Graphon Mean Field Games with a Representative Player: Analysis and Learning Algorithm

Fuzhong Zhou [1] **Chenyu Zhang** [2] **Xu Chen** [3] **Xuan Di** [3]

Abstract

We propose a discrete time graphon game formulation on continuous state and action spaces using a representative player to study stochastic games with heterogeneous interaction among agents. This formulation admits both philosophical and mathematical advantages, compared to a widely adopted formulation using a continuum of players. We prove the existence and uniqueness of the graphon equilibrium with mild assumptions, and show that this equilibrium can be used to construct an approximate solution for finite player game on networks, which is challenging to analyze and solve due to curse of dimensionality. An online oracle-free learning algorithm is developed to solve the equilibrium numerically, and sample complexity analysis is provided for its convergence.

1. Introduction

Many real-world applications, such as flocking (Perrin et al., 2021), epidemiology (Cui et al., 2022), and autonomous driving (Huang et al., 2020) involve multiagent systems, where agents optimize individual cumulative rewards by selecting sequential actions in an (in)finite horizon, while interacting strategically among one another. In discrete time, such finite player games form Markov games (Littman, 1994; Solan & Vieille, 2015; Yang et al., 2018b). At a Nash equilibrium (NE), nobody can improve her payoff by unilaterally switching her individual action policies. The NE is challenging to solve when the population size grows due to curse of dimensionality (Wang et al., 2020). To address such a challenge, mean field formulations are proposed to model players interacting with others only via an aggregate population, usually a population measure, instead of individual states or actions directly.

Mean field games (MFGs) (Huang et al., 2006; Lasry & Lions, 2007) is a type of mean field model which describes the limiting behavior of its corresponding finite player game as the number of players is large, and their analytical properties are now well-studied (Carmona & Delarue, 2018). The model is build upon the assumption that the interaction among players are homogeneous, in the sense that all players follow the same state distribution. As one interacts with the population only through the measure, all players react in a same manner and considering one representative is straightforward and appropriate.

As a generalization to MFGs, graphon mean field games (GMFGs or graphon games) are developed (Caines & Huang, 2021; Gao et al., 2021; Aurell et al., 2022; Cui & Koeppl, 2022; Tangpi & Zhou, 2024) to tackle the limiting behavior of finite player games with *heterogeneous* agents who interact *asymmetrically*, deemed as games on networks. In such network games, the interactions are given by a weighted graph (network), where each player is represented by a vertex and the interaction intensities among players are depicted by edge weights. Each player reacts to an interaction-weighted average of other individuals' empirical state measure, which is made precise in Section 3. As a limiting model, GMFGs models a continuum of players whose interaction intensities are given by a graphon $W \in L_1[0,1]^2$, which is a natural limit of finite graphs and can be deemed as a weighted graph on infinitely many vertices labeled by the continuum $[0,1]$. Rather than depending on states of some specific individuals, a player in the game reacts only to an average of the population state distribution, which is weighted as individuals of different types (different vertices in the graphon) exert heterogeneous influence on the current player.

GMFGs cover a wider range of real-world applications than MFGs, as it allows more flexibility with heterogeneous interaction. They are applicable to problems in finance, economics, and engineering, including for instance high-frequency trading, social opinion dynamics and autonomous vehicle driving. Because the equilibrium of GM-

[1]Department of Industrial Engineering and Operations Research, Columbia University, New York, NY, USA [2]Data Science Institute, Columbia University, New York, NY, USA [3]Department of Civil Engineering and Engineering Mechanics, Columbia University, New York, NY, USA. Correspondence to: Xuan Di <sharon.di@columbia.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

FGs may not be solved explicitly in general, recent years have seen a growing trend of using learning methods for equilibria. Compared to abundant studies on learning MFGs (Cardaliaguet & Hadikhanloo, 2017; Yang et al., 2018a; Guo et al., 2019; Elie et al., 2020; Perrin et al., 2020; 2021; 2022; Laurière et al., 2022; Chen et al., 2023a;b;c), learning for graphon games (Cui & Koeppl, 2022; Zhang et al., 2024b) is relatively understudied.

A major roadblock in learning GMFGs lies in the fact that there is no consensus on what a mathematically tractable formulation of GMFGs should be, since it is not straightforward to describe the limiting behavior (or a common population measure) of large number of heterogeneous players. There are mainly two types of formulations so far. The first type, also the widely adopted one, models a game for a continuum (uncountably infinite) of players with distinct types (Caines & Huang, 2021), so-called "continuum-player" games (Carmona et al., 2021). In this formulation, each player is assigned a controlled state process, which evolves independently of other individuals' state processes, and optimize her own reward function.

Unfortunately, this formulation suffers from limitations. Theoretically, the joint measurability of state dynamics with respect to the player types and randomness under the usual product space σ-algebra is not compatible with the independence of their evolution, which potentially poses challenges for the analytical investigation of solution properties (Appendix C.2); And practically, it is difficult to develop an algorithm that directly solves a system of optimal control problems for a continuum of players. Moreover, these studies could lack consistency between formulations (that model infinitely many players) and algorithms (that only sample a single representative agent).

To tackle the aforementioned challenges, a second kind of formulation (Lacker & Soret, 2023) refers to a generic representative player who represents all types of players while interacting with the aggregate population. While the state distribution for players of different type are different, it is possible to fit their label-state pairs into a common law on the product space of labels and state paths. This formulation is amenable to theoretical guarantees and ease the algorithmic design and implementation.

In this paper, we study discrete time graphon games of the second formulation with rigorous analysis and learning methods. We start from finite player games to motivate graphon games, which in turn provide approximate equilibria for finite games in dense interaction networks. Subsequently, GMFGs and graphon games always refer to the representative-player formulation, unless otherwise specified.

Related work. A detailed comparison with the most relevant studies on learning GMFGs is demonstrated in Appendix B. **Continuum-player formulation:** In discrete time regime, (Cui & Koeppl, 2022) showed the existence of Nash equilibrium and approximate equilibrium for finite player games under Lipschitz transition kernel and graphon, and (Zhang et al., 2024b) only showed the existence of Nash equilibrium for GMFGs with entropic regularization. Both studies assumed access to an oracle that returns the population dynamics, and the latter further assumes access to an action-value function oracle that returns the optimal policies. Under these assumptions, (Zhang et al., 2024b) provides a convergence rate of their algorithm, while (Cui & Koeppl, 2022) only shows the asymptotic convergence. In continuous time regime, (Caines & Huang, 2021) focused on finite networks where each vertex represents a population. (Gao et al., 2021; Aurell et al., 2022; Tangpi & Zhou, 2024) studied linear quadratic games, and the latter two adopted rich Fubini extensions to address the measurability issue. **Representative-player formulation:** As the establisher and the only work to the best of our knowledge, (Lacker & Soret, 2023) rigorously studied the equilibrium existence uniqueness and approximate equilibrium in continuous-time, with no discussion in algorithm implementation.

Contributions. Our major contributions are:

- We propose a general-purpose graphon game framework on continuous state and action space with one representative player that admits great technical and philosophical advantages over the continuum-player formulation in most prior work.
- We present an extensive self-contained analysis on the equilibrium including existence, uniqueness, and approximate equilibrium to network games with weaker assumptions and novel proof techniques.
- We give a comprehensive discussion and clarification on various aspects of graphon games, including but not limited to the MFG reformulation, overview on measurability issue, convergence of graph sequence, and fixed point iteration.
- We provide the first fully oracle-free online algorithm that numerically solves the equilibrium, and showed a sample complexity analysis for the ready-to-implement algorithm with assumptions equivalent to or weaker than prior work.
- We conduct abundant numerical experiments with assessments that demonstrate the validity of our algorithm design.

2. Preliminaries

2.1. Notation

Let E be any Polish space (complete separable metric topological space). We use $\mathcal{P}(E)$ to represent all the probability

measures on E equipped with the weak topology, with $\Rightarrow$ being the weak convergence. Let $\mathcal{M}_+(E)$ denote the space of nonnegative Borel measures of finite variation. Denote $\|\cdot\|_{\mathrm{TV}}$ the total variation norm. Given a random element X valued in E, let $\mathcal{L}(X) \in \mathcal{P}(E)$ be the probabilistic law (distribution) of X. For any $\mu \in \mathcal{P}(E)$, we write $X \sim \mu$ if $\mathcal{L}(X) = \mu$. For simplicity, we represent the integral with $\langle \mu, \phi \rangle = \int_E \phi d\mu$ for $\mu \in \mathcal{M}_+(E)$ and measurable ϕ.

Let $\mathcal{P}_{\mathrm{unif}}([0,1] \times E)$ denote a measure on product space $[0,1] \times E$ with uniform first marginal. We always consider E to be a regular space, and thus each element $\mu \in \mathcal{P}_{\mathrm{unif}}([0,1] \times E)$ admits a disintegration $du\mu^u(dx)$ where $\mu^u(dx)$ is a kernel $[0,1] \to E$ uniquely defined for Lebesgue almost every u.

2.2. Graphon

2.2.1. Definition

A graphon W is an L_1 integrable function : $[0,1]^2 \to \mathbb{R}_+$. It represents a graph with infinitely many vertices taking labels in $[0,1]$, and the edge weight connecting vertex u and v is given by $W(u,v)$. It is a natural notion for the limit of a sequence of graphs as the size of vertices grows.

Any finite graph can be expressed equivalently as a graphon: given any graph on $n \geq 1$ vertices with non-negative edge weights, it can be equivalently expressed as a matrix $\xi \in \mathbb{R}_+^{n \times n}$, where ξ_{ij} is the edge weight between vertex i and j. We define a *step graphon associated with* ξ, denoted as W_ξ on $[0,1]^2$ as below:

$$W_\xi(u,v) := \sum_{i,j=1}^{n} \xi_{ij} 1_{\{u \in I_i^n, v \in I_j^n\}}, \tag{1}$$

where the interval of $[0,1]$ is divided into n bins with the i^{th} bin as $I_i^n := [(i-1)/n, i/n), \forall i = 1, \ldots, n-1$; $I_n^n := [(n-1)/n, 1]$.

2.2.2. Graphon Operator

Given a Polish space E and any graphon W, the graphon operator W, which maps a measure in $\mathcal{P}_{\mathrm{unif}}([0,1] \times E)$ to a function $[0,1] \to \mathcal{M}_+(E)$, is defined as follows (Lacker & Soret, 2023): for any $m \in \mathcal{P}_{\mathrm{unif}}([0,1] \times E)$,

$$\mathsf{W}m(u) := \int_{[0,1] \times E} W(u,v) \delta_x m(dv, dx), \tag{2}$$

where δ_x is Dirac delta measure at x. Intuitively, let us assume m admits disintegration $m(du, dx) = dv m_u(dx)$, and W represents a graph with infinitely many vertices where each vertex $u \in [0,1]$ bears a random value on E with distribution m_u. Then, $\mathsf{W}m(u) = \int_{[0,1] \times E} W(u,v) \delta_x m_v(dx) dv$ is an average of the distributions of the random values over all vertices, weighted by edges with u as one end. Note that $\mathsf{W}m(u) \in \mathcal{M}_+(E)$ since the weighted average may no longer be a probability measure.

2.2.3. Strong Operator Topology

Now we define the convergence of graphons in strong operator topology. We abuse the notation by denoting the usual integral operator $\mathsf{W} : L_\infty[0,1] \to L_1[0,1]$,

$$\mathsf{W}\phi(u) := \int_{[0,1]} W(u,v)\phi(v)dv, \quad \forall \phi \in L_\infty[0,1], \tag{3}$$

and it should lead to no ambiguity as graphon operators and integral operators have different domains. We say a sequence of graphons W^n converges to a limit graphon W in the strong operator topology if for any $\phi \in L_\infty[0,1]$, $\|\mathsf{W}^n\phi - \mathsf{W}\phi\|_1 \to 0$, denoted as $W^n \to W$. Convergence in strong operator topology is usually weaker than convergence in cut norm, see Appendix A.5.

3. Finite Player Games

3.1. Game Formulation

Consider a game with $n \in \mathbb{N}_+$ players. Let $\xi \in \mathbb{R}_+^{n \times n}$ be an interaction matrix with nonnegative entries, where ξ_{ij} is the interaction influence of player j onto player i for $i, j \in [n]$. Let $T \in \mathbb{N}_+$ be terminal time of the game, and $\mathbb{T} := \{0, 1, 2, \ldots, T-1\}$. At each time t, denote $\mathbf{X}_t = (X_t^1, \ldots, X_t^n) \in (\mathbb{R}^d)^n$ the state dynamics of all the players, i.e., each player's state takes value in $\mathbb{R}^d$ for some fixed $d \geq 1$, and let $\mathcal{C} := (\mathbb{R}^d)^{T+1}$ be the space of state paths. For any $x \in \mathcal{C}$, write x_t the value of path at time t. The initial states $\mathbf{X}_0$ follow a vector of initial measures $\lambda = (\lambda^1, \ldots, \lambda^n) \in (\mathcal{P}(\mathbb{R}^d))^n$. At each time every player may choose an action from the action space A, and we assume that $A \subset \mathbb{R}^d$ is compact. Let $\mathcal{A}_n$ be the collection of all feedback policies $\mathbb{T} \times (\mathbb{R}^d)^n \to \mathcal{P}(A)$, and each player's action follows a policy from this collection. For any policy $\pi^i \in \mathcal{A}_n$ chosen by player i, the state process of player i evolves by a transition kernel $P : \mathbb{T} \times \mathbb{R}^d \times \mathcal{M}_+(\mathbb{R}^d) \times A \to \mathcal{P}(\mathbb{R}^d)$ as follows

$$X_0^i \sim \lambda^i,$$
$$a_t^i \sim \pi_t^i(\mathbf{X}_t), \qquad X_{t+1}^i \sim P_t(X_t^i, M_t^i, a_t^i),$$

for $i = 1, \ldots, n$, where

$$M^i := \frac{1}{n} \sum_{j=1}^{n} \xi_{ij} \delta_{X^j} \in \mathcal{M}_+(\mathcal{C})$$

is the weighted empirical neighborhood measure of player i, and M_t^i is the time t marginal of M^i. The measure is empirical as it is an average of the Dirac measures at the realizations; in particular, M^i is a random measure. For a general matrix ξ, M^i depicts the heterogeneous interaction: for a player i, the influence ξ_{ij} from player j is different from the influence ξ_{ir} from player r for $r \neq j$. In the special case where ξ is the adjacency matrix of an unweighted complete graph, i.e., ξ has 0 on the diagonal and 1 off

diagonal, the interactions become homogeneous, and M^i becomes the simple empirical measure of the states of all other players.

At a given time step, each player chooses an action according to her policy, and her state process X is a Markov decision process (MDP), which now depends not only on her current state and action, but also the empirical weighted neighborhood measure. Note that at each time t, the policy π^i of player i may depend on each of other players' state, while the transition law P should only depend on other players by an aggregation of their states, i.e., the empirical weighted neighborhood measure.

At each time all players receive a running reward according to some function $f : \mathbb{T} \times \mathbb{R}^d \times \mathcal{M}_+(\mathbb{R}^d) \times A \to \mathbb{R}$, and they receive a terminal reward at the terminal time T according to some function $g : \mathbb{R}^d \times \mathcal{M}_+(\mathbb{R}^d) \to \mathbb{R}$. The objective of player i is to maximize her expected accumulated reward

$$J^i(\boldsymbol{\pi}) := \mathbb{E}\left[\sum_{t\in\mathbb{T}} f_t(X_t^{\boldsymbol{\pi},i}, M_t^{\boldsymbol{\pi},i}, a_t^{\boldsymbol{\pi},i}) + g(X_T^{\boldsymbol{\pi},i}, M_T^{\boldsymbol{\pi},i})\right],$$

which is a function of the policy of all players $\boldsymbol{\pi} = (\pi^1, \dots, \pi^n) \in (\mathcal{A}_n)^n$. We write $X^{\boldsymbol{\pi},i}$, $M^{\boldsymbol{\pi},i}$ and $a^{\boldsymbol{\pi},i}$ to emphasize that the state dynamic of player i depends on $\boldsymbol{\pi}$.

Definition 3.1. For any nonnegative vector $\boldsymbol{\epsilon} = (\epsilon^1, \dots, \epsilon^n) \in \mathbb{R}^n_+$, an $\boldsymbol{\epsilon}$-equilibrium of an n-player game is defined as $\widehat{\boldsymbol{\pi}} = (\widehat{\pi}^1, \dots, \widehat{\pi}^n) \in (\mathcal{A}_n)^n$ such that for any i,

$$J^i(\widehat{\boldsymbol{\pi}}) \geq \sup_{\pi\in\mathcal{A}_n} J^i(\widehat{\boldsymbol{\pi}}^{-i}, \pi) - \epsilon_i, \tag{4}$$

where $(\widehat{\boldsymbol{\pi}}^{-i}, \pi)$ denotes the vector $\widehat{\boldsymbol{\pi}}$ with i^{th} coordinate replaced by π.

3.2. Mapping n-Player Indices onto a Continuous Label Space

This part serves as a transition from finite player game defined above, to its limiting system in the next section. In the finite n-player game, we map the index of agent $i \in \{1, \cdots, n\}$ onto a continuous label space $[0, 1]$, by assigning player i a label $u_i \in I_i^n := [(i-1)/n, i/n), \forall i = 1, \dots, n-1$ and $u_n \in I_n^n := [(n-1)/n, 1]$.

We demonstrate that the empirical weighted neighborhood measure M^i can be expressed in terms of the graphon operator. Let W_ξ be the step graphon (1) associated with interaction matrix ξ, then the interaction between player i and j can be expressed by $\xi_{ij} = W_\xi(u_i, u_j)$. Define the empirical label-state joint measure

$$S := \frac{1}{n}\sum_{i=1}^n \delta_{(u_i, X^i)} \in \mathcal{P}([0, 1] \times \mathcal{C}), \tag{5}$$

which is an empirical measure of the label-state pairs of all players. Then we have for $i = 1, \dots, n$,

$$M^i = \frac{1}{n}\sum_{j=1}^n \xi_{ij}\delta_{X^j} = \int W(u_i, v)\delta_x S(dv, dx) = \mathsf{W}_\xi S(u_i). \tag{6}$$

This demonstrates that the graphon operator is a generalization of the weighted neighborhood measure when there are infinitely many players: with W being the interaction among a continuum of players, and μ being their population label-state joint measure, $\mathsf{W}\mu(u)$ is the weighted neighborhood measure for the player of label $u \in [0, 1]$.

