uniform}\{0, \cdots, s\} & \text{w.p} \quad 0.9 \\ 0 & \text{w.p} \quad 0.1 \end{cases}$$

For GMBL, we set $n_0 = 500$ and we run the outer loop for 500 iterations. For TMFQL we run Q Learning for 1000 time steps, and preform 5000 iterations of the outer loop. For O-TMFQ Learning, run the outer loop for 5000 iterations with different number of agents. For better sample efficiency, in both TMFQL and O-TMFQ Learning we initialize the Q function at each iteration with Q of the previous iteration. We used a logarithmically decaying learning rate, we decay the learning rate from 10^{-3} to 10^{-2}.

For IQL we initialize each agent with a Q function and use the same parameters as O-TMFQ. We simulate a variant of MFQ where each agent estimates the meanfield to be the average state of a subset of the population and uses it to parameterize its Q function. We define this subset to be 512 agents chosen at random and kept constant for the duration of the simulation. Each agent also obtains samples from these 512 agents to updates its Q-Function. The other parameters are same as O-TMFQ.

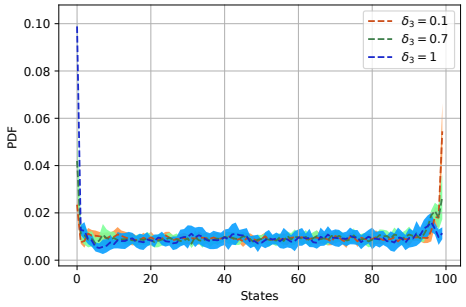

Figure 7: PDF of O-TMFQ-Learning

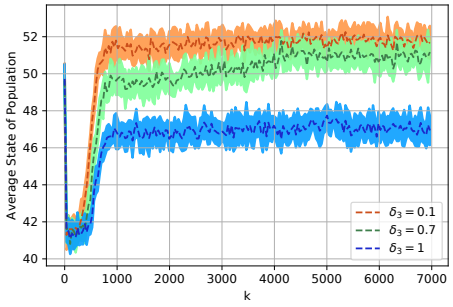

Figure 8: Mean state of O-TMFQ-Learning

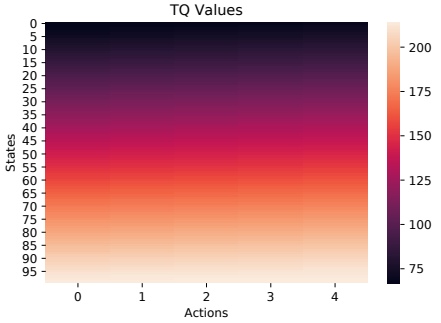

Figure 9: Heat map of TQ Values for $\delta_3 = 0.1$.

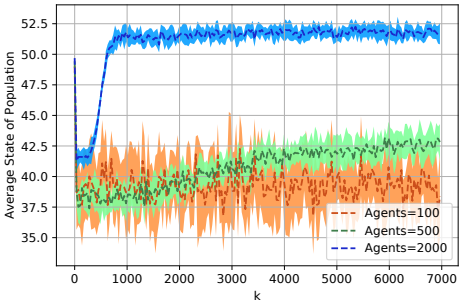

Figure 10: Mean state of O-TMFQ-Learning: $\delta_3 = 0.1$

F.2 Amazon Mechanical Turk (MTurk) and other Gig Economy Marketplaces

We consider Amazon Mechanical Turk (MTurk) as an example of a Gig economy marketplace, which includes firms like Uber and Airbnb. MTurk is a crowd sourcing market, wherein human workers are recruited to perform so-called Human Intelligence Tasks (HITs). These HITs may take the form of labeling data sets or other tasks that are simple from a human's perspective, but might be difficult for machine learning to directly undertake. The workers are called Turkers, and each has a quality score that depends on previous HITs undertaken. The firm that originates these HITs may specify the price that it is willing to pay, as well as the minimum quality of the Turkers that it desires.

There is a natural alignment of effort employed by Turkers in MTurk, since higher efforts translate into HITs done right, which in turn results in a higher quality, which finally results in firms willing to pay more per HIT. Thus, if mean field quality is high, there is an incentive to perform HITs well and enhance ones' own quality. This notion of incentive alignment applies to essentially all Gig economy marketplaces—the reputation of the agent directly enhances its reward, while the reputation of the marketplace as a whole (i.e., its mean field) draws customers willing to pay into the system, and so enhances the reward of the agent.

The formal system description is analogous to Infection Spread considered earlier, but we focus here on the application scenario. Thus, we have that each agent (Turker) has his/her quality state, and the strategic action is the choice of how much work to put into a HIT assigned to that Turker. Higher effort implies higher cost, but also implies a higher improvement in the quality. The overall reward is a combination that depends on the Turker's quality as well as the mean field quality.

Let s denote the quality of a Turker's profile, and let action a denote the effort the turker puts in to maintain the quality of the profile; this may include number of jobs successfully completed, time taken to complete a job etc. Let $c(a) = \delta_3 a$ denote the cost incurred in performing action a. The quality perceived by an entity offering jobs depends on both the quality of the individual Turker and the population as a whole (via the mean field). Thus, the reward to a Turker is a function of the perceived quality and cost incurred in taking action. We define state

transition and reward as follows,

$$s' = (s + a - w_1)_+ \mathbb{1}\{E_1\} + w_2 \mathbb{1}\{E_2\}$$
$$r = \delta_1 s + \delta_2 \sum_{s \in \mathcal{S}} s z(s) - \delta_3 a,$$

where E_1, E_2 are mutually exclusive events that occur with probabilities $1 - \zeta, \zeta$, respectively, and w_1, w_2 are realizations of non-negative integer random variables.

We use the following parameters for simulations,

$$|\mathcal{S}| = 100 \quad |\mathcal{A}| = 5 \quad \delta_1 = 0.5 \quad \delta_2 = 0.2 \quad \zeta = 0.1 \quad \epsilon = 0.3 \quad \gamma = 0.75$$
$$w_1 \sim \text{uniform}\{0, 1, 2, 3\} \quad w_2 \sim \begin{cases} \text{uniform}\{0, \cdots, |\mathcal{S}|\} & \text{w.p} \quad 0.9 \\ 0 & \text{w.p} \quad 0.1 \end{cases}$$

For O-TMFQ Learning, we run the outer iteration for 7000 steps. As before, for better sample efficiency we initialize the Q function at each iteration with Q of the previous iteration. We used a logarithmically decaying learning rate, we decay the learning rate from 10^{-3} to 10^{-2}.

The behavior of our RL algorithms is much the same as the earlier case, and is shown in Figures 7–10. Figure 7 shows the pdf of the final mean field distribution obtained by performing O-TMFQL with 2000 agents with different values of δ_3 while figure 8 shows the average state of the population. Observe that as the cost of action δ_3 increases, agents take lower actions and are hence distributed towards lower states. Figure 9 is a heat map of the final TQ value function. Figure 10 shows the mean state evolution with different number of agents for $\delta_3 = 0.1$. Observe that the convergence is poor with lesser number of agents.