
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

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 $\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r^i(s_t, a_t)]$, 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 $\mu : \mathcal{S} \rightarrow \mathcal{P}(\mathcal{A})$, 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 $V_{\mu, z}(s) = \mathbb{E}[\sum_{k=0}^{\infty} \gamma^k r(s_k, a_k, z) | s_0 = s]$, 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 $f : X \rightarrow \mathbb{R}$ 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 $f : X \times Y \rightarrow \mathbb{R}$ 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 $T : X \rightarrow Y$ 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 complementarities 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 $a \in \{1, 2, 3, \dots, |\mathcal{A}|\}$ (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

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

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_i s 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_i s 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 $\mu : \mathcal{S} \rightarrow \mathcal{P}^\epsilon(\mathcal{A})$, 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

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

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

We define trembling-hand-perfect 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 $\mu_z^* = \pi_{Q_z^*}^\epsilon$.

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

hand strategy π_Q^ϵ and the function $G(Q)$ as

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

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

Note that π_Q^ϵ 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

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

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., $Q_{m, z} \rightarrow Q_z^*$. 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** $k = 0, 1, 2, \dots$ **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., $\mu_k = \pi_{Q_{z_k}^*}^\epsilon$
 - 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 $\mu_k \rightarrow \mu^*$ and $z_k \rightarrow z^*$ as $k \rightarrow \infty$ 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:

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

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 $Q_{t,z_k} \rightarrow Q_{z_k}^*$ as $t \rightarrow \infty$.

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 ?? (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

```

1: Initialization: Initial mean field  $z_0$ 
2: for  $k = 0, 1, 2, \dots$  do
3:   Initialize time step  $t \leftarrow 0$ . Initialize  $s_0, Q_{0,z_k}$ 
4:   repeat
5:     Take action  $a_t \sim \pi_{Q_{t,z_k}}^\epsilon(s_t)$ , observe reward
       and the next state  $s_{t+1} \sim P(\cdot|s_t, a_t, z_k)$ 
6:     Update  $Q_{t,z_k}$  according to TQ-learning (6)
7:      $t \leftarrow t + 1$ 
8:   until  $\|Q_{t,z_k} - Q_{t-1,z_k}\|_\infty < \epsilon_1$ 
9:   Let  $Q_{z_k} = Q_{t,z_k}$  and the strategy  $\mu_k = \pi_{Q_{z_k}}^\epsilon$ 
10:   $z_{k+1} = \text{Next-MF}(z_k, \mu_k)$ 
11: 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 $\mu_k \rightarrow \mu^*$ and $z_k \rightarrow z^*$ as $k \rightarrow \infty$ 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., $\mu_1 = \pi_{Q_1}^\epsilon, \mu_2 = \pi_{Q_2}^\epsilon$. 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*

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

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)).

¹Covering 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 $\mu_1 = \pi_{Q_1}^\epsilon, \mu_2 = \pi_{Q_2}^\epsilon$. 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 $(C_4 + C_5D) < 1$, where $D = (C_1 + \gamma C_2)/(1 - \gamma)$.

Corollary 1. Let Assumption 1 and Assumption 2 hold. Then, we obtain the bound on T_0 as in (7) with $B = (C_5 + 1)/(1 - (C_4 + C_5D))$, 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 ??.

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 $P_{z_{k+1}, \cdot}$, 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 $P_{z_{k+1}, \cdot}$, 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

- 1: Initialize mean field z_0 and strategy μ_0
 - 2: **for** $k = 1, 2, \dots$ **do**
 - 3: Reset *memory buffer*
 - 4: **for** agents $i = 1, 2, \dots, I$ **do**
 - 5: Given current state s_k^i , take action $a_k^i \sim \mu_{k-1}(s_k^i)$, get reward r_k^i , observe the next state s_{k+1}^i
 - 6: Add the sample $(s_k^i, a_k^i, r_k^i, s_{k+1}^i)$ to the *memory buffer*
 - 7: **end for**
 - 8: Perform TQ-learning on the *memory buffer* to obtain Q_{z_k} and $\mu_k = \pi_{Q_{z_k}}^\epsilon$
 - 9: **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

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

where $B = \frac{((1-\zeta)2A)^{k_0}(C_3\epsilon)}{(1-\zeta)(A-1)}$, $A = \max\{1 + C_2, C_3D\}$, $D = (C_1 + \gamma C_2)/(1 - \gamma)$, $V_{\max} = 1/(1 - \gamma)$, $\beta = (1 - \gamma)/2$, $w \in (1/2, 1)$, 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.

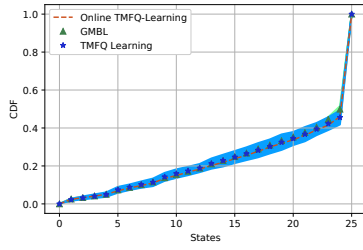
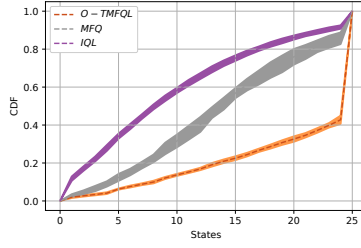
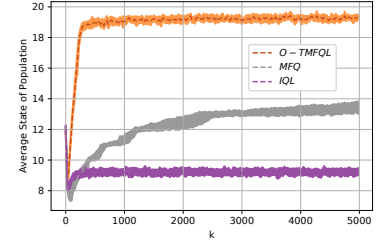
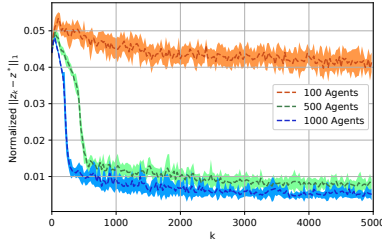
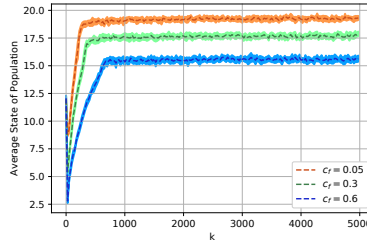

 Figure 1: TMFE with $c_f = 0.1$

 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

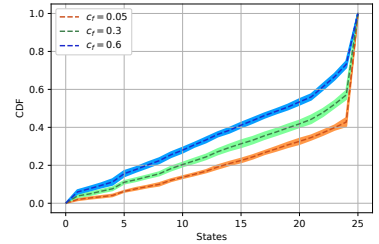


Figure 6: CDF of O-TMFQ-Learning

6 Experiments

We consider *Infection Spread* described in Section 2. State space is $\mathcal{S} = \{0, \dots, 24\}$ with 0 being the lowest health level. Action space is $\mathcal{A} = \{0, \dots, 4\}$ 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.
- 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. Hadikhannloo. 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.
- 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.
- 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.