4. Representative-Player Graphon Games

4.1. Game Formulation

Given a graphon $W \in L^1_+[0, 1]^2$ representing the interaction intensity among a continuum-type of players labeled in $[0, 1]$, with $W(u, v)$ being the interaction intensity between player u and player v, we define the graphon game associated with W for a single representative player as follows. Let the state and action space be defined as in Section 3. Let $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ be a filtered probability space that supports an $\mathcal{F}_0$-measurable random variable U uniform on $[0, 1]$, and an adapted Markov process X valued in $\mathbb{R}^d$. We understand U as the label for the representative player, and X as her state dynamic. The initial label-state law of the representative player is given by $\lambda := \mathcal{L}(U, X_0) \in \mathcal{P}_{\text{unif}}([0, 1] \times \mathbb{R}^d)$. The term "label-state" always refer to the joint measure of a player's label and state pair (U, X). As in mean field games, we abstract all other players into a measure $\mu \in \mathcal{P}_{\text{unif}}([0, 1] \times \mathcal{C})$, i.e., each non-representative player should admit μ as her label-state joint measure, and the representative player only reacts to the population via this measure. Let μ be fixed. Let $\mu_t \in \mathcal{P}_{\text{unif}}([0, 1] \times \mathbb{R}^d)$ be the marginal of μ under image $(u, x) \mapsto (u, x_t)$.

Let $\mathcal{V}_U$ be the collection of all the open-loop policies, i.e., all the adapted process valued in $\mathcal{P}(A)$. Let $\mathcal{A}_U$ denotes the collection of all the closed-loop (Markovian) policies, i.e., measurable functions $\mathbb{T} \times [0, 1] \times \mathbb{R}^d \to \mathcal{P}(A)$. $\mathcal{A}_U$ is usually a proper subset of $\mathcal{V}_U$, unless the filtration is generated by U and X. For any $\pi \in \mathcal{V}_U$, the label-state pair (U, X) follows the transition dynamic $(U, X_0) \sim \lambda$ and at each $t \in \mathbb{T}$,

$$a_t \sim \pi_t, \qquad X_{t+1} \sim P_t(X_t, \mathsf{W}\mu_t(U), a_t),$$

for the same $\{P_t\}_{t\in\mathbb{T}}$ as in the finite player game introduced in Section 3. In words, the representative player is uniformly assigned a label U at time 0, and her later state transition depends on her current state, action and weighted neighborhood measure $\mathsf{W}\mu_t(U) \in \mathcal{P}(\mathbb{R}^d)$. Recall (6) in the finite player case, μ is now a generalization of S defined in (5) when there are infinitely many types of players. We may consider the disintegration $\mu(du, dx) = du\mu^u(dx)$, where du is the Lebesgue measure and $[0, 1] \ni u \mapsto \mu^u \in \mathcal{P}(\mathcal{C})$ is a probabilistic kernel.

Then $\mathsf{W}\mu(u) = \int_{[0,1]} W(u,v) \int_{\mathcal{C}} \delta_x \mu^v(dx) dv$. The inner integral is the path measure of player v, and the outer integral depicts an average of state distributions over all labels v, weighted by their interaction with the representative player when her label is $u \in [0,1]$.

Let $f : \mathbb{T} \times \mathbb{R}^d \times \mathcal{M}_+(\mathbb{R}^d) \times A \to \mathbb{R}$ be the running reward and $g : \mathbb{R}^d \times \mathcal{M}_+(\mathbb{R}^d) \to \mathbb{R}$ be the terminal reward. The objective of the representative player is to choose a policy $\pi \in \mathcal{V}_U$ to maximize

$$J_W(\mu, \pi) := \mathbb{E}\Big[\sum_{t\in\mathbb{T}} f_t(X_t^\pi, \mathsf{W}\mu_t(U), a_t^\pi) + g(X_T^\pi, \mathsf{W}\mu_T(U))\Big].$$

Note that the expectation is w.r.t. all random elements on $\mathcal{F}$, i.e., (U, X) and π, and we use X^π, a^π to emphasize that they depend on the policy π.

Definition 4.1. We say that the measure-policy pair $(\widehat{\mu}, \widehat{\pi}) \in \mathcal{P}_{\text{unif}}([0,1] \times \mathcal{C}) \times \mathcal{V}_U$ is a W-equilibrium if

$$J_W(\widehat{\mu}, \widehat{\pi}) = \sup_{\pi \in \mathcal{V}_U} J_W(\widehat{\mu}, \pi), \tag{7}$$

$$\widehat{\mu} = \mathcal{L}(U, X^{\widehat{\pi}}). \tag{8}$$

$\widehat{\mu}$, $\widehat{\pi}$ are called the equilibrium population measure and equilibrium optimal policy respectively.

Intuitively, the game is formulated for a representative player, while all other players are abstracted into a label-state joint measure μ. The representative player interacts with the population only through the weighted neighborhood measure $\mathsf{W}\mu(U)$, according to which she takes action to optimize her reward. The proposed graphon game is a strict generalization of MFGs, and it degenerates to an MFG when the graphon $W \equiv 1$. This is made precise in Appendix A.3. We will give a comprehensive comparison between our formulation and the continuum-player graphon game in Appendix C.

Remark 4.2. We define an infinite horizon version of graphon game with time-invariant dynamics and rewards in Appendix A.1. The analysis in the rest of this section could be easily adapted to the infinite horizon formulation by eliminating the time dependency of functions.

Remark 4.3 (GMFG as MFG with augmented state space). The graphon games defined here could be transformed into classical MFGs with an augmented state space, by imposing the label space $[0,1]$ as an additional dimension to the state space. However, this does not simplify the analysis or proof, and it is not appropriate to adapt existing MFG results directly (See Appendix A.2).

4.2. Existence of Equilibrium

Assumption 4.4. 1. The action space A is a compact subspace of $\mathbb{R}^d$.
2. The running rewards $\{f_t\}_{t\in\mathbb{T}}$ and terminal reward g are bounded and jointly continuous.
3. The initial distribution $\lambda \in \mathcal{P}_{\text{unif}}([0,1] \times \mathbb{R}^d)$ admits disintegration $\lambda(du, dx) = du\lambda_u(dx)$, and the following collection of measures is tight[1]:
$$\{\lambda_u\}_{u\in[0,1]} \subset \mathcal{P}(\mathbb{R}^d).$$
4. For each $t \in \mathbb{T}$, the following collection of measures is tight:
$$\zeta_t := \{P_t(x, m, a)\}_{(x,m,a)\in\mathbb{R}^d\times\mathcal{M}_+(\mathbb{R}^d)\times A} \subset \mathcal{P}(\mathbb{R}^d).$$
5. For each $t \in \mathbb{T}$, $P_t(x, m, \cdot)$ is continuous in A for every $(x, m) \in \mathbb{R}^d \times \mathcal{M}_+(\mathbb{R}^d)$.

An example case where Assumption 4.4(3, 4) are trivially satisfied is that there exists some compact subspace $\mathcal{X} \subset \mathbb{R}^d$ such that the collection ζ_t are uniformly supported on $\mathcal{X}$, i.e., the Markov process X takes values in the state space $\mathcal{X}$. Also note that we do not assume the graphon W to be continuous.

Theorem 4.5. *Suppose Assumption 4.4 holds. Then there exists a W-equilibrium $(\widehat{\mu}, \widehat{\pi})$. Moreover, the equilibrium optimal policy $\widehat{\pi}$ can be chosen to be a closed-loop policy.*

The theorem is proved with probabilistic compactification and Kakutani-Fan-Glicksberg fixed point theorem in Appendix E.

4.3. Uniqueness of Equilibrium

Assumption 4.6. 1. The state transition law P does not depend on the measure argument. Then it reads $P_t : \mathbb{R}^d \times A \to \mathcal{P}(\mathbb{R}^d)$ for $t \in \mathbb{T}$.
2. For each $t \in \mathbb{T}$, the running reward f_t is separable in the measure and action argument: there exists $f_t^1 : \mathbb{R}^d \times \mathcal{M}_+(\mathbb{R}^d) \to \mathbb{R}$ and $f_t^2 : \mathbb{R}^d \times A \to \mathbb{R}$ such that $f_t(x, m, a) = f_t^1(x, m) + f_t^2(x, a)$.
3. The optimal policy is unique. More specifically, for each $\mu \in \mathcal{P}_{\text{unif}}([0,1] \times \mathcal{C})$, the supremum $\sup_{\pi\in\mathcal{V}_U} J_W(\mu, \pi)$ is attained uniquely.
4. The functions f_t^1 and g satisfy the Larsy-Lions Monotonicity condition, in the following sense: for any $m_1, m_2 \in \mathcal{P}_{\text{unif}}([0,1] \times \mathbb{R}^d)$, and $t \in \mathbb{T}$,
$$\int_{[0,1]\times\mathbb{R}^d} \big(g(x, \mathsf{W}m_1(u)) - g(x, \mathsf{W}m_2(u))\big) (m_1 - m_2)(du, dx) \le 0,$$
$$\int_{[0,1]\times\mathbb{R}^d} \big(f_t^1(x, \mathsf{W}m_1(u)) - f_t^1(x, \mathsf{W}m_2(u))\big) (m_1 - m_2)(du, dx) \le 0.$$

Assumption 4.6 are the graphon game analogies to the classic uniqueness assumptions for mean field games, see for ex-

[1] Recall the definition of tightness: for arbitrary index set I and Polish space E, a collection of probability measures $\{P_i\}_{i\in I} \subset \mathcal{P}(E)$ is tight if for any $\epsilon > 0$, there exists some compact measurable subset $K \subset E$ such that $\inf_{i\in I} P_i(K) > 1 - \epsilon$.

ample, (Carmona & Delarue, 2018, Section 3.4) and (Lacker, 2018, Section 8.6).

Theorem 4.7. *Suppose Assumptions 4.4 and 4.6 hold. Then the graphon game admits a unique W-equilibrium.*

The proof follows a standard argument in MFG literature, see Appendix F.

4.4. Approximate Equilibrium for Finite Player Games

Let $\widehat{\pi} : \mathbb{T} \times [0,1] \times \mathbb{R}^d \to \mathcal{P}(A)$ be the equilibrium optimal closed-loop policy of the graphon game associated with graphon W, and we construct an n-player game policy from $\widehat{\pi}$ as follows. Assign player i the policy

$$\pi^{n,\mathbf{u}^n,i}(t, x_1, \dots, x_n) := \widehat{\pi}(t, u_i^n, x_i), \tag{9}$$

and $\boldsymbol{\pi}^{n,\mathbf{u}^n} := (\pi^{n,\mathbf{u}^n,1}, \dots, \pi^{n,\mathbf{u}^n,n}) \in (\mathcal{A}_n)^n$. Define

$$\epsilon_i^n(\mathbf{u}^n) := \sup_{\beta \in \mathcal{A}_n} J_i(\boldsymbol{\pi}^{n,\mathbf{u}^n,-i}, \beta) - J_i(\boldsymbol{\pi}^{n,\mathbf{u}^n}), \tag{10}$$

and $\boldsymbol{\epsilon}^n(\mathbf{u}^n) := (\epsilon_1^n(\mathbf{u}^n), \dots, \epsilon_n^n(\mathbf{u}^n))$. $\epsilon_i^n(\mathbf{u}^n)$ is the largest reward improvement player i could achieve by changing her own policy, when all other players follow policies $\boldsymbol{\pi}^{n,\mathbf{u}^n}$. By definition 3.1, $\boldsymbol{\pi}^{n,\mathbf{u}^n}$ is an $\boldsymbol{\epsilon}^n(\mathbf{u}^n)$-equilibrium of the n-player game. We need the following additional assumptions.

Assumption 4.8. 1. $\xi^n \in \mathbb{R}_+^{n \times n}$ is a sequence of matrix with 0 diagonals such that $W_{\xi^n} \to W$ in strong operator topology, and

$$\lim_{n\to\infty} \frac{1}{n^3} \sum_{i,j=1}^n (\xi_{ij}^n)^2 = 0. \tag{11}$$

2. For each $t \in \mathbb{T}$, the transition dynamic $P_t : \mathbb{R}^d \times \mathcal{M}_+(\mathbb{R}^d) \times A \to \mathcal{P}(\mathbb{R}^d)$ is jointly continuous for all $t \in \mathbb{T}$.

The next main result demonstrates that the n-player game policy $\boldsymbol{\pi}^{n,\mathbf{u}^n}$ constructed from the graphon game equilibrium optimal policy $\widehat{\pi}$ forms an approximate equilibrium, and it converges to the true equilibrium in an average sense as the number of players $n \to \infty$.

Theorem 4.9. *Suppose Assumptions 4.4 and 4.8 hold. For each $n \in \mathbb{N}_+$, let $\mathbf{U}^n := (U_1^n, \dots, U_n^n)$ where $U_i^n \sim \mathrm{unif}(I_i^n)$ and U_i^n is independent of U_j^n for $i \neq j$. Then we have*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\epsilon_i^n(\mathbf{U}^n)] = 0. \tag{12}$$

The proof is in Appendix G. Equation (12) can be equivalently written as $\epsilon_{I^n}^n(\mathbf{U}^n) \to 0$ in probability, where $I^n \sim \mathrm{unif}([n])$. Intuitively, for randomly assigned labels $\mathbf{U}^n$, and a player I^n uniformly chosen on $[n]$, the error is small. As the number of player $n \to \infty$, the collection of $\mathbf{U}^n$ and player label I^n such that the error cannot be controlled becomes a measure 0 set.

Remark 4.10. Equation (11) is a very mild graph denseness condition and is satisfied by many commonly-encountered finite graphs. The assumption $W_{\xi^n} \to W$ also poses denseness restrictions on the underlying graphs of interaction matrix ξ^n, as the existence of a graphon limit implicitly implies that the sequence of finite graphs are dense enough. We give some examples and a detailed discussion on dense graph sequences in Appendix A.5.

5. Learning Scheme and Sample Complexity

We now develop a scheme for learning the stationary equilibrium of infinite-horizon graphon games (Appendix A.1). Throughout the section we assume finite state space $\mathcal{X}$ and action space A.

5.1. Finite Classes of Label Space

To handle the continuous label space algorithmically, one generally needs function approximation techniques such as linear function approximation or neural networks, which is beyond the scope of this work. For the development and analysis of our algorithms, we discretize the label space $[0,1]$ into D classes of types of players $\mathcal{U} \subset [0,1]$ such that $|\mathcal{U}| = D < \infty$. We denote $\mathcal{U} := \{u_1, \dots, u_D\}$, and define projection mapping $\Pi_D : [0,1] \to \mathcal{U}$. Denote the inverse image $I_{u_d} := \Pi_D^{-1}(u_d) \subset [0,1]$. A simple example is the uniform quantization: $[0,1]$ is divided into D bins $\{I_d^D\}_{d=1}^D$, and Π_D maps each bin to its midpoint:

$$\Pi_D(u) = \sum_{i=1}^D \frac{2i-1}{2D} 1_{\{u \in I_i^D\}}. \tag{13}$$

As we are only able to learn measures on the finite discretization $\mathcal{U}$, we define $\boldsymbol{\Pi}_D : \mathcal{P}(\mathcal{X})^{\mathcal{U}} \to \mathcal{P}_{\mathrm{unif}}([0,1] \times \mathcal{X})$ as follows: for any $M = \{M^{u_d}\}_{d=1}^D$, $\boldsymbol{\Pi}_D M$ is the measure $\mathrm{Leb} \otimes \nu$, where ν is a probabilistic kernel given by $\nu(u) := \sum_{d=1}^D M^{u_d} 1_{\{u \in I_{u_d}\}}$.

5.2. Approximate Fixed-Point Iteration

Our learning scheme follows fixed-point iteration (FPI), which is widely used for learning (G)MFGs (Guo et al., 2019; Cui & Koeppl, 2022; Zhang et al., 2024b). An FPI represents an update of the game: given the population measure, the representative player first finds the optimal policy in reaction to this population, i.e., $\Gamma_1 : \mathcal{P}_{\mathrm{unif}}([0,1] \times \mathcal{X}) \to \mathcal{A}_U$, $\Gamma_1(\mu) := \mathrm{argmax}_{\pi \in \mathcal{A}_U} J_W(\mu, \pi)$. As everyone in the population reacts similarly, the population is then updated to the induced state distribution of the acquired policy, i.e., $\Gamma_2 : \mathcal{A}_U \times P_{\mathrm{unif}}([0,1] \times \mathcal{X}) \to \mathcal{P}_{\mathrm{unif}}([0,1] \times \mathcal{X})$, $\Gamma_2(\pi, \mu) := \mathcal{L}(U, X^\pi)$. Then, the FPI is given by $\Gamma(\mu) := \Gamma_2(\Gamma_1(\mu), \mu)$, and the equilibrium population measure $\widehat{\mu}$ satisfies $\widehat{\mu} = \Gamma(\widehat{\mu})$. However, the FPI operators can be hard to implement. As the environment (P and f) is unknown,

Γ_1 and Γ_2 are not directly accessible and need to be approximated. The general approximate FPI scheme is presented in Algorithm 1.

Algorithm 1 provides a general framework that can incorporate various learning algorithms for the two evaluation steps as subroutines, with (i) and (ii) approximating Γ_1 and Γ_2 respectively. If access to a state process generator (called an oracle) is assumed, we may generate the state variable under any control and population measure for arbitrary times, and Algorithm 1 recovers the algorithms used in prior work (Cui & Koeppl, 2022; Zhang et al., 2024b).

Algorithm 1 Approximate FPI for GMFGs

Initialize policy estimate $\{\pi_d^0\}_{d=1}^D$ and population estimate $\{M_d^0\}_{d=1}^D$ for all label classes $d \in [D]$
for $k \leftarrow 0$ to $K-1$ **do**
  for $d \leftarrow 1$ to D **do**
    (i) Evaluate approximate optimal policy π_d^{k+1} in reaction to M^k (ii) Evaluate approximate population measure M_d^{k+1} induced by π_d^{k+1}
  end for
end for
Return $\{\pi_d^K\}$ and $\{M_d^K\}$

We next provide the first non-asymptotic analysis for D-class FPI scheme given the following assumptions.

Assumption 5.1. 1. The transition kernel and reward function are uniformly L_P, L_f Lipschitz w.r.t. the measure argument respectively[2]:

$$\sup_{x,a} |f(x,m_1,a) - f(x,m_2,a)| \le L_f \|m_1 - m_2\|_{\mathrm{TV}},$$

$$\sup_{x,a} \|P(x,m_1,a) - P(x,m_2,a)\|_{\mathrm{TV}} \le L_P \|m_1 - m_2\|_{\mathrm{TV}}.$$

2. There exists L_d such that

$$\sup_{u,v\in[0,1]} |W(u,v) - W(\Pi_D(u),v)| \le L_d/D.$$

3. The FPI operator Γ is a contraction mapping: there exists $\kappa \in (0,1)$ such that $\|\Gamma(\mu_1) - \Gamma(\mu_2)\|_{\mathrm{TV}} \le (1-\kappa)\|\mu_1 - \mu_2\|_{\mathrm{TV}}$ for any $\mu_1, \mu_2 \in \mathcal{P}([0,1]\times\mathcal{X})$.

