Last-iterate Convergence in Regularized Graphon Mean Field Game

Jing Dong[1,*], **Baoxiang Wang**[1,3,*], **Yaoliang Yu**[2,3,*]

[1] Chinese University of Hong Kong, Shenzhen,
[2] University of Waterloo,
[3] Vector Institute
jingdong@link.cuhk.edu.cn, bxiangwang@cuhk.edu.cn, yaoliang.yu@uwaterloo.ca

Abstract

To model complex real-world systems, such as traders in stock markets, or the dissemination of contagious diseases, graphon mean-field games (GMFG) have been proposed to model many agents. Despite the empirical success, our understanding of GMFG is limited. Popular algorithms such as mirror descent are deployed but remain unknown for their convergence properties. In this work, we give the first last-iterate convergence rate of mirror descent in regularized monotone GMFG. In tabular monotone GMFG with finite state and action spaces and under bandit feedback, we show a last-iterate convergence rate of $O(T^{-1/4})$. Moreover, when exact knowledge of costs and transitions is available, we improve this convergence rate to $O(T^{-1})$, matching the existing convergence rate observed in strongly convex games. In linear GMFG, our algorithm achieves a last-iterate convergence rate of $O(T^{-1/5})$. Finally, we verify the performance of the studied algorithms by empirically testing them against fictitious play in a variety of tasks.

Introduction

In many real-world applications, the presence of complex systems including many interacting individuals or components is indispensable. These systems manifest in various forms, from the intricate networks of neurons within human brains (Bullmore and Sporns 2009, 2012; Avena-Koenigsberger, Misic, and Sporns 2018), to the dynamic interactions of traders in stock markets (Bakker et al. 2010; Bian, Xu, and Li 2016), and to the dissemination of contagious diseases throughout societies (Newman 2002; Pastor-Satorras et al. 2015). Due to the large number of interacting individuals or components, these systems pose significant challenges for modeling. Mean field games (MFG) (Caines, Huang, and Malhamé 2006; Lasry and Lions 2007) have emerged as a highly effective approach for addressing this complexity, offering both scalability and robust theoretical guarantees in these multi-agent systems. MFG operates on the principle of weak interactions, positing that each individual's influence on the overall system is negligible. This framework has been successfully applied to many real-world tasks, including social networks (Yang et al. 2018), and swarm robotics (Cui et al. 2023).

The MFG framework leverages the assumption of agent homogeneity and has demonstrated success across various applications. However, this assumption becomes a hindrance when dealing with heterogeneous agents. To address this limitation, the Graphon mean field games (GMFG) (Parise and Ozdaglar 2019; Aurell et al. 2022) framework has been introduced as an extension of MFGs to accommodate heterogeneous agent modeling. The GMFG framework captures agent interactions through a graphical structure and is shown to be successful in applications like modeling investment decisions in financial markets (Tangpi and Zhou 2024). Despite the empirical success of MFG and GMFG, our theoretical understanding of this framework remains limited.

In monotone MFGs (Lasry and Lions 2007), and under the access to the exact cost and transition functions, (Perrin et al. 2020) proposed a continuous time fictitious play algorithm, where the averaged iterates policy converge to a Nash equilibrium in $O(T^{-1})$ iterations. For discrete-time monotone GMFGs with access to exact cost and transition functions, (Zhang et al. 2023) proposed a mirror descent-based algorithm that converges to the Nash equilibrium in $O(T^{-1/2})$ iterations.

However, in real-world applications, approximating continuous-time dynamics can be challenging, and exact knowledge of cost and transition functions may not be feasible. Agents typically only receive bandit feedback, meaning they observe the cost and transitions associated with the states and actions they have visited. Meanwhile, the existing approaches only guarantee the convergence of the time average of the joint action profile, rather than the last-iterate convergence, the convergence of the joint action profile. Last-iterate convergence holds greater appeal as it offers a descriptive account of the evolution of players' overall behavior. In contrast, while the trajectory of players' joint action converge in the time-average sense, it may exhibit cycling, which is not suitable for practical deployment (Mertikopoulos, Papadimitriou, and Piliouras 2018). The following question thus arises.

How fast can discrete-time algorithms converge (in the last iterate) to a Nash equilibrium in GMFGs with bandit feedback?

[*] Authors are listed alphabetically, all authors are corresponding authors.

In this work, we focus on the mirror descent-based algorithm, which has been empirically verified to be successful in GMFGs (Pérolat et al. 2022; Zhang et al. 2023). In tabular monotone GMFGs (finite state and actions space) and under bandit feedback, we show a $O(T^{-1/4})$ last-iterate convergence rate. When the exact knowledge of cost and transitions is present, we show that the convergence rate can be improved to $O(T^{-1})$, matching the existing convergence rate in strongly convex (but not mean field) games. To address scenarios involving large or even infinite state spaces, we extend our analysis to linear GMFGs, where costs and transitions adhere to a linear structure. In this context, we achieve a last-iterate convergence rate of $O(T^{-1/5})$. We validate the effectiveness of the studied algorithm by empirically comparing them against the fictitious play in four different environments.

Related Works

Mean Field Game (MFG) To address the challenge of modeling a large number of agents in a game, the Mean Field Game (MFG) was proposed by (Caines, Huang, and Malhamé 2006; Lasry and Lions 2007). It considers the limit case of a continuous distribution of homogeneous agents (all anonymous and with symmetric interest) and reduces the problem to the characterization of the optimal behavior of a single representative agent. The classic approaches include the numerical approximation approach for partial differential equation (Achdou and Capuzzo-Dolcetta 2010; Achdou, Camilli, and Capuzzo-Dolcetta 2012; Achdou et al. 2020), and the more recent deep reinforcement learning approaches (Cui and Koeppl 2021a; Lauriere et al. 2022; Fabian, Cui, and Koeppl 2023).

Recent efforts also introduced the traditional fictitious play (FP) algorithm and combined it with machine learning techniques (Perrin et al. 2020). While FP achieves impressive results and is shown to be convergent (Geist et al. 2022), it is hard to scale due to its low computational efficiency, as it requires computing the best response at every iteration. To address this, the policy mirror descent algorithm is proposed, and its asymptotic convergence in continuous time is studied (Pérolat et al. 2022). Under a monotonicity assumption and regularization, the average iterate of the discrete-time mirror descent algorithm is shown to enjoy linear convergence under the tabular case and with function approximation (but with access to an approximation subroutine) (Zhang et al. 2023).

Graphon Mean Field Game (GMFG) To capture the heterogeneity among agents, the Graphon mean field game (GMFG) has been proposed by (Parise and Ozdaglar 2019), where the heterogeneous interaction between agents is described by graphon. By using a contraction condition, (Cui and Koeppl 2021b) proposed an algorithm to efficiently approximate the Nash equilibrium. The average iterate of the mirror descent algorithm is then shown to be convergent (Fabian, Cui, and Koeppl 2023) (asymptotically) and in finite time (Zhang et al. 2023). To our best knowledge, there is no last-iterate convergence guarantee of algorithms for (graphon) mean field games.