Assumption 5.1(2) ensures label classes $\mathcal{U}$ are a good approximation of the label space $[0,1]$. An example that satisfies this is the uniform quantization Π_D in (13) if the graphon is Lipschitz continuous in the first argument. The contraction mapping along with the Lipschitzness assumptions are limited but unfortunately necessary in the complexity analysis. We give a brief discussion on this assumption and different types of fixed-point theorems in Appendix A.6.

Suppose Assumption 5.1 holds, Algorithm 1 with exact evaluation steps needs at most $D = O(\kappa^{-1}\epsilon^{-1})$ classes and $K = O(\kappa^{-1}\log\epsilon^{-1})$ iterations to achieve an ϵ-approximate equilibrium. This claim is formalized in Theorem 5.4.

[2]Finite signed measures on finite space can be equivalently expressed as a vector, and the total variation norm is equivalent to the ℓ_1 norm.

5.3. Online Oracle-Free Learning

An oracle is defined to be a state process generator that could return the distribution of a player's next state, or a device that could collect the next state for a large number of players at the same time regardless of communication costs and asynchronous delays. An oracle-free algorithm (Angiuli et al., 2022) is one that does not involve an oracle. In the following, we present an online oracle-free subroutine for the approximate evaluation steps in Algorithm 1 by specifying a concrete implementation of the two evaluation steps (i) and (ii). Specifically, we use SARSA (Sutton & Barto, 2018), a value-based reinforcement learning method, for policy estimation, and Markov chain Monte Carlo (MCMC) for population estimation.

For policy estimation, we maintain a Q-function: $\mathcal{U}\times\mathcal{X}\times A \to \mathbb{R}$, with entry $Q_d(x,a)$ estimating the expected return starting with the state-action pair (x,a) conditional on label being u_d. Let $\mathcal{Q}$ be the collection of all Q-functions. To obtain the policy from a Q-function, we assume access to a Lipschitz continuous policy operator $\Gamma_\pi : \mathcal{Q} \to \mathcal{A}_U$, i.e., for any $Q_1, Q_2 \in \mathcal{Q}$, there exists a constant L_π such that

$$\sup_{u,x} \|(\Gamma_\pi(Q_1) - \Gamma_\pi(Q_2))(u,x)\|_{\mathrm{TV}} \le L_\pi \|Q_1 - Q_2\|_2 . \quad (14)$$

An example policy operator satisfying (14) is the softmax function, with its temperature parameter controlling the constant L_π (Gao & Pavel, 2017). Given Γ_π, SARSA converges to the Q-function corresponding to the optimal policy in $\Gamma_\pi(\mathcal{Q}) \subset \mathcal{A}_U$ (Zou et al., 2019).

Remark 5.2. Utilizing a Lipschitz continuous policy operator, Γ_1 returns the optimal Q-function instead of a policy; and Assumption 5.1(3) can be relaxed to only requiring Γ_1 and Γ_2 to be Lipschitz continuous with constants L_1 and L_2. Then, we can choose a sufficiently smooth policy operator such that $L_\pi L_1 L_2 < 1$, making the FPI operator $\Gamma(\mu) \coloneqq \Gamma_2(\Gamma_\pi(\Gamma_1(\mu)), \mu)$ contractive.

For population estimation, we maintain an M-function: $\mathcal{U} \to \mathcal{P}(\mathcal{X})$, with entry M_d estimating the population measure of the representative player conditional on label u_d. Since $\mathcal{U}\times\mathcal{X}\times A$ is a finite space, both Q- and M-functions can be represented by tables. Being fully online, SARSA and MCMC can update the Q- and M-functions using the same online samples without the need of any oracle. Specifically, we execute H updates for the evaluation subroutine of Algorithm 1. At each step $\tau = 0, \dots, H-1$, the representative agent with label u_d at x_τ samples its action $a_\tau \sim \Gamma_\pi(Q_d^{k,\tau})$, reward $r_\tau = f(x_\tau, \mathbf{W\Pi}_D M^{k,0}(u_d), a_\tau)$, next state $x_{\tau+1} \sim P(x_\tau, \mathbf{W\Pi}_D M^{k,0}(u_d), a_\tau)$, and next action $a_{\tau+1} \sim \Gamma_\pi(Q_d^{k,\tau})$. Using these observations, the Q-

and M-functions are updated simultaneously as follows:

$$\begin{aligned} Q_d^{k,\tau+1}(x_\tau, a_\tau) \leftarrow & (1-\alpha_\tau) Q_d^{k,\tau}(x_\tau, a_\tau) \\ & + \alpha_\tau \left(r_\tau + \gamma Q_d^{k,\tau}(x_{\tau+1}, a_{\tau+1}) \right), \\ M_d^{k,\tau+1} \leftarrow & (1-\beta_\tau) M_d^{k,\tau} + \beta_\tau \delta_{x_{\tau+1}}, \end{aligned} \tag{15}$$

where the Q- and M-functions are indexed by the outer iteration k and the inner evaluation step t, and α_τ and β_τ are step sizes. Substituting the Q-function $Q_d^{k,\tau}$ with the optimal Q-function $Q^{\mu^{k,\tau}}$ associated with the population measure $\mu^{k,\tau} = \mathbf{\Pi}_D M^{k,\tau}$, we recover the FPI scheme in Algorithm 1. Substituting (i) and (ii) in Algorithm 1 with H updates using Equation (15), we obtain the first fully online algorithm for learning GMFGs. Notably, our method is oracle-free in the sense that we do not assume access to an optimal policy calculator or a state process generator. Additionally, in contrast to FPI-like methods in prior work where (i) and (ii) in Algorithm 1 are executed sequentially, Equation (15) updates both policy and population concurrently using the same samples, enhancing the sample efficiency. Algorithm 2 in Appendix H is an example of a concrete realization of the aforementioned ideas.

As our method is fully online, we need the following ergodicity assumption (Zou et al., 2019).

Assumption 5.3. For any $\pi \in \Gamma_\pi(\mathcal{Q})$ and $M \in \mathcal{P}(\mathcal{X})^{\mathcal{U}}$, the Markovian state dynamic is ergodic: there exists $\nu \in \mathcal{P}_{\text{unif}}(\mathcal{U} \times \mathcal{X})$ and $c_1 > 0$, $c_2 \in (0,1)$ such that

$$\sup_x \|\mathcal{L}(X_\tau \mid X_0 = x) - \nu\|_{\text{TV}} \le c_1 c_2^\tau,$$

where the dynamic of X is determined by policy π and neighborhood measure $\mathbf{W}\mathbf{\Pi}_D M$.

Finally, we give the sample complexity of our algorithm.

Theorem 5.4. *Let $\widehat{\mu}$ be the stationary equilibrium measure of the infinite horizon GMFG. Suppose Assumptions 5.1 and 5.3 hold. For any initial estimate $M^{0,0} \in \mathcal{P}(\mathcal{X})^{\mathcal{U}}$, Algorithm 1, combined with Equation (15) and step sizes $\alpha_\tau, \beta_\tau \asymp 1/\tau$, finds an ϵ-approximate equilibrium distribution $M^{K,H}$ such that $\mathbb{E}\|\mathbf{\Pi}_D M^{K,H} - \widehat{\mu}\|_{\text{TV}} \le \epsilon$, with the number of iteration being at most*

$$K = O\left(\kappa^{-1} \log \epsilon^{-1}\right), \quad D = O\left(\kappa^{-1} \epsilon^{-1}\right),$$
$$H = O\left(\kappa^{-3} \epsilon^{-3} \log \epsilon^{-1}\right),$$

giving a total sample complexity of $O\left(\kappa^{-5} \epsilon^{-4} \log^2 \epsilon^{-1}\right)$.

We present a paraphrase of Theorem 5.4 in Theorem H.3 that incorporates the dependence on constants in Assumptions 5.1 and 5.3. The proof as well as more details of our method are deferred to Appendix H.

6. Numerical Experiments

In this section, we apply our learning algorithm to three graphon game examples, namely, Flocking-, SIS- and Invest-Graphon. We first briefly introduce each game scenario, and present only the algorithm performance and graphon mean field equilibrium (GMFE) for Flocking-Graphon due to space limit. The problem formulation of each game is in Appendix I. The detailed numerical results are in Appendix J, including algorithm performance (e.g., exploitability, convergence) and visualizations for GMFE. All numerical experiments are conducted on Mac Air M2.

Flocking-Graphon The Flocking-Graphon game (Lacker & Soret, 2023) studies the flocking behavior in a system where each agent makes decisions on its velocity, which in turn determines its position. We consider the game with $\mathcal{X} = [0,1]$, and a time horizon $\mathbb{T} = [0,1]$, under proper discretization. An agent with label u and position x randomly picks its velocity at time t according to the policy $\pi_t(u,x) \in \mathcal{P}(A)$ for action space $A = [0,1]$. Each agent aims to minimize its own running cost determined by the velocity control and the agent's deviation from the population.

SIS-Graphon The SIS-Graphon game (Cui & Koeppl, 2022) models an epidemic scenario where agents can choose taking precautions to avoid being infected. The infected probability is determined by the agents' action (i.e., take precaution or not) and the number of infected neighbors.

Invest-Graphon In the Invest-Graphon game model (Cui & Koeppl, 2022), each firm aims to maximize its own profit, which is determined by the firm's investment strategies and other firms' product quality.

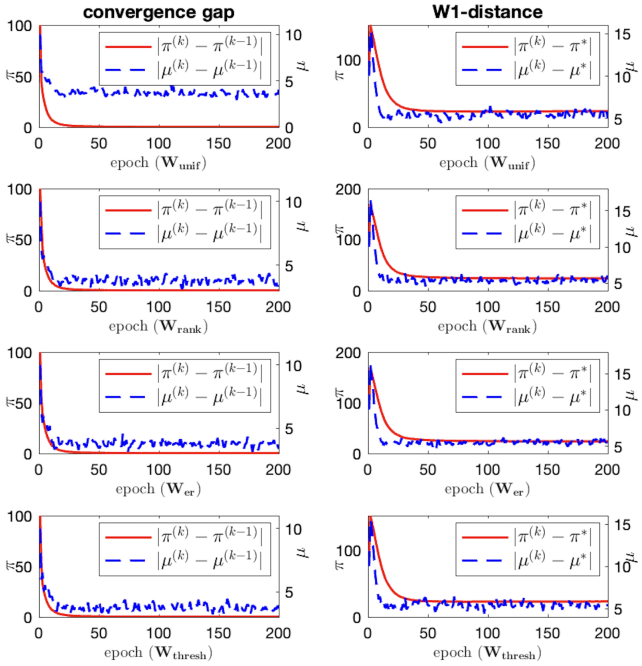

Figure 1. Algorithm performance (Flocking-Graphon)

We test four types of graphons: uniform attachment graphon ($W_{\text{unif}}(u,v) = 1 - \max(u,v)$), ranked attachment graphon ($W_{\text{rank}}(u,v) = 1 - uv$), Erdös-Rényi graphon ($W_{\text{er}}(u,v) = p$) and threshold graphon ($W_{\text{thresh}}(u,v) = \mathbf{1}_{u+v<1}$). Figure 1 demonstrates the algorithm performance to solve the Flocking-Graphon. The x-axis denotes the epoch index k. We visualize the convergence gaps $|\mu^{(k)} - \mu^{(k-1)}|$, $|\pi^{(k)} - \pi^{(k-1)}|$, and the W1-distances $|\mu^{(k)} - \mu^*|$, $|\pi^{(k)} - \pi^*|$, which measures the closeness between the benchmark solution (π^*, μ^*) and results at each epoch. The benchmark solution is obtained by the equivalent class method in (Cui & Koeppl, 2022). The results show that it takes around 50 epochs for our algorithm to converge. The convergence performance remains consistent for all four graphons.

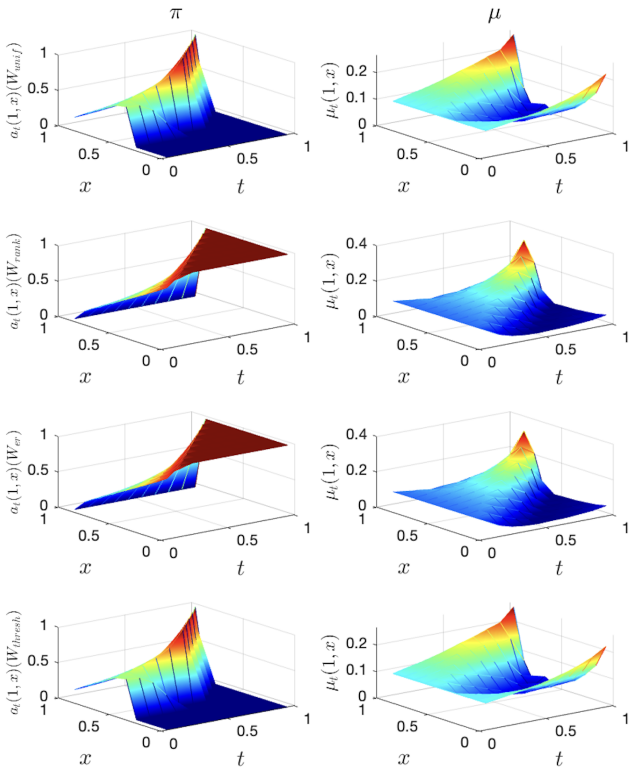

Figure 2. GMFE (Flocking-Graphon)

Figure 2 shows the obtained GMFE for Flocking-Graphon. We visualize the policy and state density of the agent with label $U = 1$ at equilibrium in a 3D plot. The x-axis denotes the space domain $\mathcal{X}$, and the y-axis is the time horizon $\mathbb{T}$. Agent with each label is initialized at $t = 0$ uniformly over $\mathcal{X}$. Note that the GMFE is time-dependent, which is obtained by adapting our learning algorithm to solve graphon games with finite horizons (See Algorithm 3 in Appendix H). The z-axis is the spatial-temporal velocity control $\alpha_t(1,x)$ and population density $\mu_t(1,x)$ of the agent with label 1. The numerical results show that GMFEs associated with W_{unif} and W_{thresh} are similar. The flock behavior, i.e., the phenomenon that players gather together without central planning at some location as time goes by, occurs at position $x = 0.6$ and the population density μ reaches a red peak around 0.35 with velocity around 0.2. When the agent's velocity reaches the maximum velocity $\alpha_{\max} = 1$ (dark red), the population quickly dissipates (dark blue) and no flock behavior occurs.

7. Conclusion

We offer a new general formulation of graphon games with one representative player in continuous state and action space. A comprehensive analysis on the equilibrium properties is proved with assumptions milder than previous work. We present a general approximate fixed-point iteration framework, and design an oracle-free algorithm along with the sample complexity analysis.

Impact Statement

This work is motivated by the theoretical challenges in the analysis of graphon games. As a generalization to mean field games, graphon games is capable of modeling heterogeneous interactions among gaming participants, and this flexibility allows it to cover a broader range of applications in finance, economics, engineering, including for example high-frequency trading, social opinion dynamics and autonomous vehicle driving. By addressing rigorously the technical issues faced by games on networks, this work proposes a conceptually and mathematically concise formulation. The analysis provides concrete theoretical foundation for the mathematical properties, on top of which the algorithms empower the solvability of the system. With the comprehensive and self-consistent technical analysis, this work is capable of modeling system with large amount of agents and remain computationally efficient.

References

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

Angiuli, A., Fouque, J.-P., and Laurière, M. Unified reinforcement Q-learning for mean field game and control problems. *Mathematics of Control, Signals, and Systems*, pp. 1–55, 2022.

Aurell, A., Carmona, R., and Laurière, M. Stochastic graphon games: II. the linear-quadratic case. *Applied Mathematics & Optimization*, 85(3):39, 2022.

Billingsley, P. *Probability and Measure*. John Wiley and Sons, 3rd edition, 1995.

Bris, P. L. and Poquet, C. A note on uniform in time mean-

field limit in graphs. *arXiv preprint arXiv:2211.11519*, 2022.

Brunick, G. and Shreve, S. Mimicking an Itô process by a solution of a stochastic differential equation. *The Annals of Applied Probability*, pp. 1584–1628, 2013.

Caines, P. E. and Huang, M. Graphon mean field games and their equations. *SIAM Journal on Control and Optimization*, 59(6):4373–4399, 2021.

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

Carmona, R. and Delarue, F. *Probabilistic Theory of Mean Field Games with Applications I*. Springer, 2018.

Carmona, R., Cooney, D. B., Graves, C. V., and Laurière, M. Stochastic graphon games: I. the static case. *Mathematics of Operations Research*, 47(1):750–778, 2021.

Chen, X., Fu, Y., Liu, S., and Di, X. Physics-informed neural operator for coupled forward-backward partial differential equations. In *ICML Workshop on the Synergy of Scientific and Machine Learning Modeling*, 2023a.

Chen, X., Liu, S., and Di, X. A hybrid framework of reinforcement learning and physics-informed deep learning for spatiotemporal mean field games. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pp. 1079–1087, 2023b.

Chen, X., Liu, S., and Di, X. Learning dual mean field games on graphs. In *Proceedings of the 26th European Conference on Artificial Intelligence, ECAI*, 2023c.

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

Cui, K. and Koeppl, H. Learning graphon mean field games and approximate Nash equilibria. In *International Conference on Learning Representations*, 2022.

Cui, K., KhudaBukhsh, W. R., and Koeppl, H. Hypergraphon mean field games. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 32(11), 2022.

Delattre, S., Giacomin, G., and Luçon, E. A note on dynamical models on random graphs and Fokker–Planck equations. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 165:785–798, 2016.

Elie, R., Perolat, J., Laurière, M., Geist, M., and Pietquin, O. On the convergence of model free learning in mean field games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 7143–7150, 2020.

Erdős, P. and Rényi, A. On random graphs. I. *Publicationes Mathematicae*, 6(3–4):290–297, 1959.

Erdős, P. and Rényi, A. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5:17–61, 1960.

Fabian, C., Cui, K., and Koeppl, H. Learning sparse graphon mean field games. In *International Conference on Artificial Intelligence and Statistics*, pp. 4486–4514. PMLR, 2023.

Gao, B. and Pavel, L. On the properties of the softmax function with application in game theory and reinforcement learning. *arXiv preprint arXiv:1704.00805*, 2017.

Gao, S., Tchuendom, R. F., and Caines, P. E. Linear quadratic graphon field games. *Communications in Information and Systems*, 21(3):341–369, 06 2021.

Guo, X., Hu, A., Xu, R., and Zhang, J. Learning mean-field games. *Advances in Neural Information Processing Systems*, 32, 2019.

Huang, K., Di, X., Du, Q., and Chen, X. A game-theoretic framework for autonomous vehicles velocity control: Bridging microscopic differential games and macroscopic mean field games. *Discrete and Continuous Dynamical Systems - Series B*, 25(12):4869–4903, 2020.

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

Jabin, P.-E., Poyato, D., and Soler, J. Mean-field limit of non-exchangeable systems. *arXiv preprint arXiv:2112.15406*, 2021.

Lacker, D. Mean field games via controlled martingale problems: Existence of Markovian equilibria. *Stochastic Processes and their Applications*, 23(4):2856–2894, 2015.

Lacker, D. Mean field games and interacting particle systems. *preprint*, 2018.

Lacker, D. and Soret, A. A label-state formulation of stochastic graphon games and approximate equilibria on large networks. *Mathematics of Operations Research*, 48 (4):1987–2018, 2023.

Lacker, D., Ramanan, K., and Wu, R. Local weak convergence for sparse networks of interacting processes. *The Annals of Applied Probability*, 33(2):843 – 888, 2023.

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

Laurière, M., Perrin, S., Girgin, S., Muller, P., Jain, A., Cabannes, T., Piliouras, G., Pérolat, J., Elie, R., Pietquin, O., and Geist, M. Scalable deep reinforcement learning algorithms for mean field games. In *International Conference on Machine Learning*, pp. 12078–12095. PMLR, 2022.

Littman, M. L. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pp. 157–163. Elsevier, 1994.

Lovász, L. *Large Networks and Graph Limits*, volume 60. American Mathematical Soc., 2012.

Mitrophanov, A. Y. Sensitivity and convergence of uniformly ergodic Markov chains. *Journal of Applied Probability*, 42(4):1003–1014, 2005.

Perrin, S., Pérolat, J., Laurière, M., Geist, M., Elie, R., and Pietquin, O. Fictitious play for mean field games: Continuous time analysis and applications. *Advances in Neural Information Processing Systems*, 33:13199–13213, 2020.

Perrin, S., Laurière, M., Pérolat, J., Geist, M., Élie, R., and Pietquin, O. Mean field games flock! The reinforcement learning way. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pp. 356–362, 2021.

Perrin, S., Laurière, M., Pérolat, J., Élie, R., Geist, M., and Pietquin, O. Generalization in mean field games by learning master policies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 9413–9421, 2022.

Solan, E. and Vieille, N. Stochastic games. *Proceedings of the National Academy of Sciences*, 112(45):13743–13746, 2015.

Sun, Y. The exact law of large numbers via Fubini extension and characterization of insurable risks. *Journal of Economic Theory*, 126:31–69, 2006.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

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

Wang, L., Yang, Z., and Wang, Z. Breaking the curse of many agents: Provable mean embedding Q-iteration for mean-field reinforcement learning. In *International Conference on Machine Learning*, pp. 10092–10103. PMLR, 2020.

Yang, J., Ye, X., Trivedi, R., Xu, H., and Zha, H. Deep mean field games for learning optimal behavior policy of large populations. In *International Conference on Learning Representations*, 2018a.

Yang, Y., Luo, R., Li, M., Zhou, M., Zhang, W., and Wang, J. Mean field multi-agent reinforcement learning. In *International conference on machine learning*, pp. 5571–5580. PMLR, 2018b.

Zhang, C., Chen, X., and Di, X. A single online agent can efficiently learn mean field games. *arXiv preprint arXiv:2405.03718*, 2024a.

Zhang, F., Tan, V., Wang, Z., and Yang, Z. Learning regularized monotone graphon mean-field games. *Advances in Neural Information Processing Systems*, 36, 2024b.

Zou, S., Xu, T., and Liang, Y. Finite-sample analysis for sarsa with linear function approximation. *Advances in Neural Information Processing Systems*, 32, 2019.

Organization of Appendix

The appendix is outlined as follows.

Appendix A is a discussion section serving as a supplement to the concepts in the main paper. The topics include: the infinite horizon version of graphon game formulation (Appendix A.1), formulating the graphon game into a mean field game with augmented state space (Appendix A.2), the degeneration of graphon game to mean field games with trivial graphon (Appendix A.3), time-variant interaction intensities (Appendix A.4), dense graph sequences and examples (Appendix A.5), fixed point theorems and the contraction mapping assumption (Appendix A.6).

Appendix B provides a list of tables where we compare our novelties and improvements to prior work.

In Appendix C, we define the continuum-player formulation (Appendix C.1) and discuss in detail the aforementioned measurability issue residing in continuum-player formulation in Appendix C.2. We then give a toy example in Appendix D to demonstrate the difference of the two formulations.

The following three sections are dedicated to the proof of analysis properties. The existence of equilibrium (Theorem 4.5) is proved in Appendix E. The uniqueness of equilibrium (Theorem 4.7) is proved in Appendix F. The approximate equilibrium (Theorem 4.9) is proved in Appendix G.

In Appendix H.1 we give a concrete realization of algorithm discussed in Section 5.3, and the rest of Appendix H is dedicated to the proof of the sample complexity of learning algorithms (Theorem 5.4). Finally, we give the detailed problem setups for the numerical examples in Appendix I, and show the numerical results in Appendix J.

A. Additional Discussion

A.1. Infinite Horizon Formulation

In this section we define the infinite horizon version of the representative-player graphon game, as appose the finite horizon version defined in Section 4.1. All analysis results in Section 4 regarding existence, uniqueness and approximate equilibrium holds by adjusting the assumptions accordingly.

Let the graphon $W \in L^1_+[0,1]^2$ be given and fixed. Let $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ be a filtered probability space that support an $\mathcal{F}_0$-measurable random variable U uniform on $[0,1]$, and a Markov process X valued in $\mathbb{R}^d$. We understand U as the label for the representative player, and X as her state dynamic. Let the flow of label-state joint measures be $\mu \in \mathcal{P}_{\text{unif}}([0,1] \times \mathcal{C})$, where the path space $\mathcal{C} = \prod_{i=0}^{\infty} \mathbb{R}^d$ is now a countable product of $\mathbb{R}^d$. $\mu_t \in \mathcal{P}_{\text{unif}}([0,1] \times \mathbb{R}^d)$ is the marginal under image $(u,x) \mapsto (u, x_t)$. Let the initial joint law $\lambda := \mathcal{L}(U, X_0) \in \mathcal{P}_{\text{unif}}([0,1] \times \mathbb{R}^d)$ be given.

We still let $\mathcal{A}_U$ denotes the collection of *time-variant* closed-loop (Markovian) policies $\mathbb{N}_+ \times [0,1] \times \mathbb{R}^d \to \mathcal{P}(A)$. For any $\pi \in \mathcal{A}_U$, (U, X) follows the transition dynamic

$$
\begin{aligned}
&(U, X_0) \sim \lambda, \\
&a_t \sim \pi_t(U, X_t), \qquad X_{t+1} \sim P(X_t, \mathsf{W}\mu_t(U), a_t),
\end{aligned}
$$

at any time $t \in \mathbb{N}_+$. Note that the transition law P is time-invariant. Let $f : \mathbb{R}^d \times \mathcal{M}_+(\mathbb{R}^d) \times A \to \mathbb{R}$ be the running reward and $\gamma \in (0,1)$ be a known discount factor. The objective of the representative player is to choose $\pi \in \mathcal{A}_U$ to maximize

$$
J_W(\mu, \pi) = \mathbb{E}\bigg[\sum_{t=0}^{\infty} \gamma^t f(X_t^\pi, \mathsf{W}\mu_t(U), a_t)\bigg].
$$

Definition A.1. We say that $(\widehat{\mu}, \widehat{\pi}) \in \mathcal{P}_{\text{unif}}([0,1] \times \mathcal{C}) \times \mathcal{A}_U$ is a W-equilibrium if

$$
\begin{aligned}
J_W(\widehat{\mu}, \widehat{\pi}) &= \sup_{\pi \in \mathcal{V}_U} J_W(\widehat{\mu}, \pi), \\
\widehat{\mu} &= \mathcal{L}(U, X^{\widehat{\pi}}) \quad .
\end{aligned}
$$

If we do not fix an initial distribution λ, we may define a stationary equilibrium which is time independent:

Definition A.2. We say that $(\widehat{\mu}, \widehat{\pi}) \in \mathcal{P}_{\text{unif}}([0,1] \times \mathbb{R}^d) \times \mathcal{A}_U$ is a stationary W-equilibrium if

$$J_W(\widehat{\mu}, \widehat{\pi}) = \sup_{\pi \in \mathcal{A}_U} J_W(\widehat{\mu}, \pi),$$
$$\widehat{\mu} = \mathcal{L}(U, X_t^{\widehat{\pi}}), \qquad \forall t \geq 0,$$

where $\mathcal{A}_U$ now denotes the collection of time-invariant closed-loop policies $[0,1] \times \mathbb{R}^d \to \mathcal{P}(A)$.

Note that we need an additional ergodicity assumption of the Markov chain to show the existence of stationary W-equilibrium with the same proof argument in Appendix E, i.e., the Markov chain admits a stationary distribution under any policy. A sufficient condition for ergodicity is given in Assumption 5.3.

A.2. Game with Augmented State Space

We give another view of a graphon game by reformulating it into a mean field game with an augmented state space. We view the label U as a coordinate of the state, and it remains at the same value a.s. Let $\overline{X}_t = \binom{U}{X_t} \in \mathbb{R}^{d+1}$, where the state process space is augmented by one additional dimension. Any fixed $\mu \in \mathcal{P}_{\text{unif}}([0,1] \times \mathcal{C})$ can now be equivalently regarded as an element in $\mathcal{P}(\mathcal{C}^{d+1})$ where $\mathcal{C}^{d+1} = (\mathbb{R}^{d+1})^{T+1}$ is the augmented path space. More specifically, we denote $\mu_t := \mu \circ (U, X_t)^{-1} \in \mathcal{P}(\mathbb{R}^{d+1})$, where $\circ$ is the pushforward. Given any graphon closed-loop policy $\pi \in \mathcal{A}_U$, define a mean field closed-loop policy $\overline{\pi}$ and a mean field Markovian transition law $\overline{P}$ as follows. For every $\overline{x} = \binom{u}{x} \in \mathbb{R}^{d+1}$,

$$\overline{\pi}_t(\overline{x})(da) := \pi_t(u, x)(da),$$
$$\overline{P}_t(\overline{x}, m, a)(d\overline{y}) := \delta_u(dv) P_t(x, \mathsf{W}m(u), a)(dy), \quad \forall \overline{y} = \begin{pmatrix} v \\ y \end{pmatrix},$$

respectively and let $\overline{\lambda}(d\overline{y}) := \lambda(dv, dy)$ for any $\overline{y} = \binom{v}{y}$ be the mean field initial condition. Then $\overline{X}$ satisfies the dynamic

$$\overline{X}_0 \sim \overline{\lambda},$$
$$a_t \sim \overline{\pi}_t(\overline{X}_t), \qquad \overline{X}_{t+1} \sim \overline{P}_t(\overline{x}, \mu_t, a_t).$$

Define similarly for every $\overline{x} = \binom{u}{x} \in \mathbb{R}^{d+1}$ the reward functions

$$\overline{f}_t(\overline{x}, m, a) := f_t(x, \mathsf{W}m(u), a),$$
$$\overline{g}(\overline{x}, m) := g(x, \mathsf{W}m(u)),$$

for all $t \in \mathbb{T}$. The objective is recast into

$$J(\overline{\pi}) := \mathbb{E}\left[\sum_{t \in \mathbb{T}} \overline{f}_t(\overline{X}_t^{\overline{\pi}}, \mu_t, a_t) + \overline{g}(\overline{X}_T^{\overline{\pi}}, \mu_T)\right].$$

Thus, we have obtained a classic mean field game problem associated with the new coefficients $\overline{\lambda}, \overline{P}_t, \overline{f}_t, \overline{g}$. Note that they are implicitly dependent on W.

However, it is worth noticing that in most of the proofs for graphon games, this translation into mean field games with augmented state space does not simplify the mathematical analysis, and it is not appropriate to adapt the mean field game results directly. There are two main reasons (Lacker & Soret, 2023):

Firstly, it requires the graphon $W \in L_+^1[0,1]^2$ to be continuous. To see this, recall that most of the results for classic mean field games assume the joint continuity of reward function, see e.g., (Carmona & Delarue, 2018; Lacker, 2018). In particular, $\overline{f}_t(\overline{x}, m, a) := f_t(x, \mathsf{W}m(u), a)$ is assumed to be continuous in the augmented state variable $\overline{x} = (u, x)^\top$. This requires that the graphon operator $\mathsf{W}\mu$ viewed as a function

$$[0,1] \ni u \mapsto \mathsf{W}\mu(u) \in \mathcal{M}_+(E)$$

should be continuous, for any $\mu \in \mathcal{P}_{\text{unif}}([0,1] \times E)$, which is satisfied by a continuous graphon. However, the graphon is in general not a continuous function. Indeed, many commonly encountered convergent graph sequence tends to a discontinuous graphon limit, see for instance examples in (Lovász, 2012, Section 11.4).

Second, in the analysis of approximate equilibrium, the model setting for the finite player game does not fit into this augmented state space framework. Consider an n-player game associated with interaction matrix $\xi \in \mathbb{R}_+^{n\times n}$, and assign player i the label $u_i \in I_i^n$. Recall the setting for finite player games in Section 3, the running reward can be written as $f_t(X_t^i, \mathsf{W}_\xi S_t(u_i), a_t^i)$, where S is the empirical label-state measure defined in (5). On the other hand, let $\overline{X}_t^i = \binom{u_i}{X_t^i}$, and the running cost of player i in the aformentioned augmented state space framework is

$$\overline{f}_t(\overline{X}^i, S_t, a_t^i) := f_t(X^i, \mathsf{W}S_t(u_i), a_t^i),$$

which is different from the original problem, as in the finite player game the graphon W needs to be replaced with the step graphon W_ξ. However, it is not possible to incorporate this change in the augmented state space framework. As a result the augmented state space transformation fails to provide an approximate equilibrium result, which is a strong justification of the reasonableness of graphon game formulation.

In the continuous time setting (Lacker & Soret, 2023), the augmented state space formulation provides an equivalent forward-backward PDE system for the graphon game, and thus provides another perspective to the problem formulation.

Actually the continuum-player graphon games may be transformed to a mean field game with augmented state space similarly, and many previous studies on continuum-player formulation relied on this (Cui & Koeppl, 2022; Zhang et al., 2024b) to show existence of equilibrium. However, they not only suffer from the two limitations mentioned above, but also encounter a critical measurability issue that representative-player formulation does not have, and this leads to difficulties in the proof. We will discuss this point in detail in Appendix C.

A.3. Degeneration to Mean-Field Games under a Trivial Graphon

When the graphon $W \equiv 1$, the interactions among players are symmetric, and we illustrate that our graphon game formulation degenerates to the classic mean field game.

Let the initial distribution λ, which is a product measure with the path space marginal λ°, be given. Fix a population measure $\mu \in \mathcal{P}_{\text{unif}}([0,1] \times \mathcal{C})$ and assume it takes the product measure form: $\mu(du, dx) = du \times \nu(dx)$ for some $\nu \in \mathcal{P}(\mathcal{C})$. The graphon operator applied on μ degenerates to ν:

$$\mathsf{W}\mu(u) = \int_{[0,1]\times\mathcal{C}} \delta_x \mu(dv, dx) = \int_{\mathcal{C}} \delta_x \nu(dx) = \nu, \qquad \forall u \in [0,1].$$

For any closed-loop policy $\pi \in \mathcal{A}_U$ that depends only on the state variable, i.e., π is a function $\mathbb{T} \times \mathbb{R}^d \to \mathcal{P}(A)$, (U, X) follows the transition dynamic: $U \sim \text{unif}[0,1]$, $X_0 \sim \lambda^\circ$ and

$$a_t \sim \pi_t, \qquad X_{t+1} \sim P_t(X_t, \nu_t, a_t).$$

Note that U and X are now independent. The objective of the representative player becomes

$$J_W(\nu, \pi) = \mathbb{E}\bigg[\sum_{t\in\mathbb{T}} f_t(X_t^\pi, \nu_t, a_t) + g(X_T^\pi, \nu_T)\bigg],$$

where the expectation is w.r.t X only, thanks to the independence between U and X. In this way our label-state graphon game formulation degenerates to a classic mean field game problem. The equilibrium measure and controls indeed does not depend on the label, and thus it is safe to restrict to μ being a product measure and policies not depending on label at the beginning.

A.4. Time-Variant Interaction Intensity

It is possible to consider time-variant interaction intensities in our framework when the time horizon is finite. In the definition of finite player games (Section 3), we may replace ξ with a sequence of matrix $\{\xi^t\}_{t=0}^T$, where ξ^t is the interaction intensity

of the n players at time t. The empirical weighted neighborhood measure of player i then becomes $M_t^i = \frac{1}{n}\sum_{i=1}^n \xi_{ij}^t \delta_{X_t^j}$, and it can be equivalently written as $W_{\xi^t} S_t(u_i)$, in the notation of Section 3.

In the graphon game setting (Section 4.1), we may work on a sequence of graphon $\{W_t\}_{t=0}^T$, where W_t is the interaction among a continuum-type of players at time t. Note that the sequence $\{W_t\}_{t=0}^T$ should be non-random. By replacing every $\mathsf{W}\mu_t$ with $\mathsf{W}_t\mu_t$, it is ready to check that the existence (Section 4.2) and uniqueness (Section 4.3) results still hold. As for the approximate equilibrium result (Section 4.4), we may change Assumption 4.8(1) into the following: $W_{\xi^{n,t}} \to W_t$, and

$$\lim_{n\to\infty} \sup_{t\in\mathbb{T}} \frac{1}{n^3} \sum_{i,j=1}^n (\xi_{ij}^{n,t})^2 = 0..$$

Then the approximate equilibrium result still holds. We present the main paper in terms of a time-invariant graphon W to avoid distraction from the main point we want to address.

A.5. Graph Sequence and the Convergence Assumption 4.8

Conceptually, a graph is dense if nearly every pair of vertices are connected by an edge. However, rigorously, the denseness of graph is ill-defined, and different results require different denseness conditions.

We first demonstrate that assumption (11) is indeed very mild. We may write $\mathrm{Tr}((\xi^n)^2) = \sum_{i,j=1}(\xi_{ij}^n)^2$ where $\mathrm{Tr}(\cdot)$ is the trace, and this is referred as second moment of square matrix. Here are several examples on commonly-encountered interaction matrix on networks.