Last-iterate Convergence in Monotone (not Mean Field) Game In strongly convex games with full gradient feedback, the linear last-iterate convergence rate is established (Tseng 1995; Liang and Stokes 2019; Zhou et al. 2020). When the gradient feedback is with a zero-mean noise, (Jordan, Lin, and Zhou 2022) gave a $O(T^{-1})$ last-iterate convergence rate. When only bandit feedback is available, (Bervoets, Bravo, and Faure 2020) established an asymptotic convergence rate if the equilibrium is unique. Subsequently, (Bravo, Leslie, and Mertikopoulos 2018) improved this convergence rate of $O(T^{-1/3})$, while the proposed algorithm also ensured the no-regret property. Later works by (Lin et al. 2021) further improved the last-iterate convergence rate to $O(T^{-1/2})$ using the self-concordant barrier function.

Preliminary

We consider a GMFG defined as $\left(\mathcal{I}, \mathcal{S}, \mathcal{A}, \{P\}_{h\in[H]}, \{c\}_{h\in[H]}, \{W_h\}_{h\in[H]}, \mu_1, H\right)$ with infinitely many agents. Each agent corresponds to a point $\alpha \in \mathcal{I}$. Let ν be a positive measure on $\mathcal{I}$. The state and action space ($\mathcal{S}$ and $\mathcal{A}$) are the same for each agent. We further assume the state space is compact and the action space is finite. The interaction among agents at time h is captured through graphon W_h, a symmetric function such that $W_h(\alpha,\beta) = W_h(\beta,\alpha)$. The transition and reward of each agent are affected by the collective behavior of all other agents by an aggregate z. At time h, the aggregate for agent α is defined as $z_h^\alpha = \int_{\mathcal{I}} W_h(\alpha,\beta)\mu_h^\beta d\nu(\beta)$, where μ_h^β is the state distribution of agent β. We assume each agent has access to the aggregate z_h^α. We also let $\mu^{\mathcal{I}}(s) = \lim_{N\to\infty} \frac{\sum_{j=1}^N \mathbb{I}\{s_j=s\}}{N}$ denote the state distribution of all agents. On state s_h^α and when the agent takes action a_h^α, the state transits according to $s_{h+1}^\alpha \sim P_h(\cdot \mid s_h^\alpha, a_h^\alpha, z_h^\alpha)$. The agent α will also incur a cost of $c_h(s_h^\alpha, a_h^\alpha, z_h^\alpha)$.

Define the value functions as

$$
\begin{aligned}
&V_h^\alpha\left(s^\alpha, \pi^\alpha, \mu^{\mathcal{I}}\right) \\
&= \mathbb{E}_{\pi^\alpha}\left[\sum_{t=h}^{H} c_h\left(s_t^\alpha, a_t^\alpha, z_h^\alpha\left(\mu^{\mathcal{I}}\right)\right) \mid s_h^\alpha = s^\alpha\right], \qquad (1)\\
&Q_h^\alpha\left(s^\alpha, a^\alpha, \pi^\alpha, \mu^{\mathcal{I}}\right) \\
&= c_h\left(s^\alpha, a^\alpha, z_h^\alpha\right) + \mathbb{E}_{\pi^\alpha, P_h}\left[V_{h+1}^\alpha\left(s_{h+1}^\alpha, \pi^\alpha, \mu^{\mathcal{I}}\right)\right. \\
&\quad \left. \mid s_h^\alpha = s^\alpha, a_h^\alpha = a^\alpha\right]. \qquad (2)
\end{aligned}
$$

Following the standard settings studied in GMFG and Markov games, we investigate the convergence with the regularized value function, which enables faster convergence (Zhang et al. 2023; Cen, Wei, and Chi 2021; Shani, Efroni, and Mannor 2020). The λ-regularized value functions are

defined as

$$V_h^{\lambda,\alpha}\left(s^\alpha,\pi^\alpha,\mu^{\mathcal{I}}\right)=\mathbb{E}_{\pi^\alpha}\left[\sum_{t=h}^{H}c_h\left(s_t^\alpha,a_t^\alpha,z_t^\alpha\left(\mu^{\mathcal{I}}\right)\right)+\lambda\ln\pi_t^\alpha(a_t^\alpha\mid s_t^\alpha)\mid s_h^\alpha=s^\alpha\right], \tag{3}$$

$$Q_h^{\lambda,\alpha}\left(s^\alpha,a^\alpha,\pi^\alpha,\mu^{\mathcal{I}}\right)=c_h\left(s^\alpha,a^\alpha,z_h^\alpha\right)+\mathbb{E}_{\pi^\alpha,P_h}\left[V_{h+1}^{\lambda,\alpha}\left(s_{h+1}^\alpha,\pi^\alpha,\mu^{\mathcal{I}}\right)\mid s_h^\alpha=s^\alpha,a_h^\alpha=a^\alpha\right]. \tag{4}$$

Without loss of generality, we further assume the rewards are bounded between $[0,1]$. Then, $\left\|Q_h^{\lambda,\alpha}\left(s^\alpha,\cdot,\pi_t^\alpha,\mu_t^{\mathcal{I}}\right)\right\|_\infty\le H$, for any h,α,s. We define cumulative reward as

$$J^\alpha(\pi,\mu_1)=\mathbb{E}_{\mu_1}\left[V_1(s^\alpha,\pi,\mu)\right],$$
$$J^{\lambda,\alpha}(\pi,\mu_1)=\mathbb{E}_{\mu_1}\left[V_1^\lambda(s^\alpha,\pi,\mu)\right].$$

A common solution concept in GMFG is Nash equilibrium, which equilibrium state where no agent can gain in value by unilaterally changing its action. Formally, the Nash equilibrium is defined as follows.

Definition 1 (Nash equilibrium)**.** *An NE of the λ-regularized MFG is a pair $\left(\pi^{*,\mathcal{I}},\mu^{*,\mathcal{I}}\right)$ that satisfies*

- *Agent rationality:*

$$J^{\lambda,\alpha}\left(\pi^{*,\alpha},\mu^{*,\mathcal{I}}\right)=\min_{\widetilde{\pi}^\alpha\in\Pi^H}J^{\lambda,\alpha}\left(\widetilde{\pi}^\alpha,\mu^{*,\mathcal{I}}\right),$$

  for all $\alpha\in\mathcal{I}$ up to a zero measure set on $\mathcal{I}$ with respect to ν.
- *Distribution consistency: The distribution flow $\mu^{*,\mathcal{I}}$ is equal to the distribution flow induced by implementing the policy $\pi^{*,\mathcal{I}}$.*

To ensure the existence of a Nash equilibrium in a λ-regularized GMFG, we maintain the following assumptions of the game.

Assumption 1. *The GMFG satisfies*

- *The cost function and the transition function are continuous.*
- *The graphon is a continuous function.*

Remark 1. *The above model also includes games with finitely many players. For a finite graph, $\mathcal{G}=(V,E)$ with N nodes denoting the agents, and E denotes the set of edges that models the relationship between agents. We can partition a unit interval $[0,1]$ into N intervals, $I_1,\dots,I_N$ of equal length. Then we can let the graphon W assign a constant value on each square $I_i\times I_j$, $i,j\in V$. It is equal to one if there is an edge between i,j in $\mathcal{G}$, and zero otherwise. Although this is not continuous, one can smooth it so that it is continuous (Fabian, Cui, and Koeppl 2023).*

Theorem 1 (Theorem 4.4 (Zhang et al. 2023))**.** *Under Assumption 1, for all $\lambda\ge 0$, there exists a Nash equilibrium in a λ-regulairzed GMFG.*

To ensure the uniqueness of Nash equilibrium, we further maintain the following weakly monotonicity assumption of the game. The following condition is a generalization of the monotonicity condition in games (Lin et al. 2020, 2021; Duvocelle et al. 2023), and is commonly seen in literature in GMFG (Zhang et al. 2023).

Assumption 2 (Weakly monotone condition)**.** *A GMFG is said to be weakly monotone if for any $\rho^{\mathcal{I}},\tilde{\rho}^{\mathcal{I}}\in\Delta(\mathcal{S}\times\mathcal{A})^{\mathcal{I}}$ and their marginalizations on the states $\mu^{\mathcal{I}},\tilde{\mu}^{\mathcal{I}}\in\Delta(\mathcal{S})^{\mathcal{I}}$, we have*

$$\int_{\mathcal{I}}\sum_{a\in\mathcal{A}}\int_{\mathcal{S}}\left(\rho^\alpha(s,a)-\widetilde{\rho}^\alpha(s,a)\right)\left(c_h\left(s,a,z_h^\alpha\left(\mu^{\mathcal{I}}\right)\right)-c_h\left(s,a,z_h^\alpha\left(\widetilde{\mu}^{\mathcal{I}}\right)\right)\right)\mathrm{d}s\,\mathrm{d}\nu(\alpha)\ge 0,$$

for all t. It is strictly weakly monotone if the inequality is strict when $\rho^{\mathcal{I}}\neq\widetilde{\rho}^{\mathcal{I}}$.

When a λ-regularized GMFG admits Assumption 2, there exists a unique Nash equilibrium and satisfies the following property. An example of such a weakly monotone game is the multi-population predator-prey model described in (Pérolat et al. 2022).

Proposition 1 (Proposition 5.3 of (Zhang et al. 2023))**.** *If a λ-regularized GMFG satisfies the weakly monotone condition, then for any two policies $\pi^{\mathcal{I}},\widetilde{\pi}^{\mathcal{I}}\in\widetilde{\Pi}$ and their induced distribution flows $\mu^{\mathcal{I}},\widetilde{\mu}^{\mathcal{I}}\in\widetilde{\Delta}$, we have*

$$\int_{\mathcal{I}}J^{\lambda,\alpha}\left(\pi^\alpha,\mu^{\mathcal{I}}\right)+J^{\lambda,\alpha}\left(\widetilde{\pi}^\alpha,\tilde{\mu}^{\mathcal{I}}\right)-J^{\lambda,\alpha}\left(\widetilde{\pi}^\alpha,\mu^{\mathcal{I}}\right)-J^{\lambda,\alpha}\left(\pi^\alpha,\widetilde{\mu}^{\mathcal{I}}\right)\mathrm{d}\nu(\alpha)\ge 0.$$

If the λ-regularized GMFG satisfies the strictly weakly monotone condition, then the inequality is strict when $\pi^{\mathcal{I}}\neq\widetilde{\pi}^{\mathcal{I}}$.

Algorithms

In this section, we introduce our algorithm for solving Nash equilibrium in regularized GMFG. Our algorithm extends the celebrated mirror-descent algorithm, for which its efficiency in solving Nash equilibrium has been demonstrated. The last-iterate convergence properties of mirror descent has been investigated in many works (Cen, Wei, and Chi 2021; Lin et al. 2021; Cai et al. 2023; Duvocelle et al. 2023).

At each iteration t, the agent α execute $\{\pi_{t,h}^\alpha\}_{h=1}^H$ for H steps and receive costs $\{c_h(s_h^\alpha,a_h^\alpha,z_h^\alpha)\}_{h=1}^H$. Then, dependent on the information available, the agent computes a gradient $\hat{g}_{t,h}(s^\alpha,\cdot)$ and updates it with a mirror descent step. Algorithm 1 provides a summary of our algorithm.

We now discuss how one can construct the gradient estimator in a full information setting, tabular bandit feedback setting, and linear GMFG setting.

Tabular GMFG with full information feedback When the cost function and the transition kernel are known to the agent, the agent can set $\hat{g}_{t,h}(s^\alpha,\cdot)=Q_{t,h}^{\lambda,\alpha}\left(s^\alpha,\cdot,\pi_{t,h}^\alpha,\mu^{\mathcal{I}}\right)$, which can be computed via value iteration,

$$Q_{t,h}^{\lambda,\alpha}\left(s^\alpha,\cdot,\pi_t^\alpha,\mu^{\mathcal{I}}\right)=c_h\left(s^\alpha,\cdot,z_h^\alpha\right)+P_h\left(s^\alpha,\cdot,z_h^\alpha\right)V_{h+1}^{\lambda,\alpha},$$

Algorithm 1: Tabular online mirror descent for λ-regularized GMFG

Input: Learning rate $\{\eta_t\}_{t=1}^T$, regularization constant λ

for $t = 1, \dots, T$ **do**
  for $h = H, \dots, 1$ **do**
    Execute $\pi_{t,h}^\alpha$ and receive costs $c_h(s_h^\alpha, a_h^\alpha, z_h^\alpha)$;
  end
  for $h = H, \dots, 1$ **do**
    Compute gradient $\hat{g}_{t,h}(s_h^\alpha, \cdot)$ according to Equation (5), or (8);

$$\begin{aligned}&\pi_{t+1,h}^\alpha(\cdot \mid s_h^\alpha)\\ =\ &\arg\min_{\pi^\alpha} \eta_t \langle \hat{g}_{t,h}(s_h^\alpha, \cdot)\\ &+\lambda \log(\pi_t^\alpha), \pi^\alpha(\cdot \mid s_h^\alpha)\rangle\\ &+ D_{\mathrm{KL}}\left(\pi^\alpha(\cdot \mid s_h^\alpha), \pi_{t,h}^\alpha(\cdot \mid s_h^\alpha)\right)\end{aligned}$$

  end
end

then $V_{t,h}^{\lambda,\alpha}\left(s^\alpha, \pi_t^\alpha, \mu^{\mathcal{I}}\right) = \left\langle Q_h^{\lambda,\alpha}\left(s^\alpha, \cdot, \pi_{t,h}^\alpha, \mu^{\mathcal{I}}\right), \pi_{t,h}^\alpha(\cdot \mid s^\alpha)\right\rangle$.