Complete graph. Let $\xi_{ij}^n = 1$ for each $i \neq j$, and thus ξ^n is the adjacency matrix of a complete graph, and this recovers the mean field case where the players interact symmetrically. We have $W_{\xi^n} \equiv 1$ for all n, and thus $W_{\xi^n} \to W$ for $W \equiv 1$. We have

$$\frac{1}{n^3} \sum_{i,j=1}^n (\xi_{ij}^n)^2 \leq \frac{1}{n} \to 0.$$

Threshold graph. Consider a threshold graph on n vertices where vertex i and j are connected by an edge if $i + j < n$, and let $\xi_{ij}^n = 1_{i+j<n}$. It is easy to see that W_{ξ^n} converges in cut norm to a limit defined by $W(u, v) := 1_{u+v<1}$. It is ready to check that (11) is satisfied.

Random walk on graph. Consider a graph on n vertices where vertex i has degree d_i^n. Let $\xi_{ij}^n = \frac{n}{d_i^n} 1_{i\sim j}$, where $1_{i\sim j}$ is 1 if i and j are connected by an edge and 0 otherwise. Then ξ/n is a Markovian transition matrix of the random walk on the graph. We have

$$\frac{1}{n^3} \sum_{i,j=1}^n (\xi_{ij}^n)^2 = \frac{1}{n} \sum_{i,j=1}^n \frac{1}{(d_i^n)^2} 1_{i\sim j} = \frac{1}{n} \sum_{i=1}^n \frac{1}{d_i^n},$$

and the assumption holds if $\sum_{i=1}^n \frac{1}{d_i^n} \to 0$. Intuitively this means the average of degrees diverges. In particular, if $d_1^n = \cdots = d_n^n = d^n$, ξ^n becomes an interaction matrix on a d^n-regular graph, and we just need $\frac{1}{d_n} \to 0$, i.e., the degree d^n diverges to satisfy (11). However, not even every sequence of regular graphs has a graphon limit, and we will discuss this below.

Erdös-Rényi graph. Consider an Erdös-Rényi graph $G^n(p_n)$ (Erdős & Rényi, 1959) on n vertices, where every edge is connected with Bernoulli(p_n). Let $\xi_{ij}^n = \frac{1}{p_n} 1_{i\sim j}$, it is not hard to show that (11) holds in probability as long as $np_n \to \infty$. We understand np_n as the expected degree of any vertex, and this is an important quantity of Erdös-Rényi graphs that also implies connectivity (Erdős & Rényi, 1960). Moreover, when $p_n \to p$ for some $p \in (0, 1)$, $W_{\xi^n} \to W$ for $W \equiv 1$ in probability.

All examples mentioned above merely requires a diverging-average-degree type condition to be considered dense enough for our results to hold. These denseness conditions are attracting more and more awareness in the stochastic community and particularly in the studies of stochastic differential equation dynamics and heterogeneous propagation of chaos on networks (Delattre et al., 2016; Jabin et al., 2021; Bris & Poquet, 2022).

The assumption $W_{\xi^n} \to W$ is also a denseness condition as the existence of a graphon limit implicitly implies that the converging graph sequence is generally dense. Actually in the sparse setting, vertices in a local neighborhood interact strongly with each other and do not become negligible as the number of vertices goes to infinity (Lacker et al., 2023). The propagation of chaos results also fail in this regime. Nevertheless, not every dense graph sequence admits a graphon limit, since the sequence is also required to preserve similar network structures. This can be formalized by graph homomorphism (Lovász, 2012, Chapter 5).

It is worth noticing that a sequence of sparse graphs may converge if they are sampled from the limiting graphon. There are studies (Fabian et al., 2023) adopting this setting. However conceptually this means the finite player games are constructed from the graphon game, which is different from the view we take, that graphon games are motivated by finite player games.

Finally, we demonstrate that the convergence of graphon in strong operator topology is weaker than converging in other norm. Recall the definition of integral operator in (3):

$$\mathsf{W}\phi(u) := \int_{[0,1]} W(u,v)\phi(v)dv, \quad \forall \phi \in L_\infty[0,1],$$

which maps $L_\infty[0,1]$ to $L_1[0,1]$. The integral operator norm is given by $\|\mathsf{W}\|_{\infty\to 1} := \sup_{\|\phi\|_\infty \le 1} \|\mathsf{W}\phi\|_1$, where $\|\cdot\|_p$ is the L_p norm. It is known to be equivalent to the cut norm (Lovász, 2012, Lemma 8.11) by

$$\|W\|_\square \le \|\mathsf{W}\|_{\infty\to 1} \le 4\|W\|_\square \tag{16}$$

where the cut norm of a graphon is defined by

$$\|W\|_\square := \sup_{S,T\subset[0,1]} \left| \int_{S\times T} W(u,v)dudv \right|,$$

for measurable subsets S, T. It is immediate from (16) that the convergence in strong operator topology is weaker than converging in the cut norm. Indeed, W^n converging to W in L_1 also implies $W^n \to W$, see Lemma E.3.

A.6. Fixed Point Theorems and Strong Assumptions for Sample Complexity Analysis

There are two main stream fixed point theorems. The first type is based on contraction mapping, that if an operator is contraction in norm, then it admits a fixed point. An example of this category is the well-known Banach fixed point theorem. The second type, on the other hand, is usually based on the compact properties of the range space and operator, this includes Brouwer's fixed point theorem (compact, convex range space and continuous operators), Schauder's fixed point theorem (closed, bounded, convex range space and compact operators), and Kakutani-Fan-Glicksberg fixed point theorem for set-value functions, which is the one we will use in the proof of equilibrium existence (Appendix E). Compact-based fixed point theorems require weaker assumptions, but they fail to indicate how to find a fixed point rather than telling its theoretical existence.

Contraction based fixed point theorems usually need stronger assumptions, and the contraction in norm property is hard to verify practically. However, it provides clear approaches to find the fixed point when one exists: starting from an appropriate initial point, we may iteratively apply the operator and the result is guaranteed to converge to a fixed point. This iteration comes naturally from the forward-backward structure of games: players give their best response, and the population changes according to everyone's best response.

Many algorithms on mean field type games are based on the contraction fixed point theorem and iteration (Guo et al., 2019; Cui & Koeppl, 2022), including our work. These algorithms usually try to approximate the contraction mapping with estimations (since the environment is usually unknown) in order to demonstrate the convergences of algorithm. Thus, the contraction mapping and Lipschitzness assumptions are unfortunately necessary in the complexity analysis (see Assumption 5.1(3)), even though we do not make such assumptions in the pure mathematical analysis in Section 4.

There is a trade-off between the nice properties of the algorithm and a relatively weaker assumption. In our work, we focus on an algorithm that is online and oracle-free with convergence guarantees, which is proved based on existing complexity results of stochastic approximation methods that require stronger assumptions (see for example Lemma H.5 and Lemma H.6). On the other end of the trade-off, we may consider algorithms that require weaker assumptions, but do not enjoy theoretical convergence guarantees. Designing sharper quantitative convergence result for stochastic approximation methods is beyond the scope of this work.

B. Comparison of Related Work

In this section we give a comparison with related prior studies on discrete time and representative-player graphon games. The comparison is facilitated from four aspects: the problem formulation (Table 1), the analysis results showed (Table 2), the assumptions needed for existence and approximate equilibrium (Table 3), and the properties of the proposed learning algorithms (Table 4).

Table 1. Comparison of formulation

Reference	Player-type	Time domain	State space	Action space	Entropic regularization
(Cui & Koeppl, 2022)	Continuum	Discrete	Finite	Finite	✗
(Fabian et al., 2023)[3]	Continuum	Discrete	Finite	Finite	✗
(Zhang et al., 2024b)	Continuum	Discrete	Compact	Finite	✓
(Lacker & Soret, 2023)	Representative	Continuous	$\mathbb{R}^d$	Compact	✗
This paper	Representative	Discrete	$\mathbb{R}^d$	Compact	✗

Table 2. Comparison of analysis results

Reference	Existence	Uniqueness	Approximate Equilibrium
(Cui & Koeppl, 2022)	✓	✗	✓
(Fabian et al., 2023)	✓	✗	✓
(Zhang et al., 2024b)	✓	✓	✗
(Lacker & Soret, 2023)	✓	✓	✓
This paper	✓	✓	✓

Table 3. Comparison of analysis assumptions

Reference	**Existence assumptions**			**Approx. eqbm. assumptions**	
Perspective	Graphon	Transition kernel	Reward function	Transition kernel	Graphon convergence
(Cui & Koeppl, 2022)	Jointly Lipschitz	Jointly Lipschitz	Jointly Lipschitz	Jointly Lipschitz	In cut norm
(Fabian et al., 2023)	Jointly Lipschitz	Jointly Lipschitz	Jointly Lipschitz	Jointly Lipschitz	In cut norm
(Zhang et al., 2024b)	Jointly cont.	Jointly cont.	Jointly cont.	NA	NA
This paper	L_1-integrable	Cont. in action	Jointly cont.	Jointly cont.	In strong operator norm

Table 4. Comparison of algorithm analysis

Reference	Oracle-free	MFG criterion	Graphon	Complexity analysis
(Cui & Koeppl, 2022)	✓	Contractive	Block-wise jointly Lipschitz	✗
(Fabian et al., 2023)	✗	Monotone	Block-wise jointly Lipschitz	✗
(Zhang et al., 2024b)	✗	Monotone	Jointly Lipschitz	✓[4]
This paper	✓	Contractive	Assumption 5.1(2)	✓

As a comparison with the Lacker & Soret's pioneering work in representative graphon game (2023), the discrete time game model proposed in this paper is more realistic and strongly associated with the applications and lends itself well to stochastic algorithm design as existing algorithms and analytical results for stochastic approximation methods are mostly given on a discrete time domain.

In addition, the discrete time formulation, defined with transition kernels directly, covers a broader range of state dynamics. Consider a partition of the continuous-time domain $[0, T]$ into $N + 1$ time slices with a time step $\Delta := T/N$, and $t_k := k\Delta$, then the SDE can be discretized on the time slices $\{t_k\}_{k=0}^N$, resulting in a discrete time Markov process. In

[3]We note that (Fabian et al., 2023) considers a sparse type of graphon convergence with a factor that mitigates the sparseness, and adopts a sampling regime where the finite player interactions are sampled from a graphon. This is different from the other work where the graphon games are considered to be the limit model of finite player network games.

[4]The sample complexity in (Zhang et al., 2024b) is partial, as it only accounts for the backward procedure, i.e., the best response of players to a fixed population distribution, without discussing the complexity of obtaining an estimate of the induced population measure in the forward procedure.

other words, every state Ito's process of the form $dX_s = b(s, X_s, \alpha_s)ds + \sigma(s, X_s)dB_s$ can be discretized into the form $X_{k+1} \sim P_k(X_k, \mathsf{W}\mu_k(U), \alpha_k)$ for some Markovian transition kernel P_k. On the other hand, not every discrete time Markov process (with a general transition kernel P_k) has a continuous time analog in the form of an Ito's process. It is actually a trade-off: we sacrifice information on time domain by evaluating the state process only on discrete time slices rather than a complete path on $[0, T]$, but our framework may cover a broader range of possible state dynamics.

C. Comparing Representative-Player Games and Continuum-player Games

C.1. Continuum-Player Graphon Games

In this section we give a review on continuum-player graphon games in previous work. Consider a game with a continuum of players, labeled with $u \in [0, 1]$, and we assume the label space $[0, 1]$ is equipped with Borel-σ-algebra and Lebesgue measure. Each player u admits a state process X^u valued in $\mathbb{R}^d$. Fix a population measure, which is a collection $\mu = \{\mu^u\}_{u\in[0,1]}$, and it is usually assumed that $u \mapsto \mu^u$ is a probabilistic kernel, i.e., $u \mapsto \mu^u(B)$ is a measurable function for any Borel subset $B \subset \mathcal{C}$. Let $\mathcal{A}$ be the collection of all the feedback (closed-loop) policies $\mathbb{T} \times \mathbb{R}^d \to \mathcal{P}(A)$, and assume player u adopts a policy $\pi^u \in \mathcal{A}$. The state process follows

$$\begin{aligned} &X_0^u \sim \lambda^u, \\ &a_t^u \sim \pi_t^u(X_t^u), \qquad X_{t+1}^u \sim P_t(X_t^u, \mathsf{W}\mu_t(u), a_t), \end{aligned}$$

for some initial condition $\lambda^u \in \mathcal{P}(\mathbb{R}^d)$. Here we regard μ as a measure constructed by $du \times \mu^u(dx)$, where the assumption μ being a kernel come into place. It is also common to write $\mathsf{W}\mu_t(u)$ as $\int_{[0,1]} W(u, v)\mu_t^v dv$.

Note that all the players' state dynamics are independent, in the following sense: for every $u \in [0, 1]$, X^u is independent of X^v for every $v \in [0, 1]$. Indeed, this independence leads to a significant measurability issue that many proofs ignore, and we will give a detailed discussion in Appendix C.2.

Each player u aims to maximize an objective function

$$J^u(\mu, \pi^u) := \mathbb{E}\bigg[\sum_{t\in\mathbb{T}} f_t(X_t^{u,\pi^u}, \mathsf{W}\mu_t(u), a_t^u) + g(X_T^{u,\pi^u}, \mathsf{W}\mu_T(u))\bigg],$$

where we denote X^{u,π^u} to emphasize the process X^u is controlled by policy π^u. The equilibrium is defined as a pair $(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\pi}}) := (\{\widehat{\mu}^u\}_{u\in[0,1]}, \{\widehat{\pi}^u\}_{u\in[0,1]}) \in \mathcal{P}(\mathcal{C})^{[0,1]} \times \mathcal{A}^{[0,1]}$ such that

$$\begin{aligned} J^u(\widehat{\boldsymbol{\mu}}, \widehat{\pi}^u) &= \sup_{\pi\in\mathcal{A}} J^u(\widehat{\boldsymbol{\mu}}, \pi), \\ \widehat{\mu}^u &= \mathcal{L}(X^{u,\widehat{\pi}^u}), \end{aligned}$$

for almost every $u \in [0, 1]$. This is called "continuum-player formulation" since it involves a continuum of players.

C.2. Technical: The Non-Measurability Issue

Mathematically, the continuum-player formulation suffers from significant measurability difficulties. For completeness, we first cite (Sun, 2006, Proposition 2.1) as follows:

Proposition C.1. *Consider index space $(I, \mathcal{I}, \lambda)$ and probability space $(\Omega, \mathcal{F}, P)$. Consider function $f : I \times \Omega \to E$ for some Polish space E. Suppose f is measurable on the product space $(I \times \Omega, \mathcal{I} \otimes \mathcal{F}, \lambda \otimes P)$, equipped with the usual product σ-algebra, and for λ-almost every $j \in I$, f_j is independent of f_i for λ-almost every $i \in I$. Then, for λ-almost every $i \in I$, f_i is a constant random variable.*

Intuitively, the product σ-algebra $\mathcal{I} \otimes \mathcal{F}$ fails to support a large amount of information when we require both the joint measurability of f, and the independence between f_i and f_j. This would lead to a problem when we consider a continuum of players, even if the state space is a finite space rather than $\mathbb{R}^d$, and even for a static game.

More precisely, let $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ be a probability space that supports a collection of stochastic processes $\{X^u : u \in [0, 1]\}$, where X^u is a process on $\{0, 1, \ldots, T\}$ valued in $\mathcal{X}$ (which could be $\mathbb{R}^d$ or a finite state space). X^u represents the state

process of player with label u. From time $t-1$ to t, $\{X_t^u\}$ are generated independently for every $u \in [0,1]$, and thus the mapping $(u,\omega) \to X^u(\omega)$ is not measurable on the typical product σ-algebra; similarly, $(u,\omega) \mapsto a_t^u(X_t^u(\omega))$ is not measurable. This measurability issue leads to significant difficulties in the proof, as the objective reward function may involve these mappings. For instance, as one attempts to transform the continuum-player graphon game into a mean field game with augmented state space, the objective becomes

$$\mathbb{E}\int_{[0,1]}\left[\sum_{t\in\mathbb{T}}\overline{f}_t\left(\begin{pmatrix} u \\ X_t^{u,\pi^u}\end{pmatrix},\mu_t,a_t\right)+\overline{g}\left(\begin{pmatrix} u \\ X_T^{u,\pi^u}\end{pmatrix},\mu_T\right)\right]du,$$

where the integral with respect to (u,ω) over $[0,1]\times\Omega$ is not well-defined since the integrand is not measurable. A similar argument demonstrates why we cannot aggregate the objective of all the players in a continuum-player graphon game, where the integral $\int_{[0,1]} J^u(\boldsymbol{\mu},\pi^u)du$ is not well-defined. Thus, the continuum-player graphon game is not mathematically equivalent to our representative-player formulation.

This technical issue can be addressed by carefully enlarging the σ-algebra with rich Fubini extensions (Sun, 2006, Section 2), allowing it to hold more information while ensuring the joint measurability and independence (Aurell et al., 2022; Tangpi & Zhou, 2024). However, this approach is restricted to linear-quadratic problems.

On the contrary, our graphon game formulation considers only one representative player. Recall that for any $\mu \in \mathcal{P}_{\text{unif}}([0,1]\times\mathcal{C})$, the conditional law of X given U yielded by disintegration is a uniquely defined Lebesgue almost surely. Thus, it encodes less information by only considering almost every label u, but this provides great technical convenience and allows us to consider the game for one representative player (Lacker & Soret, 2023).

C.3. Philosophical: Clarification on the Representative Player

As demonstrated above, the representative-player graphon game we present in Section 4.1 and the continuum-player graphon game in Appendix C.1 are not mathematically equivalent.

Conceptually, our representative-player formulation inherits the spirit of mean field games. We recall that there is only one representative player in the mean field game, and all other players are abstracted into a population measure in $\mathcal{P}(\mathcal{C})$. Similarly, our game formulation is for one representative player, and the difference is that now the representative player is in addition assigned a random label, while all other players are abstracted into a label-state joint population measure on $\mathcal{P}_{\text{unif}}([0,1]\times\mathcal{C})$.

In other words, our graphon game formulation is defined directly for a representative player as in classic mean field games. This should be distinguished from reformulating a continuum-player graphon game into an MFG with augmented state space (in a similar way of Appendix A.2), which is a proof technique to adapt existing MFG results into a graphon game setting, and it is possible to avoid this technique as in our proofs. Using this technique does not change the fact that the problems remain for a continuum of players: the theorems are stated for the proposed continuum-player graphon game, and the measurability issue is not avoided. Thus, we do not regard this concept as "representative-player".