Tabular GMFG with bandit feedback Under the tabular bandit feedback model, the agent does not have exact knowledge of the cost and transition kernel. Instead, they can only observe the cost corresponding to the state, action, and state distribution that they have visited. In this case, we compute the gradient as follows, which is similar to the gradient estimator on multi-agent tabular Markov games (Jin et al. 2022). Let $k = N_{t,h}(s)$ be the number of times state s is visited at step h up to time t. Then we first approximate the value function as

$$\begin{aligned}\hat{V}_h^{\lambda,\alpha}\left(s_h^\alpha, \pi_t^\alpha, \mu^{\mathcal{I}}\right) = \max\{&H+1-h,\\ &(1-\beta_k)\hat{V}_h^{\lambda,\alpha}\left(s_h^\alpha, \pi_t^\alpha, \mu^{\mathcal{I}}\right)\\ &+\beta_k\left(c_h(s_h^\alpha, a_h^\alpha, z_h^\alpha)\right.\\ &\left.+\hat{V}_{t,h+1}^{\lambda,\alpha}\left(s_{h+1}^\alpha, \pi_t^\alpha, \mu^{\mathcal{I}}\right)\right)\}.\end{aligned}$$

As we have no access to the cost associated with each action, we estimate the gradient with respect to all actions using importance sampling

$$\hat{g}_{t,h}(s_h^\alpha, a^\alpha) = \frac{\mathbb{I}\{a_h^\alpha = a^\alpha\}\left(c_h + \hat{V}_{t,h+1}^{\lambda,\alpha}\left(s_{h+1}^\alpha, \pi_t^\alpha, \mu^{\mathcal{I}}\right)\right)}{\pi_{t,h}^\alpha(a^\alpha \mid s_h^\alpha) + \gamma_t}. \tag{5}$$

To avoid unbounded gradient estimation and to encourage exploration, we add a γ_t factor. For the initialization step, we set $\hat{V}(s, \pi_0^\alpha, \mu^{\mathcal{I}}) = 0$, for all s.

Linear GMFG with bandit feedback When the state and action space are large, maintaining a tabular value function may become infeasible. Therefore, we consider learning within a linearly parameterized GMFG, which extends the tabular GMFG and accommodates the potential enormity of state and action spaces.

Definition 2 (Linear GMFG)**.** *A linear GMFG has a linearly structured transition*

$$P_h(\cdot \mid s_h^\alpha, a_h^\alpha, z_h^\alpha) = \theta_h^* \phi(s_h^\alpha, a_h^\alpha, z_h^\alpha),$$

$\forall h, s_h^\alpha \in \mathcal{S}, a_h^\alpha \in \mathcal{S}$ *where* ϕ *is a known feature mapping. Further, we assume*

1. $\sup_{s,a,z} \|\phi(s,a,z)\|_2 \le 1$*, and*
2. $\|v^\top \theta_h^*\| \le \sqrt{d}$*, for any* $\|v\|_\infty \le 1$ *and all* h*.*

Given the linear structure of the transition kernel, it remains to estimate θ_h accurately to compute the value function. Let $\delta_h(s^\alpha)$ be a one-hot vector that has zero everywhere except that the entry corresponding to s^α is one, and denote $\epsilon_h^\alpha = P_h(\cdot \mid s_h^\alpha, a_h^\alpha, z_h^\alpha) - \delta_h(s_{h+1}^\alpha)$. Conditioned on the history generated on all previous episodes up to episode t, $\mathcal{H}_{t,h}$, we have $\mathbb{E}[\epsilon_h^\alpha \mid \mathcal{H}_{t,h}] = 0$. Therefore δ acts as an unbiased estimate of P_h.

At iteration t, step h, we consider using all previous interactions $D_{t,h} = \{s_{j,h}^\alpha, a_{j,h}^\alpha, z_{j,h}^\alpha\}_{j=1}^{t-1}$to estimate the θ_h^*. Once we have an estimate $\hat{\theta}_h$, we can use value iteration to compute the value function. With the dataset $D_{t,h}$ one could estimate P_h using ridge regression

$$\begin{aligned}\hat{\theta}_{t,h} = \arg\min_{\theta_h} &\sum_{j=1}^{t-1} \left\|\theta_h \phi\left(s_{j,h}^\alpha, a_{j,h}^\alpha, z_{j,h}^\alpha\right) - \delta_h\left(s_{j,h+1}^\alpha\right)\right\|\\ &+ \|\theta_h\|^2,\end{aligned}$$

for which the closed-form solution is

$$\hat{\theta}_{t,h} = \sum_{j=1}^{t-1} \delta_h\left(s_{j,h+1}^\alpha\right) \phi\left(s_{j,h}^\alpha, a_{j,h}^\alpha, z_{j,h}^\alpha\right)^\top \left(\Lambda_t\right)^{-1}, \tag{6}$$

$$\Lambda_t = \sum_{j=1}^{t-1} \phi\left(s_{j,h}^\alpha, a_{j,h}^\alpha, z_{j,h}^\alpha\right) \phi\left(s_{j,h}^\alpha, a_{j,h}^\alpha, z_{j,h}^\alpha\right)^\top + I. \tag{7}$$

Using the estimate, we then update the value function using value iteration.

$$\begin{aligned}&\hat{Q}_{t,h}^{\lambda,\alpha}\left(s_h^\alpha, a_h^\alpha, \pi_t^\alpha, \mu^{\mathcal{I}}\right)\\ &= c_h\left(s_h^\alpha, a_h^\alpha, z_h^\alpha\right) + \phi(s^\alpha, a^\alpha)^\top \hat{\theta}_{t,h}^\top V_{h+1}^{\lambda,\alpha}\left(s_{h+1}^\alpha, \pi_t^\alpha, \mu^{\mathcal{I}}\right).\end{aligned}$$

Similar to the tabular case, as we only have information on the state and action we have visited, we estimate the gradient using importance sampling,

$$\hat{g}_{t,h}(s_h^\alpha, a^\alpha) = \frac{\mathbb{I}\{a_h^\alpha = a^\alpha\}\hat{Q}_{t,h}^{\lambda,\alpha}\left(s_h^\alpha, a_h^\alpha, \pi_t^\alpha, \mu^{\mathcal{I}}\right)}{\pi_{t,h}^\alpha(s_h^\alpha) + \gamma_t}. \tag{8}$$

Convergence Analysis

In this section, we present our main results on the last-iterate convergence to the Nash equilibrium in a regularized GMFG. To measure the distance to the equilibrium, we use the following convergence metric.

Convergence metric Define $D\left(\pi_t^{\mathcal{I}}\right)$ as

$$\int_{\mathcal{I}} \sum_{h=1}^{H} \mathbb{E}_{\mu_h^{*,\alpha}} \left[D_{\mathrm{KL}} \left(\pi^{*,\alpha}(\cdot \mid s_h^\alpha), \pi_t^\alpha(\cdot \mid s_h^\alpha) \right) \right] d\nu(\alpha) \,.$$

Note that this measures the weighted KL divergence between the policy computed and the Nash equilibrium, where the weights are the Nash equilibrium distribution flow μ^*. At equilibrium, the metric is zero. We also note that this is used in (Zhang et al. 2023).

Tabular GMFG with full information feedback When we have access to the exact cost and transition function, Theorem 2 shows we can converge linearly.