As a comparison, using one representative to define a graphon game directly at the beginning, as in (Lacker & Soret, 2023) and this paper, brings novelty compared to prior work. This is similar to the concept that classic MFG literature defines a game for a representative player from the start, rather than modelling a game for a continuum of players and using techniques to reformulate it in the proofs.

D. A Toy Example on the Difference between Two Formulations

In this section we compare two types of graphon game formulations on a toy example, inspired by the motivating example in (Cui & Koeppl, 2021). The two types of formulations of graphon games lead to the same equilibrium in this particular one-shot game, while the representative-player graphon game is simpler in formulation. When a finite player game contains larger and continuous state and action spaces with more complex settings, our formulation would demonstrate more advantages in both analysis and computation. Note that this toy example focuses on demonstrating the difference in formulation, and the measurability issue mentioned in Appendix C.2 is not the main point here as the example is simple enough to be solved explicitly, and no technical proofs are involved.

The interaction is defined by a threshold graph, where $\xi_{ij}^n = 1_{i+j<n}$. It is easy to see that W_{ξ^n} converges in cut norm to a limit defined by $W(u,v) := 1_{u+v<1}$. Note that this graphon is discontinuous.

D.1. n-player Game

Consider a one-shot (single-stage) game for n players, and let the state and action space be $\mathcal{X} = A = \{-1, 1\}$, understood as left and right. Each player simultaneously chooses either left or right, and is punished by the weighted average of proportion of players that chose the same action. Precisely,

$$a^i = \begin{cases} 1 & \text{w.p.} \quad p^i; \\ -1 & \text{w.p.} \quad 1 - p^i, \end{cases} \qquad X^i = a^i,$$

where p^i is the probability player i choose right (state 1), and this characterizes the policy. Let $\boldsymbol{p} = (p^1, \dots, p^n)$ and let the terminal reward be $g(x, m) = -\langle m, 1_x \rangle$, where 1_x is the indicator function. Player i aims to maximize

$$\begin{aligned} J^i(\boldsymbol{p}) &= -\mathbb{E}\bigg(\sum_{j=1}^{n} \xi_{ij}^n 1_{X^i = X^j} \bigg) \\ &= -\sum_{j=1}^{n-i} \Big(p^i p^j + (1 - p^i)(1 - p^j) \Big). \end{aligned}$$

It can be verified that one equilibrium is given by $p^1 = \cdots = p^n = \frac{1}{2}$.

D.2. Representative-Player Formulation

Consider a one-shot game for a single player, and let the state and action space be $\mathcal{X} = A = \{-1, 1\}$. Any population measure $\mu \in \mathcal{P}_{\text{unif}}([0, 1] \times \mathcal{X})$ can be characterized by a function $q(u) := \mu(u, \{1\})$, $\forall u \in [0, 1]$. Let this population measure be fixed. The graphon operator is given by

$$\mathsf{W}\mu(u) = \int_{[0,1]} W(u, v)(q(v)\delta_1 + (1 - q(v))\delta_{-1})dv \in \mathcal{M}_+(\{-1, 1\}),$$

where δ is Dirac delta measure. The player is randomly assigned a label $U \sim \text{unif}[0, 1]$, and let π be her policy. Equivalently the policy can be characterized by $p(u) := \pi(u)(\{1\})$. Then she follows the dynamic

$$a = \begin{cases} 1 & \text{w.p.} \quad p(U); \\ -1 & \text{w.p.} \quad 1 - p(U), \end{cases} \qquad X = a.$$

The objective is

$$\begin{aligned} J(q, p) &= -\mathbb{E}\Big(\langle \mathsf{W}\mu(U), 1_X \rangle \Big) \\ &= -\mathbb{E}\bigg(\int_{[0,1]} W(U, v)(q(v)1_{X=1} + (1 - q(v))1_{X=-1})dv \bigg) \\ &= -\int_{u+v<1} (q(v)p(u) + (1 - q(v))(1 - p(u)))dvdu. \end{aligned}$$

Solving this as a calculus of a variation problem provides a necessary condition $\int_0^u q(v)dv = \frac{1}{2}$, $\forall u \in [0, 1]$, and thus the equilibrium is given by $p(u) = \frac{1}{2}$ for a.e. u, and $q(v) = \frac{1}{2}$ for a.e. v.

D.3. Continuum-Player Formulation

Consider a static game for a continuum of players with the same setting, and let the population measure be $\boldsymbol{q} := \{q^u\}_{u \in [0,1]}$ for $q^u = \mu^u(\{1\})$. Each player $u \in [0, 1]$ admits a policy π^u as the probability choosing 1, so we may write $p^u := \pi^u(\{1\})$ and denote $\boldsymbol{p} := \{p^u\}_{u \in [0,1]}$. Then the player u chooses the action

$$a^u = \begin{cases} 1 & \text{w.p.} \quad p^u; \\ -1 & \text{w.p.} \quad 1 - p^u, \end{cases} \qquad X^u = a^u,$$

and optimize the objective

$$\begin{aligned} J^u(\boldsymbol{q}, p^u) &= -\mathbb{E}\Big(\langle \mathsf{W}\mu(u), 1_X\rangle\Big) \\ &= -\mathbb{E}\Big(\int_{[0,1]} W(u,v)(q^v 1_{X=1} + (1-q^v)1_{X=-1})dv\Big) \\ &= -\int_0^{1-u} (q^v p^u + (1-q^v)(1-p^u))dv. \end{aligned}$$

It is immediate that the equilibrium is given by $p^u = \frac{1}{2}$, and $q^v = \frac{1}{2}$ for almost every u, v. Note that the measurability issue is not a concern for this specific example, since it can be solved directly and thus doesn't involve technical analysis.

E. Proof for Existence

E.1. Preliminary Lemmas

Lemma E.1 ((Lacker, 2015, Lemma A.2)). *Let X_1 and X_2 be Polish spaces. Define the coordinate projections $\Pi_i : \mathcal{X}_1 \times \mathcal{X}_2 \to \mathcal{X}_i$ for $i = 1, 2$. Then a set $S \subset \mathcal{P}(\mathcal{X}_1 \times \mathcal{X}_2)$ is tight if and only if the sets $S_1 = \{\mu \circ \Pi_1^{-1} : \mu \in S\}$ and $S_2 = \{\mu \circ \Pi_2^{-1} : \mu \in S\}$ are tight in $\mathcal{P}(X_1)$ and $\mathcal{P}(X_2)$ respectively.*

Lemma E.2 ((Lacker, 2015, Corollary A.5)). *Let E, F, G be complete, separable metric spaces. $\phi : E \times F \times G \to \mathbb{R}$ is a bounded measurable function, with $\phi(x, \cdot, \cdot)$ being jointly continuous for any $x \in E$. Then the following mapping is continuous:*

$$G \times \mathcal{P}(E \times F) \ni (z, P) \mapsto \int_{E\times F} \phi(x, y, z)P(dx, dy).$$

Lemma E.3. *Let W^n, W be graphons. If $W^n \xrightarrow{L_1} W$, then, $W^n \longrightarrow W$.*

Proof. Given any $\psi \in L_\infty[0, 1]$,

$$\begin{aligned} \|\mathsf{W}^n\psi - \mathsf{W}\psi\|_1 &= \int_{[0,1]} \Big|\mathsf{W}^n\psi(u) - \mathsf{W}\psi(u)\Big|du \\ &= \int_{[0,1]} \Big|\int_{[0,1]} W^n(u,v)\psi(v) - W(u,v)\psi(v)dv\Big|du \\ &\le \|\psi\|_\infty \int_{[0,1]^2} \Big|W^n(u,v) - W(u,v)\Big|dvdu \\ &= \|\psi\|_\infty \|W^n - W\|_1 \longrightarrow 0. \end{aligned}$$

□

Lemma E.4 ((Lacker & Soret, 2023, Lemma 4.2)). *] Let E be any Polish space, and W be any graphon.*

1. *For a.e. $u \in [0, 1]$, the following map is continuous:*

$$\mathcal{P}_{\text{unif}}([0, 1] \times E) \ni \mu \mapsto \mathsf{W}\mu(u) \in \mathcal{M}_+(E).$$

2. *Suppose the map $[0, 1] \ni u \mapsto \mathsf{W}(u, v)dv \in \mathcal{M}_+([0, 1])$ is continuous, then for any $\mu \in \mathcal{P}_{\text{unif}}([0, 1] \times E)$,*

$$[0, 1] \ni u \mapsto \mathsf{W}\mu(u) \in \mathcal{M}_+(E)$$

is continuous.

Lemma E.5. *Let E, F be complete, separable metric space, and F is a regular measurable space. Consider a sequence of probability measures on the product space $\{\nu_n\} \subset \mathcal{P}(E \times F)$. Suppose that ν_n admits disintegration*

$$\nu_n(dx, dy) = \mu_n(dx)K(x, dy),$$

for some common kernel K, which is continuous as a mapping $E \to \mathcal{P}(F)$, i.e., any sequence $x_n \to x$ implies $K_{x_n} \Rightarrow K_x$. Then if $\nu_n \Rightarrow \nu$, ν admits a disintegration $\nu(dx, dy) = \mu(dx)K(x, dy)$ for some $\mu \in \mathcal{P}(E)$.

Proof. Let Π_1 be the projection to first coordinate, which is a continuous mapping. By the continuous mapping theorem, the pushforward of a weak convergence measure sequence under continuous mapping converge weakly:

$$\mu_n \coloneqq \nu_n \circ \Pi_1^{-1} \Rightarrow \nu \circ \Pi_1^{-1} =: \mu.$$

Suppose ν admits disintegration $\nu(dx, dy) = \mu(dx)\bar{K}(x, dy)$ for some $\bar{K}$. Given any bounded and jointly continuous $\phi : E \times F \to \mathbb{R}$, the mapping $E \ni x \mapsto \int_F \phi(x, y)K(x, dy) \in \mathbb{R}$ is bounded and continuous since for any $x_n \to x$,

$$\begin{aligned}
&\left| \int_F \phi(x_n, y)K(x_n, dy) - \int_F \phi(x, y)K(x, dy) \right| \\
&\leq \left| \int_F \phi(x_n, y)K(x_n, dy) - \int_F \phi(x, y)K(x_n, dy) \right| + \left| \int_F \phi(x, y)K(x_n, dy) - \int_F \phi(x, y)K(x, dy) \right|,
\end{aligned}$$

which converges to 0. Finally, $\langle \nu_n, \phi \rangle \to \langle \nu, \phi \rangle$, and on the other hand,

$$\langle \nu_n, \phi \rangle = \int_{E \times F} \phi(x, y)K(x, dy)\mu_n(dx) \longrightarrow \int_{E \times F} \phi(x, y)K(x, dy)\mu(dx),$$

which holds for any bounded continuous ϕ. We conclude that K is a version of $\bar{K}$. □

E.2. Existence of Equilibrium

Given any function $\phi : E \times A \to F$ for Polish space E, F and a measure $\pi \in \mathcal{P}(A)$, we may also abuse the notation by writing ϕ as a function $E \times \mathcal{P}(A) \to F$, defined by $\phi(x, \pi) = \langle \pi, \phi(x, \cdot) \rangle$ for each $x \in E$.

Throughout the proof we fix a graphon W, and denote $\mathcal{V} = \mathcal{P}(A)^T$ the space of all policies. We fix an arbitrary policy $\pi \in \mathcal{V}$, and construct the label-state joint measure of the representative player controlled by π as follows. Recall that at time t given $U = u, X_t = x, \alpha_t = a$, the law of next state X_{t+1} follows the probabilistic kernel $[0, 1] \times \mathbb{R}^d \times A \to \mathbb{R}^d$:

$$\mathcal{L}(X_{t+1}|X_t = x, U_t = u, \alpha_t = a)(dy) = P_t(dy|x, \mathsf{W}\mu_t(u), a), \qquad \forall y \in \mathbb{R}^d,$$

and the control process α_t follows

$$\mathcal{L}(\alpha_t)(da) = \pi_t(da), \qquad \forall a \in A.$$

We may thus consider

$$\widehat{P}_t^{\pi,\mu}(dy|u, x) := \mathcal{L}(X_{t+1}|X_t = x, U_t = u)(dy) = \int_A P_t(dy|x, \mathsf{W}\mu_t(u), a)\pi_t(da), \qquad \forall y \in \mathbb{R}^d,$$

and we use the superscript to emphasize that the law is controlled by the policy π. Note that $\mathcal{V}_U \ni \pi \mapsto \widehat{P}_t^{\pi,\mu}(u, x) \in \mathcal{P}(\mathbb{R}^d)$ is measurable. The collection of kernels $\{\widehat{P}_t^{\pi}\}_{t \in \mathbb{T}}$ (recall $\mathbb{T} = \{0, 1, \ldots, T-1\}$) along with the initial law λ implies a label-state joint law in $\mathcal{P}_{\text{unif}}([0, 1] \times \mathcal{C})$

$$\widehat{P}^{\pi,\mu}(du, dx) := \mathcal{L}(U, X)(du, dx) = \lambda(du, dx_0) \prod_{t \in \mathbb{T}} \widehat{P}_t^{\pi,\mu}(dx_{t+1}|u, x_t), \qquad \forall (u, x) \in [0, 1] \times \mathcal{C},$$

which is the label-state joint measure of the representative player, when her state dynamic is controlled by π. Since the space $[0, 1] \times \mathcal{C}$ is a standard measurable space, this is understood as a regular version of the kernel from $\mathcal{V}$ to $[0, 1] \times \mathcal{C}$.

Lemma E.6. *Under Assumption 4.4(5), for any $\mu \in \mathcal{P}_{\text{unif}}([0,1]\times\mathcal{C})$, $\pi \mapsto \widehat{P}^{\pi,\mu}$ is continuous. In particular, $\pi_t \mapsto \widehat{P}_t^{\pi,\mu}(u,x)$ is continuous for every $(u,x)\in[0,1]\times\mathbb{R}^d$.*

Proof. Let $\{\pi^n\}\subset\mathcal{V}$ be any sequence of policies such that $\pi^n \Rightarrow \pi$ for some $\pi\in\mathcal{V}$. For any $\phi : [0,1]\times\mathcal{C}\to\mathbb{R}$ bounded continuous,

$$\begin{aligned}
&\int_{[0,1]\times\mathcal{C}} \phi(u,x)\widehat{P}^{\pi^n,\mu}(du,dx)\\
&= \int_{[0,1]\times\mathbb{R}^d}\Big[\int_{A^T}\int_{(\mathbb{R}^d)^T}\phi(u,x_0,\dots,x_T)\prod_{t\in\mathbb{T}}P_t(dx_{t+1}|x_t,\mathsf{W}\mu_t(u),a_t)\pi^n(da_0,\dots,da_{T-1})\Big]\lambda(du,dx_0)\\
&=: \int_{[0,1]\times\mathbb{R}^d}\Big[\int_{A^T}\psi(a_0,\dots,a_{T-1})\pi^n(da_0,\dots,da_{T-1})\Big]\lambda(du,dx_0),
\end{aligned}$$

where

$$\psi(a_0,\dots,a_{T-1}) := \int_{(\mathbb{R}^d)^T}\phi(u,x_0,\dots,x_T)\prod_{t\in\mathbb{T}}P_t(dx_{t+1}|x_t,\mathsf{W}\mu_t(u),a_t).$$

We know that $a_t\mapsto P_t(dx_{t+1}|x_t,\mathsf{W}\mu_t(u),a_t)$ is continuous for each $t\in\mathbb{T}$ by Assumption 4.4(5), and since $(\mathbb{R}^d)^T$ is separable, with the standard measure theory argument for weak convergence on a product space, for instance, (Billingsley, 1995, Chapter 2), the map ψ is continuous. Thus, $\langle\pi^n,\psi\rangle\to\langle\pi,\psi\rangle$, and

$$\begin{aligned}
&\int_{[0,1]\times\mathcal{C}}\phi(u,x)\widehat{P}^{\pi^n,\mu}(du,dx)\\
\longrightarrow&\int_{[0,1]\times\mathbb{R}^d}\Big[\int_{A^T}\psi(a_0,\dots,a_{T-1})\pi(da_0,\dots,da_{T-1})\Big]\lambda(du,dx_0)\\
&=\int_{[0,1]\times\mathcal{C}}\phi(u,x)\widehat{P}^{\pi,\mu}(du,dx).
\end{aligned}$$

□

Define the probability space $\Omega := \mathcal{V}\times[0,1]\times\mathcal{C}$, equipped with the product σ-algebra. A typical element of Ω is (π,u,x), where we understood them as a policy, a label of the representative player, and the player's path, respectively. Let the coordinate maps be Λ, U, X respectively. The filtration is given by $\mathcal{F}_t = \sigma\{\Lambda|_{[t]\times A}, U, \{X_s\}_{0\le s\le t}\}$.

The collection of admissible laws $\mathcal{R}(\mu)$ is defined as the set

$$\mathcal{R}(\mu) := \{R\in\mathcal{P}(\Omega) : R \text{ admits disintegration } R(d\pi,du,dx) = R_\Lambda(d\pi)\widehat{P}^{\pi,\mu}(du,dx) \text{ for some } R_\Lambda\in\mathcal{P}(\mathcal{V})\}.$$

Define a random variable $\Xi^\mu:\Omega\to\mathbb{R}$ by

$$\Xi^\mu(\pi,u,x) := \sum_{t\in\mathbb{T}}\int_A f_t(\mathsf{W}\mu_t(u),x_t,a)\pi_t(da) + g(x_T,\mathsf{W}\mu_T(u)) \tag{17}$$

where μ_t is the marginal obtained as the image by $(u,x)\mapsto(u,x_t)$. In particular, given a policy $\pi\in\mathcal{V}$, let $R^{(\pi)}(d\tilde\pi,du,dx) := \delta_\pi(d\tilde\pi)\widehat{P}^{\tilde\pi,\mu}(du,dx)$ be an element of $\mathcal{R}(\mu)$, where δ is the Dirac measure. It holds that the objective can be rewritten as

$$J_W(\mu,\pi) = \langle R^{(\pi)},\Xi^\mu\rangle.$$

Thus, the expectation $\langle R,\Xi^\mu\rangle$ is a reformulation of the objective, and a single player's objective is to find the collection of measures that maximize this expectation:

$$\mathcal{R}^*(\mu) := \{R^*\in\mathcal{R}(\mu) : \langle R^*,\Xi^\mu\rangle \ge \langle R,\Xi^\mu\rangle,\ \forall R\in\mathcal{R}(\mu)\} \tag{18}$$

Define the correspondence (i.e., set valued function, see (Aliprantis & Border, 2006) for an overview) $\Phi : \mathcal{P}([0,1]\times\mathcal{C}) \to 2^{\mathcal{P}([0,1]\times\mathcal{C})}$, given by

$$\Phi(\mu) := \{R \circ (U,X)^{-1} : R \in \mathcal{R}^*(\mu)\}.$$

The existence of $W-$equilibrium is divided into two steps: we first show the existence of an optimizer to the optimization problem (18) over the probability measures, i.e., $\mathcal{R}^*(\mu)$ is non-empty for any μ; Next, to obtain a $W-$equilibrium, we aim to find a fixed point for the correspondence Φ.