Theorem 2. *Let $\eta_t = t^{-1}$. We have*

$$D\left(\pi_{t+1}^{\mathcal{I}}\right) \leq \frac{H^3}{\lambda t} \,.$$

We first fix a tuple h, α, s_h^α. To obtain Theorem 2, we first use the mirror descent update rule to obtain the following relationship on $D_{\mathrm{KL}}\left(\pi^{*,\alpha}(\cdot \mid s_h^\alpha), \pi_{t+1,h}^\alpha(\cdot \mid s_h^\alpha)\right)$,

$$\begin{aligned}
&D_{\mathrm{KL}}\left(\pi^{*,\alpha}(\cdot \mid s_h^\alpha), \pi_{t+1,h}^\alpha(\cdot \mid s_h^\alpha)\right) \\
&\leq D_{\mathrm{KL}}\left(\pi^{*,\alpha}(\cdot \mid s_h^\alpha), \pi_{t,h}^\alpha(\cdot \mid s_h^\alpha)\right) + \frac{\eta_t^2 H^2}{2} \\
&\quad + \eta_t \left\langle \pi^{*,\alpha}(\cdot \mid s_h^\alpha) - \pi_{t,h}^\alpha(\cdot \mid s_h^\alpha), \right. \\
&\quad \left. Q_h^{\lambda,\alpha}\left(s_h^\alpha, \cdot, \pi_t^\beta, \mu^{\mathcal{I}}\right) + \lambda \log(\pi_{t,h}^\alpha(\cdot \mid s_h^\alpha)) \right\rangle \,.
\end{aligned}$$

We then use the third term and the monotonicity condition to obtain a recursion on $D_{\mathrm{KL}}\left(\pi^{*,\alpha}(\cdot \mid s_h^\alpha), \pi_{t+1,h}^\alpha(\cdot \mid s_h^\alpha)\right)$. The observation is that the λ parameter acts as a regularization for making the game more convex. Under λ regularization, notice that $Q_h^{\lambda,\alpha}\left(s_h^\alpha, \cdot, \pi^\alpha, \mu_t^{\mathcal{I}}\right) + \lambda \log(\pi^\alpha(\cdot \mid s_h^\alpha))$ is the gradient. In the following, we let the expectation to be a conditional expectation that condition on $s_1^\alpha = s^\alpha$. One can then show that

$$\begin{aligned}
&\mathbb{E}_{\pi^{*,\alpha}, \mu_t^{\mathcal{I}}} \left[\sum_{h=1}^{H} \left\langle Q_h^{\lambda,\alpha}\left(s_h^\alpha, \cdot, \pi^\alpha, \mu_t^{\mathcal{I}}\right) \right. \right. \\
&\quad \left. \left. + \lambda \log(\pi^\alpha(\cdot \mid s_h^\alpha)), \pi_h^{*,\alpha}(\cdot \mid s_h^\alpha) - \pi_h^\alpha(\cdot \mid s_h^\alpha) \right\rangle \right] \\
&\leq \left(V_1^{\lambda,\alpha}\left(s^\alpha, \pi^{*,\alpha}, \mu_t^{\mathcal{I}}\right) - V_1^{\lambda,\alpha}\left(s^\alpha, \pi^\alpha, \mu_t^{\mathcal{I}}\right) \right) \\
&\quad - \lambda \mathbb{E}_{\pi^{*,\alpha}, \mu_t^{\mathcal{I}}} \left[\sum_{h=1}^{H} D_{\mathrm{KL}}\left(\pi^{*,\alpha}(\cdot \mid s_h^\alpha), \pi_h^\alpha(\cdot \mid s_h^\alpha)\right) \right] .
\end{aligned}$$

Using the monotonicity condition, one can show that the first term is non-positive. Taking summation over H and integrating over all agents, we have

$$\begin{aligned}
&\int_{\mathcal{I}} \sum_{h=1}^{H} \mathbb{E}_{\mu_h^{*,\alpha}} \left[D_{\mathrm{KL}}\left(\pi^{*,\alpha}(\cdot \mid s_h^\alpha), \pi_{t+1,h}^\alpha(\cdot \mid s_h^\alpha)\right) \right] d\nu(\alpha) \\
&\leq (1 - \eta_t \lambda) \int_{\mathcal{I}} \sum_{h=1}^{H} \mathbb{E}_{\mu_h^{*,\alpha}} \left[D_{\mathrm{KL}}\left(\pi^{*,\alpha}(\cdot \mid s_h^\alpha), \right. \right. \\
&\quad \left. \left. \pi_{t,h}^\alpha(\cdot \mid s_h^\alpha)\right) \right] d\nu(\alpha) + \eta_t^2 H^3 \,.
\end{aligned}$$

The λ-regularization can also be interpreted as it regularize the game to be strongly convex with respect to KL divergence. As a result, one can anticipate the algorithm's convergence rate to be akin to established methods for strongly monotone games, often converging at a linear rate (Lin et al. 2020; Cen, Wei, and Chi 2021; Jordan, Lin, and Zhou 2022).

Tabular GMFG with bandit feedback We now consider the case where we only observe the cost $c_h(s, a, z)$ for the state, action, and state distribution that we have visited. In this case, we use importance sampling with implicit exploration (Eq.5) to estimate the gradient. We show that our algorithm then achieves the following last-iterate guarantee.

Theorem 3. *Take $\eta_t = \frac{1}{t^{3/4}}$, $\gamma = \frac{1}{t^{1/4}}$, $\beta_t = \frac{H+1}{H+t}$. We have $D\left(\pi_{t+1}^{\mathcal{I}}\right)$ be upper bounded by*

$$\tilde{O}\left(\frac{A \log(t)}{t^{3/4}} + \frac{\sqrt{\log(A/\delta)}}{t^{3/4}} + \frac{H^3}{t^{1/4}} + \frac{\log(1/\delta)}{t^{1/2}} \right) ,$$

where $\tilde{O}$ hides the logarithmic dependency on S, A, H, T.

We first fix a tuple h, α, s_h^α. Then, similar to Theorem 2, using the mirror descent update, we can obtain the following relation on $D_{\mathrm{KL}}\left(\pi^{*,\alpha}(\cdot \mid s_h^\alpha), \pi_{t+1,h}^\alpha(\cdot \mid s_h^\alpha)\right)$.

$$\begin{aligned}
&D_{\mathrm{KL}}\left(\pi^{*,\alpha}(\cdot \mid s_h^\alpha), \pi_{t+1,h}^\alpha(\cdot \mid s_h^\alpha)\right) \\
&\leq D_{\mathrm{KL}}\left(\pi^{*,\alpha}(\cdot \mid s_h^\alpha), \pi_{t,h}^\alpha(\cdot \mid s_h^\alpha)\right) + \frac{\eta_t^2 H^2}{2\gamma_t^2} \\
&\quad + \eta_t \left\langle \pi^{*,\alpha}(\cdot \mid s_h^\alpha) - \pi_{t,h}^\alpha(\cdot \mid s_h^\alpha), Q_h^{\lambda,\alpha}\left(s_h^\alpha, \cdot, \pi_t^\beta, \mu^{\mathcal{I}}\right) \right. \\
&\quad \left. + \lambda \log(\pi_{t,h}^\alpha(\cdot \mid s_h^\alpha)) \right\rangle \\
&\quad + \eta_t \left\langle \pi^{*,\alpha}(\cdot \mid s_h^\alpha) - \pi_{t,h}^\alpha(\cdot \mid s_h^\alpha), \hat{g}_{t,h}(s_h^\alpha, \cdot) \right. \\
&\quad \left. - Q_h^{\lambda,\alpha}\left(s_h^\alpha, \cdot, \pi_t^\beta, \mu^{\mathcal{I}}\right) \right\rangle \,.
\end{aligned}$$