Proposition E.7. *For any $\mu \in \mathcal{P}_{\text{unif}}([0,1]\times\mathcal{C})$, the following optimization problem admits an optimizer:*

$$\sup_{R\in\mathcal{R}(\mu)} \langle R, \Xi^\mu\rangle.$$

Proof. We want to show that $R \mapsto \langle R, \Xi^\mu\rangle$ is a continuous mapping on compact space, and thus the maximum of this mapping is attained. With a direct application of Lemma E.2, we immediately conclude that the following map is jointly continuous:

$$\text{Gr}(\mathcal{R}) \ni (\mu, R) \longmapsto \langle R, \Xi^\mu\rangle \in \mathbb{R}, \tag{19}$$

where Gr denotes the graph of an operator.

It remains to prove that $\mathcal{R}(\mu)$ is compact. First we want to show $\mathcal{R}(\mu)$ is tight for any μ. By Lemma E.1, it suffices to show that the following sets are tight: $\{R\circ X^{-1} : R\in\mathcal{R}(\mu)\}$, $\{R\circ U^{-1} : R\in\mathcal{R}(\mu)\}$, and $\{R\circ\Lambda^{-1} : R\in\mathcal{R}(\mu)\}$. The last two follows immediately from the fact that $[0,1]$ and A are compact spaces.

Fix any $\epsilon' > 0$, we could always find some ϵ such that $(1-\epsilon)^{T+1} > 1-\epsilon'$. By Assumption 4.4(3) and 4.4(4), let $\{K_t\}_{t\in\mathbb{T}}$ be compact subsets of $\mathbb{R}^d$ such that

$$\inf_{u\in[0,1]} \lambda^u(K_0) > 1-\epsilon, \qquad \inf_{\widetilde{P}_t\in\zeta_t} \widetilde{P}_t(K_{t+1}) > 1-\epsilon, \quad \forall t\in\mathbb{T}.$$

Define $K = \prod_{t=0}^T K_t$, which is a compact subset of $\mathcal{C}$. For every $R\in\mathcal{R}(\mu)$, let $\widehat{P}^{\pi,\mu}(du,dx)R_\Lambda(d\pi)$ be its disintegration. Then,

$$\begin{aligned}
(R\circ X^{-1})(K) &= R(\mathcal{V}\times[0,1]\times K)\\
&= \int_{\mathcal{V}\times[0,1]\times\mathcal{C}} 1_K(x)\widehat{P}^{\pi,\mu}(du,dx)R_\Lambda(d\pi)\\
&= \int_{\mathcal{V}}\int_{[0,1]\times\mathbb{R}^d}\Big[\prod_{t=0}^{T-1}\int_A\int_{\mathbb{R}^d} 1_{K_{t+1}}(x_{t+1})P_t(dx_{t+1}|x_t,\mathsf{W}\mu_t(u),a)\pi_t(da)\Big]1_{K_0}(x_0)\lambda(du,dx_0)R_\Lambda(d\pi)\\
&\geq \int_{\mathcal{V}}\int_{[0,1]}\Big[\prod_{t=0}^{T-1}\int_A(1-\epsilon)\pi_t(da)\Big]\int_{\mathbb{R}^d}1_{K_0}(x_0)\lambda^u(dx_0)duR_\Lambda(d\pi)\\
&\geq \int_{\mathcal{V}}\int_{[0,1]}(1-\epsilon)^{T+1}duR_\Lambda(d\pi)\\
&= (1-\epsilon)^{T+1} > 1-\epsilon'.
\end{aligned}$$

Thus, we have $\inf_{R\in\mathcal{R}(\mu)}(R\circ X^{-1})(K) > 1-\epsilon'$, which implies the tightness of $\{R\circ X^{-1} : R\in\mathcal{R}(\mu)\}$. Note that if the state space $\mathcal{X}$ of dynamic X is compact, then $\{R\circ X^{-1} : R\in\mathcal{R}(\mu)\}$ being tight is immediate. By Prokhorov's theorem, $\mathcal{R}(\mu)$ is precompact.

We conclude by showing that $\mathcal{R}(\mu)$ is closed. Let $\{R_n\}\subset\mathcal{R}(\mu)$, and $R_n \Rightarrow R$. Indeed, each R_n admits disintegration $R^n_\Lambda(d\pi)\widehat{P}^{\pi,\mu}(du,dx)$ for some $R^n_\Lambda\in\mathcal{P}(\mathcal{V})$, and the kernel $\pi\mapsto\widehat{P}^{\pi,\mu}$ is continuous by Lemma E.6. Then Lemma E.5 implies that R admits disintegration $R_\Lambda(d\pi)\widehat{P}^{\pi,\mu}(du,dx)$ and thus $R\in\mathcal{R}(\mu)$. □

Next we show that the correspondence Φ admits a fixed point, and thus the graphon game admits a W-equilibrium.

Proposition E.8. *There exists a fixed point $\widehat{\mu}$ for the correspondence Φ.*

Proof. We aim to apply the Kakutani-Fan-Glicksberg fixed point theorem, which is a classic fixed point theorem for correspondences, see for instance (Aliprantis & Border, 2006, Theorem 17.55). We need to show the existence of a nonempty, convex, and compact $K \subset \mathcal{P}([0,1] \times \mathcal{C})$, such that

1. $\Phi(\mu) \subset K$ for each $\mu \in K$.
2. $\Phi(\mu)$ is nonempty and convex for each $\mu \in K$.
3. The graph $\mathrm{Gr}(\Phi) = \{(\mu, \mu') : \mu \in K, \mu' \in \Phi(\mu)\}$ is closed.

We start from defining K. Note that λ is a fixed initial measure. Let

$$K := \{\lambda \otimes \prod_{t=0}^{T-1} \widehat{P}_t : \widehat{P}_t \in \overline{\mathrm{conv}}(\zeta_t)\},$$

where $\overline{\mathrm{conv}}(\cdot)$ denotes the closed convex hull of a set, and $\otimes$ is the combinations of probabilistic kernels on the product space. K is obviously non-empty. By construction, K is the finite Cartesian product of convex sets, and thus K is convex. To show K is compact, it suffices to show $\overline{\mathrm{conv}}(\zeta_t)$ is compact for each $t \in \mathbb{T}$, since Tychonoff's theorem asserts that an arbitrary product of compact spaces is again compact. This is true because ζ_t is tight, and thus precompact by Prokhorov's theorem, and the closed convex hull of a precompact set is compact in a locally convex Hausdorff space. Again, if the value space $\mathcal{X}$ of X is compact, let $K = \mathcal{P}([0,1] \times \mathcal{C})$ and K is compact automatically.

For each $R \in \mathcal{R}(\mu)$, let it admit the disintegration $R = R_\Lambda \otimes \widehat{P}$:

$$R_\Lambda(d\pi)\widehat{P}^{\pi,\mu}(du, dx) = \Big[\lambda(du, dx_0) \prod_{t=0}^{T-1} \int_A P_t(dx_{t+1}|x_t, \mathsf{W}\mu_t(u), a)\pi_t(da)\Big] R_\Lambda(d\pi).$$

We claim that for any $t \in \mathbb{T}$,

$$\widehat{P}_t^{\pi,\mu}(dx_{t+1}|u, x_t) = \int_A P_t(dx_{t+1}|x_t, \mathsf{W}\mu_t(u), a)\pi_t(da) \in \overline{\mathrm{conv}}(\zeta_t),$$

since it is the limit of convex combinations of $P_t(\cdot|x_t, \mathsf{W}\mu_t(u), a) \in \zeta_t$. Thus, for any $(\pi, \mu) \in \mathcal{V} \times \mathcal{P}_{\mathrm{unif}}([0,1] \times \mathcal{C})$, the measure $\widehat{P}^{\pi,\mu} \in \mathcal{P}_{\mathrm{unif}}([0,1] \times \mathcal{C})$ belongs to K. The pushforward of R onto the (U, X) coordinate is

$$(R \circ (U, X)^{-1})(du, dx) = \int_{\mathcal{V}} \widehat{P}^{\pi,\mu}(du, dx) R_\Lambda(d\pi),$$

which is also the limit of a sequence of convex combinations of $\widehat{P}^{\pi,\mu}(du, dx)$, indexed by π. Thus, by the closeness and compactness of K, $R \circ (U, X)^{-1} \in K$, and thus $\Phi(\mu) \subset K$.

To show the convexity of $\Phi(\mu)$, we start with showing $\mathcal{R}(\mu)$ is convex since for any $R^1 = R^1_\Lambda \otimes \widehat{P}$ and $R^2 = R^2_\Lambda \otimes \widehat{P}$ and $\lambda \in [0,1]$, $\lambda R^1 + (1-\lambda)R^2 = (\lambda R^1_\Lambda + (1-\lambda)R^2_\Lambda) \otimes \widehat{P} \in \mathcal{R}(\mu)$. Convexity of $\mathcal{R}^*(\mu)$ follows from the linearity of $R \mapsto \langle P, \Xi^\mu \rangle$ and the convexity of $\mathcal{R}(\mu)$, and thus the convexity of $\Phi(\mu)$ follows from the linearity of map $R \mapsto R \circ (U, X)^{-1}$ and the convexity of $\mathcal{R}^*(\mu)$.

It remains to show the closeness of the graph of Φ. We first show the closeness of the following set:

$$\{(\mu, R) : \mu \in K, R \in \mathcal{R}^*(\mu)\}.$$

Let $\mu_n \Rightarrow \mu$ and $R_n \Rightarrow R$ with $\mu_n, \mu \in K$, $R_n \in \mathcal{R}^*(\mu_n)$, and $R \in \mathcal{R}$. To show that $R \in \mathcal{R}^*(\mu)$, we use the continuity condition (19), and for any $R' \in \mathcal{R}$,

$$\langle R, \Xi^\mu \rangle = \lim_{n\to\infty} \langle R_n, \Xi^{\mu_n} \rangle \geq \lim_{n\to\infty} \langle R', \Xi^{\mu_n} \rangle = \langle R', \Xi^\mu \rangle.$$

Thus, $\langle R, \Xi^\mu \rangle \geq \langle R', \Xi^\mu \rangle$ for any $R' \in \mathcal{R}$. The by the continuity of $R \mapsto R \circ (U, X)^{-1}$ and compactness of K, we have the closeness of $\mathrm{Gr}(\Phi)$. □

E.3. Closed-Loop Equilibrium Optimal Policy

In this section we show the second part of theorem 4.5, that the equilibrium optimal open-loop policy can be made closed-loop.

Proposition E.9. *Let $\mu \in \mathcal{P}_{\text{unif}}([0,1]\times\mathcal{C})$, and $R \in \mathcal{R}(\mu)$. Then, there exists a closed-loop optimal policy, in the following sense: there exists a measurable function $\pi : \mathbb{T}\times[0,1]\times\mathbb{R}^d \to \mathcal{P}(A)$, and $R^0 \in \mathcal{R}(\mu)$, such that*

1. $R^0(\Lambda_t(da) = \pi_t(U, X_t)(da),\ \forall t \in \mathbb{T}) = 1$,
2. $\int_\Omega \Xi^\mu dR^0 \geq \int_\Omega \Xi^\mu dR$,
3. $R^0 \circ (U, X_t)^{-1} = R \circ (U, X_t)^{-1},\ \forall t \in \mathbb{T}$.

Corollary E.10. *There exists a closed-loop equilibrium optimal policy to the graphon game.*

Proof. We first find a space $(\Omega^1, \mathcal{F}^1, R^1)$ supporting a random variable U^1, an adapted process X^1 valued in $\mathbb{R}^d$, and a $\mathcal{P}(A)$-valued adapted process Λ_t such that

$$(U^1, X^1_0) \sim \lambda, \quad X^1_{t+1} \sim P_t(X^1_t, \mathsf{W}\mu_t(U^1), \Lambda_t),$$
$$R^1 \circ \left(U, X^1\right)^{-1} = \mu.$$

The existence of such a space is guaranteed by the reasoning in Appendix E.2. We claim that there exists a measurable $\pi : \mathbb{T}\times[0,1]\times\mathbb{R}^d \to \mathcal{P}(A)$ such that

$$\pi_t(U^1, X^1_t) = \mathbb{E}^{R^1}(\Lambda_t \,|\, U^1, X^1_t), \qquad R^1 - a.s.\ \forall t \in \mathbb{T}.$$

More precisely, for every bounded measurable $\phi : [0,1]\times\mathbb{R}^d\times A \to \mathbb{R}$,

$$\int_A \phi(U^1, X^1_t, a)\pi_t(U^1, X^1_t)(da) = \mathbb{E}^{R^1}\left(\int_A \phi(U^1, X^1_t, a)\Lambda_t(da)\,\middle|\, U^1, X^1_t\right), \qquad R^1 - a.s., \forall t \in \mathbb{T}.$$

Define a collection of measures, $\{\eta_t\}_{t\in\mathbb{T}}$, $\eta_t \in \mathcal{P}([0,1]\times\mathbb{R}^d\times A)$ by

$$\eta_t(C) := \mathbb{E}^{R^1}\left[\int_A 1_C(t, U^1_t, X^1_t, a)\Lambda_t(da)\right].$$

Let η_t admit disintegration $\eta_t(du, dx, da) = \eta'_t(du, dx)\pi_t(u, x)(da)$, where η'_t is the marginal of η_t onto $[0,1]\times\mathbb{R}^d$. Note that actually $\eta'_t(du, dx) = \mu_t$, since for any measurable $F \subset [0,1]\times\mathbb{R}^d$,

$$\begin{aligned}\eta'_t(F) = \eta_t(F\times A) &= \mathbb{E}^{R^1}\left[\int_A 1_F(U^1_t, X^1_t)1_A(a)\Lambda_t(da)\right] \\ &= \mathbb{E}^{R^1}\left[1_F(U^1_t, X^1_t)\right] = \langle R^1 \circ (U, X^1)^{-1}, 1_F\rangle.\end{aligned}$$

Fix an arbitrary t, for any bounded measurable $h : \times[0,1]\times\mathbb{R}^d \to \mathbb{R}$,

$$\begin{aligned}&\mathbb{E}^{R^1}\left[h(U^1, X^1_t)\int_A \phi(U^1, X^1_t, a)\pi_t(U^1, X^1_t)(da)\right] \\ &= \int_{[0,1]\times\mathbb{R}^d} h(u, x)\int_A \phi(u, x, a)\pi_t(u, x)(da)\eta'_t(du, dx) \\ &= \int_{[0,1]\times\mathbb{R}^d\times A} h(u, x)\phi(u, x, a)\eta_t(du, dx, da) \\ &= \mathbb{E}^{R^1}\left[h(U^1, X^1_t)\int_A \phi(U^1, X^1_t, a)\Lambda_t(da)\right].\end{aligned}$$

By definition of conditional expectation, the claim follows.

Construct another probability space $(\Omega^2, \mathcal{F}^2, R^2)$ as follows: Let $\Omega^2 = [0,1] \times \mathcal{C}$, U^2 and X^2 are the coordinate maps, and

$$R^2 := R^1 \circ \left(U^1, X^1\right)^{-1} = \mu.$$

In the rest of the proof, we aim to show that U^2 and X^2 follow the dynamic

$$(U^2, X_0^2) \sim \lambda, \quad X_{t+1}^2 \sim P_t(X_t^2, \mathsf{W}\mu_t(U^2), \pi_t(U^2, X_t^2)).$$

Fix any bounded continuous $\psi : \mathbb{R}^d \to \mathbb{R}$. For any measurable $h : [0,1] \times \mathbb{R}^d \to \mathbb{R}_+$,

$$\begin{aligned}
&\mathbb{E}^{R^2}\left[h(U^2, X_t^2)\psi(X_{t+1}^2)\right] \\
&= \mathbb{E}^{R^1}\left[h(U^1, X_t^1)\psi(X_{t+1}^1)\right] \\
&= \mathbb{E}^{R^1}\left[h(U^1, X_t^1)\mathbb{E}^{R^1}\left(\int_A \int_{\mathbb{R}^d} \psi(y) P_t(\mathsf{W}\mu_t(U^1), X_t^1, a)(dy)\Lambda_t(da) \,\middle|\, U^1, X_t^1\right)\right] \\
&= \mathbb{E}^{R^1}\left[h(U^1, X_t^1)\int_A \int_{\mathbb{R}^d} \psi(y) P_t(\mathsf{W}\mu_t(U^1), X_t^1, a)(dy)\pi_t(U^1, X_t^1)(da)\right] \\
&= \mathbb{E}^{R^2}\left[h(U^2, X_t^2)\int_A \int_{\mathbb{R}^d} \psi(y) P_t(\mathsf{W}\mu_t(U^2), X_t^2, a)(dy)\pi_t(U^2, X_t^2)(da)\right].
\end{aligned}$$

By definition of conditional expectation, we claim that

$$\mathbb{E}^{R^2}\left[\psi(X_{t+1}^2) \,|\, U^2, X_t^2\right] = \int_A \int_{\mathbb{R}^d} \psi(y) P(t, U^2, \mathsf{W}\mu_t(U^2), X_t^2, a)(dy)\pi_t(U^2, X_t^2)(da).$$

Note that this holds for any bounded continuous ψ. Finally, let $R^0 := R^2 \circ (\{\pi_t(U^2, X_t^2)\}_{t\in T}, U^2, X^2)^{-1}$, then $R^0 \in \mathcal{R}(\mu)$, and the objective value is preserved:

$$\begin{aligned}
\int_\Omega \Xi^\mu dR^0 &= \mathbb{E}^{R^2}\left[\sum_{t\in T}\int_A f(t, U^2, X_t^2, \mathsf{W}\mu_t(U^2), a)\pi(t, U^2, X_t^2)(da) + g(X_T^2, \mathsf{W}\mu_T(U^2))\right] \\
&= \mathbb{E}^{R^1}\left[\sum_{t\in T}\int_A f(t, U^1, X_t^1, \mathsf{W}\mu_t(U^1), a)\pi_t(U^1, X_t^1)(da) + g(X_T^1, \mathsf{W}\mu_T(U^1))\right] \\
&= \mathbb{E}^{R^1}\left[\sum_{t\in T}\int_A f(t, U^1, X_t^1, \mathsf{W}\mu_t(U^1), a)\Lambda_t(da) + g(X_T^1, \mathsf{W}\mu_T(U^1))\right] \\
&= \int_\Omega \Xi^\mu dR.
\end{aligned}$$

□

Remark E.11. The proof is closely based on (Lacker, 2015), which utilized a remarkable result called Markovian projection theorem (or Mimicking theorem), originated from (Brunick & Shreve, 2013). However, the discrete time setting greatly simplifies the proof and just the definition of conditional expectation would work.