As we use gradient estimation $\hat{g}_{t,h}$ instead of the exact gradient, we would need to characterize the estimation error

$$\hat{g}_{t,h}(s_h^\alpha, \cdot) - Q_h^{\lambda,\alpha}\left(s_h^\alpha, \cdot, \pi_t^\beta, \mu^{\mathcal{I}}\right)$$

to utilize our proof outline for Theorem 2. This gradient estimation error can be further refined to the estimation error of the value function

$$\hat{V}_{h+1}^{\lambda,\alpha}\left(s_{h+1}^\alpha, \pi_t^\alpha, \mu^{\mathcal{I}}\right) - V_{h+1}^{\lambda,\alpha}\left(s_{h+1}^\alpha, \pi_t^\alpha, \mu^{\mathcal{I}}\right) \,.$$

Using the update rule, we can upper bound the error as

$$\begin{aligned}
&\hat{V}_{t,h}^{\lambda,\alpha}\left(s_h^\alpha, \pi_t^\alpha, \mu^{\mathcal{I}}\right) - V_h^{\lambda,\alpha}\left(s_h^\alpha, \pi_t^\alpha, \mu^{\mathcal{I}}\right) \\
&\leq \sum_{i=1}^{k} \mathbb{E}_{\pi_{t^i}^\alpha} \left[\beta_k^i \left(\left(P_h - \hat{P}_h^{t_i}\right) V_{h+1}^{\lambda,\alpha}\left(\pi_{t^i}^\alpha, \mu^{\mathcal{I}}\right)\left(s_h^\alpha, a^\alpha\right) \right) \right] \\
&\quad + \beta_k^i \left(\hat{V}_{t^i,h+1}^{\lambda,\alpha}\left(s_{h+1}^{t^i}, \pi_{t^i}^\alpha, \mu^{\mathcal{I}}\right) - V_{h+1}^{\lambda,\alpha}\left(s_{h+1}^{t^i}, \pi_{t^i}^\alpha, \mu^{\mathcal{I}}\right) \right) .
\end{aligned}$$

Subsequently, using a martingale concentration inequality allows us to upper bound the first term at the order of $O\left(\sqrt{1/t}\right)$. Through an induction argument and leveraging the choice of learning rate β_t, we show

that $\hat{V}_{t,h}^{\lambda,\alpha}\left(s_h^\alpha, \pi_t^\alpha, \mu^{\mathcal{I}}\right) - V_h^{\lambda,\alpha}\left(s_h^\alpha, \pi_t^\alpha, \mu^{\mathcal{I}}\right)$ is also upper bounded by $O\left(\sqrt{1/t}\right)$.

Having characterized the estimation error, we then follow the same proof outline as outlined in Theorem 2 and obtain Theorem 3 by carefully choosing the parameters.

Linear GMFG with bandit feedback Under a linearly parameterized GMFG, we use ridge regression to estimate the model parameter θ_h^* (Equation (6)). Subsequently, we utilize value iteration alongside importance sampling with exploration to estimate the gradient (Equation (8)). Our algorithm provides the following convergence guarantee for the last iterate.

Theorem 4. *Take $\eta_t = \frac{1}{t^{4/5}}, \gamma_t = \frac{1}{t^{1/5}}$. We have*

$$D\left(\pi_{t+1}^{\mathcal{I}}\right) \leq O\left(\frac{H^3}{t^{3/5}} + \frac{\sqrt{\lambda} d^2 H^2}{t^{1/5}} + \frac{A}{t^{1/5}} + \sqrt{\frac{\log(A/\delta)}{t^{2/5}}}\right),$$

where $\tilde{O}$ hides the logarithmic dependency on S, A, H, T.

Similar to the analysis of Theorem 3, the key to deriving Theorem 4 lies in characterizing the estimation error $\hat{g}_{t,h}(s_h^\alpha, a^\alpha)$. This is then affected by the estimation error of our value function, and the estimation of θ_h. Leveraging a uniform convergence lemma (Lemma **??**), we can effectively upper bound the error as:

$$\begin{aligned}
&\hat{Q}_h^{\lambda,\alpha}\left(s_h^\alpha, a_h^\alpha, \pi_t^\beta, \mu^{\mathcal{I}}\right) - Q_h^{\lambda,\alpha}\left(s_h^\alpha, a_h^\alpha, \pi_t^\beta, \mu^{\mathcal{I}}\right) \\
&= \phi(s_h^\alpha, a_h^\alpha, z_h^\alpha)^\top \left(\hat{\theta}_{t,h} - \theta_h^*\right)^\top V_{h+1}^{\lambda,\alpha}\left(s_{h+1}^\alpha, \pi^\alpha, \mu^{\mathcal{I}}\right) \\
&\leq \tilde{O}\left(dH \|\phi(s_h^\alpha, a_h^\alpha, z_h^\alpha)\|_{\Lambda_{t,h}^{-1}}\right).
\end{aligned}$$

Employing the same argument as Theorem 2, we obtain a recursion on $D(\pi_t)$, which involves a summation of this estimation error

$$\int_{\mathcal{I}} \sum_{h=1}^{H} \sum_{k=1}^{t} \eta_k \omega_t^k \mathbb{E}_{\mu_h^{*,\alpha}}\left[dH \|\phi(s_{k,h}^\alpha, a_{k,h}^\alpha, z_{k,h}^\alpha)\|_{\Lambda_{t,h}^{-1}}\right] d\nu(\alpha).$$

Lastly, we bound the summation of this term using an elliptical potential lemma and recursively apply the relationship of $D(\pi_t)$ to obtain the final bound.

Experiments

To verify the effectiveness of our algorithm, we analyze it empirically on four environments, Predator Prey, Crowd Avoidance, Crowd modeling, and Periodic Aversion. To assess the performance of our algorithm, we use exploitability, which is defined as

$$\text{exploitability}(\pi) = \int_{\mathcal{I}} J(\pi, \mu^{\mathcal{I}}) - \min_{\pi'} J(\pi', \mu^{\mathcal{I}}).$$

To ensure reproducibility, we repeat each set of environments with 5 random seeds and present the result with one standard deviation.

Basline algorithm

For the baseline algorithm, we use fictitious play, a method known for providing a robust approximation of Nash equilibrium. Fictitious play iteratively computes the best response against the distribution induced by averaging past best responses. This method is known to perform well in various games, including mean-field games. However, since it needs to compute the best response every iteration, it is much more computationally heavier than our algorithm. We use the setup of fictitious play from (Perrin et al. 2020) and follow the implementation from OpenSpiel.

Enviornment set ups

For all the environments described below, we use the OpenSpiel implementation of the games.

Crowd Modeling The Crowd Modeling game also referred to as the beach bar process, presents a simplified rendition of the renowned Santa Fe bar problem (Greenwald, Mishra, and Parikh 1997). Following the dynamic modeling and cost functions outlined in (Perrin et al. 2020) (section 4.2), we conduct our experiment with 10 states and 3 actions. A beach bar is located in one of the states. In a scorching weather condition, agents aim to position themselves in proximity to the bar while avoiding excessively crowded areas.

Crowd Avoidance The crowd avoidance problem is a simple two-population game. The game has 7 states and 5 actions (stay, up, down, left, and right). The agents will receive a cost of 1 if they collide with each other, and a cost of 0 if otherwise. We follow the implementation from OpenSpiel. While trying to avoid congestion, the agent must also avoid the forbidden states.

Predator Prey We follow the setup described in section 5.4 of (Pérolat et al. 2022). This game features three populations and bears a close resemblance to the popular outdoor game for children, Hens-Foxes-Snakes. In this context, hens endeavor to capture snakes, snakes pursue foxes, and foxes are inclined to prey upon hens. Although the population sizes are predetermined, the cost structure incentivizes agents to chase the population they dominate.

Periodic Aversion The periodic aversion game was first introduced in (Almulla, Ferreira, and Gomes 2017) and served as is an approximation of a continuous space, continuous time model introduced to study ergodic MFG with an explicit solution. We follow the implementation from OpenSpiel and the description from (Elie et al. 2020). Each agent has a position on a torus $T = [0, 1]$ with periodic boundary conditions. The cost is then determined by a combination of the current position on the torus, the action, and the congestion of the agents.

Parameters and experiment configuration

For all of our experiments, we choose the learning rate to be $\eta_t = 0.1$ and the exploration rate $\gamma_t = 0.1$. We repeat the experiments with 5 different random seeds. We ran all experiments with a 10-core CPU, with 32 GB memory.

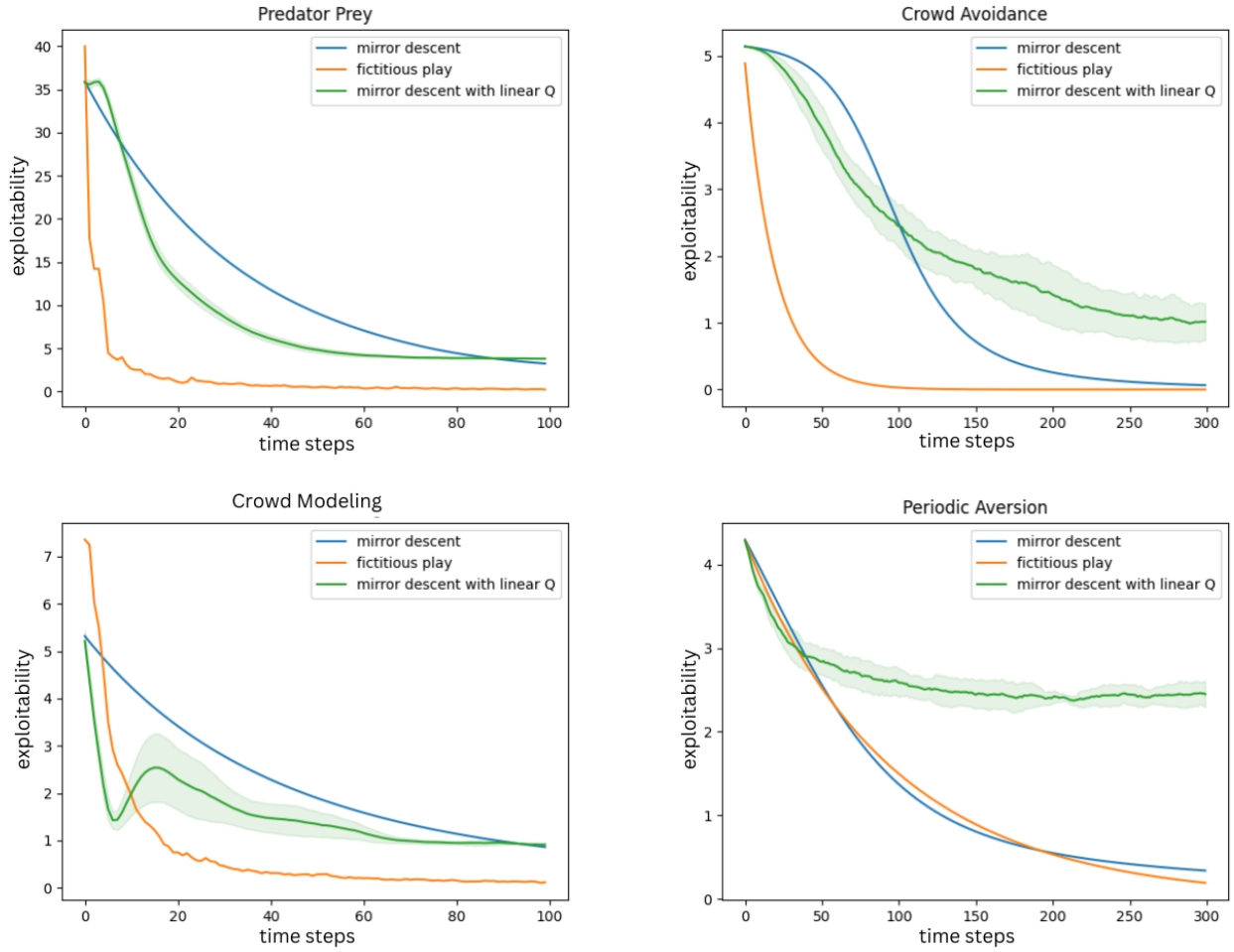

Figure 1: Experimental results for the mean field games described.

Experimental results

We show the results of the four environments described above. As evident in Figure 1, the mirror descent algorithm attains comparable performance as the fictitious play in two out of four environments, while enjoying much better computational complexity as it does not require the computation of the best response at each iteration. By approximating the value function with a linear structure, the mirror descent algorithm gains improvement in the results in two out of four environments.

Conclusion

In this work, we present the first last-iterate convergence rate for monotone GMFGs with a mirror-descent algorithm. In tabular monotone GMFGs and under bandit feedback, we obtain a $O(T^{-1/4})$ last-iterate convergence rate. Under access to the exact cost and transition functions, we improved the rate to $O(T^{-1})$. In linear GMFGs, we achieve a last-iterate convergence rate of $O(T^{-1/5})$ under bandit feedback. Our study improves the understanding of mean-field games and the commonly used algorithms by providing insights both theoretically and numerically.

Acknowledgements

We thank the reviewers and the area chairs for their constructive feedback. YY gratefully acknowledges funding support from NSERC and the Canada CIFAR AI Chairs program. Baoxiang Wang, Jing Dong are partially supported by the National Natural Science Foundation of China (62106213, 72394361), extended support projects from the Shenzhen Science and Technology Program, and support from the Longgang District Key Laboratory of Intelligent Digital Economy Security.

References

Achdou, Y.; Camilli, F.; and Capuzzo-Dolcetta, I. 2012. Mean field games: numerical methods for the planning problem. *SIAM Journal on Control and Optimization*, 50(1): 77–109.

Achdou, Y.; and Capuzzo-Dolcetta, I. 2010. Mean field

games: numerical methods. *SIAM Journal on Numerical Analysis*, 48(3): 1136–1162.

Achdou, Y.; Cardaliaguet, P.; Delarue, F.; Porretta, A.; Santambrogio, F.; Achdou, Y.; and Laurière, M. 2020. Mean field games and applications: Numerical aspects. *Mean Field Games: Cetraro, Italy 2019*, 249–307.

Almulla, N.; Ferreira, R.; and Gomes, D. 2017. Two numerical approaches to stationary mean-field games. *Dynamic Games and Applications*, 7: 657–682.

Aurell, A.; Carmona, R.; Dayanıklı, G.; and Laurière, M. 2022. Finite state graphon games with applications to epidemics. *Dynamic Games and Applications*, 12(1): 49–81.

Avena-Koenigsberger, A.; Misic, B.; and Sporns, O. 2018. Communication dynamics in complex brain networks. *Nature reviews neuroscience*, 19(1): 17–33.

Bakker, L.; Hare, W.; Khosravi, H.; and Ramadanovic, B. 2010. A social network model of investment behaviour in the stock market. *Physica A: Statistical Mechanics and its Applications*, 389(6): 1223–1229.

Bervoets, S.; Bravo, M.; and Faure, M. 2020. Learning with minimal information in continuous games. *Theoretical Economics*, 15(4): 1471–1508.

Bian, Y.-t.; Xu, L.; and Li, J.-s. 2016. Evolving dynamics of trading behavior based on coordination game in complex networks. *Physica A: Statistical Mechanics and its Applications*, 449: 281–290.

Bravo, M.; Leslie, D.; and Mertikopoulos, P. 2018. Bandit learning in concave N-person games. *Advances in Neural Information Processing Systems*.

Bullmore, E.; and Sporns, O. 2009. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature reviews neuroscience*, 10(3): 186–198.

Bullmore, E.; and Sporns, O. 2012. The economy of brain network organization. *Nature reviews neuroscience*, 13(5): 336–349.

Cai, Y.; Luo, H.; Wei, C.-Y.; and Zheng, W. 2023. Uncoupled and Convergent Learning in Two-Player Zero-Sum Markov Games. In *Advances in Neural Information Processing Systems*.

Caines, P. E.; Huang, M.; and Malhamé, R. P. 2006. Large population stochastic dynamic games: closed-loop McKean-Vlasov systems and the Nash certainty equivalence principle. *Communications in Information and Systems*, 6(3): 221–252.

Cen, S.; Wei, Y.; and Chi, Y. 2021. Fast policy extragradient methods for competitive games with entropy regularization. *Advances in Neural Information Processing Systems*.

Cui, K.; and Koeppl, H. 2021a. Approximately solving mean field games via entropy-regularized deep reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*.

Cui, K.; and Koeppl, H. 2021b. Learning Graphon Mean Field Games and Approximate Nash Equilibria. In *International Conference on Learning Representations*.

Cui, K.; Li, M.; Fabian, C.; and Koeppl, H. 2023. Scalable task-driven robotic swarm control via collision avoidance and learning mean-field control. In *International Conference on Robotics and Automation (ICRA)*.

Duvocelle, B.; Mertikopoulos, P.; Staudigl, M.; and Vermeulen, D. 2023. Multiagent online learning in time-varying games. *Mathematics of Operations Research*, 48(2): 914–941.

Elie, R.; Perolat, J.; Laurière, M.; Geist, M.; and Pietquin, O. 2020. On the convergence of model free learning in mean field games. In *The AAAI Conference on Artificial Intelligence*.

Fabian, C.; Cui, K.; and Koeppl, H. 2023. Learning sparse graphon mean field games. In *International Conference on Artificial Intelligence and Statistics*.

Geist, M.; Pérolat, J.; Laurière, M.; Elie, R.; Perrin, S.; Bachem, O.; Munos, R.; and Pietquin, O. 2022. Concave Utility Reinforcement Learning: The Mean-field Game Viewpoint. In *International Conference on Autonomous Agents and Multiagent Systems*.

Greenwald, A.; Mishra, B.; and Parikh, R. 1997. Learning in the Santa Fe bar problem.

Jin, C.; Liu, Q.; Wang, Y.; and Yu, T. 2022. V-Learning–A Simple, Efficient, Decentralized Algorithm for Multiagent RL. In *ICLR 2022 Workshop on Gamification and Multiagent Solutions*.

Jordan, M. I.; Lin, T.; and Zhou, Z. 2022. Adaptive, doubly optimal no-regret learning in games with gradient feedback. *Games with Gradient Feedback (September 8, 2022)*.

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

Lauriere, M.; Perrin, S.; Girgin, S.; Muller, P.; Jain, A.; Cabannes, T.; Piliouras, G.; Pérolat, J.; Elie, R.; Pietquin, O.; et al. 2022. Scalable deep reinforcement learning algorithms for mean field games. In *International Conference on Machine Learning*.

Liang, T.; and Stokes, J. 2019. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. In *International Conference on Artificial Intelligence and Statistics*.

Lin, T.; Zhou, Z.; Ba, W.; and Zhang, J. 2021. Doubly optimal no-regret online learning in strongly monotone games with bandit feedback. *arXiv preprint arXiv:2112.02856*.

Lin, T.; Zhou, Z.; Mertikopoulos, P.; and Jordan, M. 2020. Finite-time last-iterate convergence for multi-agent learning in games. In *International Conference on Machine Learning*.

Mertikopoulos, P.; Papadimitriou, C.; and Piliouras, G. 2018. Cycles in adversarial regularized learning. In *Symposium on discrete algorithms*.

Newman, M. E. 2002. Spread of epidemic disease on networks. *Physical review E*, 66(1): 016128.

Parise, F.; and Ozdaglar, A. 2019. Graphon games. In *Conference on Economics and Computation*.

Pastor-Satorras, R.; Castellano, C.; Van Mieghem, P.; and Vespignani, A. 2015. Epidemic processes in complex networks. *Reviews of modern physics*, 87(3): 925.

Pérolat, J.; Perrin, S.; Elie, R.; Laurière, M.; Piliouras, G.; Geist, M.; Tuyls, K.; and Pietquin, O. 2022. Scaling Mean Field Games by Online Mirror Descent. In *International Conference on Autonomous Agents and Multiagent Systems*.

Perrin, S.; Pérolat, J.; Laurière, M.; Geist, M.; Elie, R.; and Pietquin, O. 2020. Fictitious play for mean field games: Continuous time analysis and applications. *Advances in neural information processing systems*.

Shani, L.; Efroni, Y.; and Mannor, S. 2020. Adaptive trust region policy optimization: Global convergence and faster rates for regularized mdps. In *The AAAI Conference on Artificial Intelligence*.

Tangpi, L.; and Zhou, X. 2024. Optimal investment in a large population of competitive and heterogeneous agents. *Finance and Stochastics*, 1–55.

Tseng, P. 1995. On linear convergence of iterative methods for the variational inequality problem. *Journal of Computational and Applied Mathematics*, 60(1-2): 237–252.

Yang, J.; Ye, X.; Trivedi, R.; Xu, H.; and Zha, H. 2018. Learning Deep Mean Field Games for Modeling Large Population Behavior. In *International Conference on Learning Representations*.

Zhang, F.; Tan, V.; Wang, Z.; and Yang, Z. 2023. Learning Regularized Monotone Graphon Mean-Field Games. *Advances in Neural Information Processing Systems*.

Zhou, Z.; Mertikopoulos, P.; Bambos, N.; Boyd, S. P.; and Glynn, P. W. 2020. On the convergence of mirror descent beyond stochastic convex programming. *SIAM Journal on Optimization*, 30(1): 687–716.