F. Proof for Uniqueness

Let (μ, π) and (ν, ρ) be two different W-equilibria, and their Markovian state dynamic being X^π and X^ρ respectively. By construction, the processes π and ρ must be different, since otherwise X^π and X^ν would be the same, and then μ and ν will be the same as well. Therefore, by the uniqueness of optimal policy, we have

$$J_W(\mu, \pi) - J_W(\mu, \rho) > 0 \quad \text{and} \quad J_W(\nu, \rho) - J_W(\nu, \pi) > 0.$$

Note that the inequalities are strict. Adding them result in

$$J_W(\mu,\pi) - J_W(\nu,\pi) - (J_W(\mu,\rho) - J_W(\nu,\rho)) > 0 \tag{20}$$

Since the Markovian dynamic does not depend on the measure argument, when the population measure is μ, the dynamic controlled by policy ρ is the same pathwise as X^ρ. This is not true if the assumption is not satisfied, since

$$X^\pi_{t+1} \sim P_t(X^\pi_t, \mathsf{W}\mu_t(U), \pi_t), \qquad X^\rho_{t+1} \sim P_t(X^\rho_t, \mathsf{W}\nu_t(U), \rho_t),$$

and under population measure μ, the process controlled by ρ follows the dynamic $X^{'}_{t+1} \sim P_t(X^{'}_t, \mathsf{W}\mu_t(U), \rho_t)$, which is not the same as X^ρ. Continue with the proof,

$$\begin{aligned}
J_W(\mu,\pi) - J_W(\nu,\pi) = \mathbb{E}\Big[&\sum_{t\in\mathbb{T}} \big(f^1_t(X^\pi_t, \mathsf{W}\mu_t(U)) - f^1_t(X^\pi_t, \mathsf{W}\nu_t(U))\big) \\
&+ \sum_{t\in\mathbb{T}} \big(f^2_t(X^\pi_t,\pi) - f^2_t(X^\pi_t,\pi)\big) + g(X^\pi_T, \mathsf{W}\mu_T(U)) - g(X^\pi_T, \mathsf{W}\nu_T(U))\Big] \\
=&\sum_{t\in\mathbb{T}} \int_{[0,1]\times\mathbb{R}^d} \big[f^1_t(x, \mathsf{W}\mu_t(u)) - f^1_t(x, \mathsf{W}\nu_t(u))\big]\mu_t(du,dx) \\
&+ \int_{[0,1]\times\mathbb{R}^d} \big[g(x, \mathsf{W}\mu_T(u)) - g(x, \mathsf{W}\nu_T(u)\big]\mu_T(du,dx).
\end{aligned}$$

Similarly,

$$\begin{aligned}
J_W(\mu,\rho) - J_W(\nu,\rho) =&\sum_{t\in\mathbb{T}} \int_{[0,1]\times\mathbb{R}^d} \big[f^1_t(x, \mathsf{W}\mu_t(u)) - f^1_t(x, \mathsf{W}\nu_t(u))\big]\nu_t(du,dx) \\
&+ \int_{[0,1]\times\mathbb{R}^d} \big[g(x, \mathsf{W}\mu_T(u)) - g(x, \mathsf{W}\nu_T(u)\big]\nu_T(du,dx).
\end{aligned}$$

Taking difference and by the assumed Larsy-Lions monotonicity,

$$\begin{aligned}
&J_W(\mu,\pi) - J_W(\nu,\pi) - (J_W(\mu,\rho) - J_W(\nu,\rho)) \\
&= \sum_{t\in\mathbb{T}} \int_{[0,1]\times\mathbb{R}^d} \big[f^1_t(x, \mathsf{W}\mu_t(u)) - f^1_t(x, \mathsf{W}\nu_t(u))\big](\mu_t - \nu_t)(du,dx) \\
&\quad + \int_{[0,1]\times\mathbb{R}^d} \big[g(x, \mathsf{W}\mu_T(u)) - g(x, \mathsf{W}\nu_T(u)\big](\mu_T - \nu_T)(du,dx) \\
&\le 0.
\end{aligned}$$

However, this contradicts (20), and we conclude that μ and ν should be the same.

G. Proof for Approximate Equilibrium

G.1. Comparable Dynamics

Define $I^n_i := [(i-1)/n, i/n)$ for $i = 1,\dots,n-1$, $I^n_n := [(n-1)/n, 1]$, and $\mathbf{I}^n := I^n_1 \times \cdots \times I^n_n$. Let (μ,π) be a W-equilibrium, and X be the Markov chain controlled by policy π. Let X^u denote the state process conditional on $U = u$.

Fix $\forall n \in \mathbb{N}$, and any $\mathbf{u}^n = (u^n_1,\dots,u^n_n) \in [0,1]^n$. Assign player i the policy

$$\widehat{\pi}^{n,\mathbf{u}^n,i}(t,x_1,\dots,x_n) := \pi(t,u^n_i,x_i).$$

Let $\widehat{\mathbf{X}}^{n,\mathbf{u}^n} = (\widehat{X}^{n,\mathbf{u}^n,1},\dots,\widehat{X}^{n,\mathbf{u}^n,n})$ be the state dynamic of all the players:

$$\widehat{X}^{n,\mathbf{u}^n,i}_{t+1} \sim P_t(\widehat{X}^{n,\mathbf{u}^n,i}_t, \widehat{M}^{n,\mathbf{u}^n,i}_t, \widehat{\pi}^{n,\mathbf{u}^n,i}_t(\widehat{\mathbf{X}}^{n,\mathbf{u}^n}_t)), \qquad \widehat{X}^{n,\mathbf{u}^n,i}_0 \sim \lambda_{u^n_i},$$

where

$$\widehat{M}^{n,\mathbf{u}^n,i} := \frac{1}{n}\sum_{r=1}^{n} \xi^n_{ir}\delta_{\widehat{X}^{n,\mathbf{u}^n,r}},$$

and $\widehat{M}^{n,\mathbf{u}^n,i}_t$ is the time t marginal. Let $\widehat{X}^{n,\mathbf{u}^n,\beta,j}$ denote the dynamic of player j when she changes her policy from $\widehat{\pi}^{n,\mathbf{u}^n,j}$ to β. More specifically, player j follows

$$\widehat{X}^{n,\mathbf{u}^n,\beta,j}_{t+1} \sim P_t(\widehat{X}^{n,\mathbf{u}^n,j}_t, \widehat{M}^{n,\mathbf{u}^n,(\beta,j),j}_t, \beta_t), \qquad \widehat{X}^{n,\mathbf{u}^n,\beta,j}_0 \sim \lambda_{u^n_j},$$

and all other player $i \neq j$ follows

$$\widehat{X}^{n,\mathbf{u}^n,i}_{t+1} \sim P_t(\widehat{X}^{n,\mathbf{u}^n,i}_t, \widehat{M}^{n,\mathbf{u}^n,(\beta,j),i}_t, \widehat{\pi}^{n,\mathbf{u}^n,i}_t(\mathbf{X}^{n,\mathbf{u}^n,\beta,j}_t)), \qquad \widehat{X}^{n,\mathbf{u}^n,i}_0 \sim \lambda_{u^n_i},$$

where the empirical neighborhood measure is

$$\widehat{M}^{n,\mathbf{u}^n,(\beta,j),i} := \frac{1}{n}\left(\sum_{r\neq j}\xi^n_{ir}\delta_{\widehat{X}^{n,\mathbf{u}^n,r}} + \xi_{ij}\delta_{\widehat{X}^{n,\mathbf{u}^n,\beta,j}}\right),$$

and $\mathbf{X}^{n,\mathbf{u}^n,\beta,j}$ denotes the vector $\mathbf{X}^{n,\mathbf{u}^n}$ with the j^{th} element replaced by $\widehat{X}^{n,\mathbf{u}^n,\beta,j}$.

For any $u \in [0,1]$,we define $X^{\pi,u}$ to be the process with marginal $U = u$, controlled by policy π, i.e.,

$$X^{\pi,u}_{t+1} \sim P_t(X^{\pi,u}_t, \mathsf{W}\mu_t(u), \pi(t,u,X^{\pi,u}_t)) \qquad X^{\pi,u}_0 \sim \lambda_u.$$

Proposition G.1. *Assume Assumption 4.8 holds. Let $h : [0,1]\times\mathbb{R}^d\times\mathcal{M}_+(\mathbb{R}^d)\to\mathbb{R}$ be a bounded measurable function such that $h(u,\cdot,\cdot)$ is jointly continuous on $\mathbb{R}^d\times\mathcal{M}_+(\mathbb{R}^d)$ for each fixed u. Then for each $t\in\mathbb{T}$,*

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[h(U^n_i, \widehat{X}^{n,\boldsymbol{U}^n,i}_t, \widehat{M}^{n,\boldsymbol{U}^n,i}_t)] \to \mathbb{E}[h(U, X_t, \mathsf{W}\mu_t(U))]. \tag{21}$$

Proof. Expand the underlying probability space such that it supports independent random elements $(U^n_i, Y^{n,i})$, $\forall i\in[n]$, independent of $\widehat{\mathbf{X}}^{n,\mathbf{u}^n}$ and (U,X), and the law satisfies

$$\mathcal{L}(Y^{n,i}|U^n_i = u) = \mathcal{L}(X|U=u), \qquad \forall u\in I^n_i.$$

Equivalently, this means for any $u\in I^n_i$, the conditional law satisfies

$$Y^{n,i}_{t+1}|(U^n_i = u) \sim P(t, Y^{n,i}_t, \mathsf{W}\mu_t(u), \pi_t(u, Y^{n,i}_t)).$$

In particular for every measurable $\phi : [0,1]\times\mathcal{C}\to\mathbb{R}$,

$$\langle\mu,\phi\rangle = \mathbb{E}\phi(U,X) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\phi(U^n_i, Y^{n,i}). \tag{22}$$

Define the empirical neighborhood measure:

$$M^{n,i} := \frac{1}{n}\sum_{j=1}^{n}\xi^n_{ij}\delta_{Y^{n,j}} = \frac{1}{n}\sum_{j=1}^{n}W_{\xi^n}(U^n_i, U^n_j)\delta_{Y^{n,j}},$$

and the empirical label-state joint measure:

$$\mu^n := \frac{1}{n}\sum_{j=1}^{n}\delta_{(U^n_i, Y^{n,i})}.$$

The theorem is then shown in the following two stages:

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[h(U_i^n,\widehat{X}_t^{n,\mathbf{U}^n,i},\widehat{M}_t^{n,\mathbf{U}^n,i})]\to\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[h(U_i^n,Y_t^{n,i},M_t^{n,i})]\to\mathbb{E}[h(U,X_t,\mathsf{W}\mu_t(U))].$$

Step i. We first show that $\mathsf{W}_{\xi^n}\mu^n(U)\Rightarrow\mathsf{W}\mu(U)$ in probability. Fix a bounded continuous function $\phi:\mathbb{R}^d\to[-1,1]$, it suffices to show $\langle\mathsf{W}_{\xi^n}\mu^n(U),\phi\rangle\to\langle\mathsf{W}\mu(U),\phi\rangle$ in probability. This is divided into two substeps. We first claim that $\langle\mathsf{W}_{\xi^n}\mu^n(U),\phi\rangle-\mathbb{E}[\langle\mathsf{W}_{\xi^n}\mu^n(U),\phi\rangle|U]\to0$ in probability. Note that

$$\langle\mathsf{W}_{\xi^n}\mu^n(u),\phi\rangle=\frac{1}{n}\sum_{j=1}^{n}W_{\xi^n}(u,U_j^n)\phi(Y^{n,i}).$$

For $u\in I_i^n$, by the independence of $Y^{n,i}$,

$$\mathrm{var}(\langle\mathsf{W}_{\xi^n}\mu^n(U),\phi\rangle|U=u)=\mathrm{var}(\frac{1}{n}\sum_{j=1}^{n}\xi_{ij}^n\phi(Y^{n,i}))\le\frac{1}{n^2}\sum_{j=1}^{n}(\xi_{ij}^n)^2.$$

Then, by Assumption (11),

$$\begin{aligned}&\mathbb{E}\left[(\langle\mathsf{W}_{\xi^n}\mu^n(U),\phi\rangle-\mathbb{E}[\langle\mathsf{W}_{\xi^n}\mu^n(U),\phi\rangle|U])^2\right]\\&=\mathbb{E}\left[\mathrm{var}(\langle\mathsf{W}_{\xi^n}\mu^n(U),\phi\rangle|U)\right]\\&=\sum_{i=1}^{n}\int_{I_i^n}\mathrm{var}(\langle\mathsf{W}_{\xi^n}\mu^n(U),\phi\rangle|U=u)du\\&\le\frac{1}{n^3}\sum_{i,j=1}^{n}(\xi_{ij}^n)^2\to0.\end{aligned}$$

Thus, the convergence is in L^2. In the second substep we show that $\mathbb{E}[\langle\mathsf{W}_{\xi^n}\mu^n(U),\phi\rangle|U]\to\langle\mathsf{W}\mu(U),\phi\rangle$ in probability. By the independence of $(U_i^n,Y^{n,i})$,

$$\begin{aligned}\mathbb{E}[\langle\mathsf{W}_{\xi^n}\mu^n(U),\phi\rangle|U=u]&=\mathbb{E}[\frac{1}{n}\sum_{j=1}^{n}W_{\xi^n}(u,U_j^n)\phi(Y^{n,i})]\\&=\mathbb{E}[W_{\xi^n}(u,U)\phi(X)]\\&=\int_{[0,1]}W_{\xi^n}(u,v)\mathbb{E}[\phi(X)|U=v]dv,\end{aligned}$$

where we used the identity (22). Similarly,

$$\langle\mathsf{W}\mu(u),\phi\rangle=\mathbb{E}[W(u,U)\phi(X)]=\int_{[0,1]}W(u,v)\mathbb{E}[\phi(X)|U=v]dv.$$

Thus,

$$\begin{aligned}\mathbb{E}\left[|\mathbb{E}[\langle\mathsf{W}_{\xi^n}\mu^n(U),\phi\rangle|U]-\langle\mathsf{W}\mu(U),\phi\rangle|\right]&=\int_{[0,1]}\left|\int_{[0,1]}(W_{\xi^n}(u,v)-W(u,v))\mathbb{E}[\phi(X)|U=v]dv\right|du\\&=\left\|(\mathsf{W}_{\xi^n}-\mathsf{W})\phi\right\|_{L^1[0,1]}.\end{aligned}$$

By the assumption that $W_{\xi^n}\to W$ in the strong operator topology, the right-hand side goes to 0 and thus $\mathbb{E}[\langle\mathsf{W}_{\xi^n}\mu^n(U),\phi\rangle|U]\to\langle\mathsf{W}\mu(U),\phi\rangle$ in L^1. This concludes the first step.

Step ii. We next show by induction the following holds for each $t \in \mathbb{T}$:

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[h(U_i^n, \widehat{X}_t^{n,\mathbf{U}^n,i}, \widehat{M}_t^{n,\mathbf{U}^n,i})] \to \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[h(U_i^n, Y_t^{n,i}, M_t^{n,i})]. \tag{23}$$

This is trivially true at time 0, since $\widehat{X}^{n,\mathbf{U}^n,i}$ are initialized independently, we have $\mathcal{L}(U_i^n, \widehat{X}_0^{n,\mathbf{U}^n,i}) = \mathcal{L}(U_i^n, Y_0^{n,i})$, and thus

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[h(U_i^n, \widehat{X}_0^{n,\mathbf{U}^n,i}, \widehat{M}_0^{n,\mathbf{U}^n,i})] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[h(U_i^n, Y_0^{n,i}, M_0^{n,i})].$$

Now assume (23) holds for time $t-1$. We have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\Big(\mathbb{E}[h(U_i^n, \widehat{X}_t^{n,\mathbf{U}^n,i}, \widehat{M}_t^{n,\mathbf{U}^n,i})] - \mathbb{E}[h(U_i^n, Y_t^{n,i}, M_t^{n,i})]\Big)\\
&\le \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{E}[h(U_i^n, \widehat{X}_t^{n,\mathbf{U}^n,i}, \widehat{M}_t^{n,\mathbf{U}^n,i})] - \mathbb{E}[h(U_i^n, \widehat{X}_t^{n,\mathbf{U}^n,i}, M_t^{n,i})]\right)}_{\mathrm{I}}\\
&\quad + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{E}[h(U_i^n, \widehat{X}_t^{n,\mathbf{U}^n,i}, M_t^{n,i})] - \mathbb{E}[h(U_i^n, Y_t^{n,i}, M_t^{n,i})]\right)}_{\mathrm{II}}.
\end{aligned}$$

Denote $\mathcal{F}_t^n := \sigma(\{U_i^n\}_{i=1}^n, \{\widehat{X}_s^{n,\mathbf{U}^n,i}\}_{i=1}^n, \{Y_s^{n,i}\}_{i=1}^n, s \le t)$. For term I, we note that $\widehat{X}_t^{n,\mathbf{U}^n,i}$ and $\widehat{X}_t^{n,\mathbf{U}^n,j}$ are independent conditional on $\mathcal{F}_{t-1}^n$, and

$$\begin{aligned}
&\mathbb{E}[h(U_i^n, \widehat{X}_t^{n,\mathbf{U}^n,i}, \widehat{M}_t^{n,\mathbf{U}^n,i}) - h(u, \widehat{X}_t^{n,\mathbf{U}^n,i}, M_t^{n,i}) \,|\, \mathcal{F}_{t-1}^n, \widehat{X}_t^{n,\mathbf{U}^n,i}]\\
&= \int_{(\mathbb{R}^d)^{n-1}} h\Big(U_i^n, \widehat{X}_t^{n,\mathbf{U}^n,i}, \frac{1}{n}\sum_{j=1}^{n}\xi_{ij}^n\delta_{x^j}\Big)\prod_{j\neq i}\widehat{P}_{t-1}^{n,\mathbf{U}^n,j}(dx^j)\\
&\quad - \int_{(\mathbb{R}^d)^{n-1}} h\Big(U_i^n, \widehat{X}_t^{n,\mathbf{U}^n,i}, \frac{1}{n}\sum_{j=1}^{n}\xi_{ij}^n\delta_{y^j}\Big)\prod_{j\neq i}P_{t-1}^{n,j}(dy^j),
\end{aligned}$$

where we use a shorthand notation:

$$\begin{aligned}
\widehat{P}_s^{n,\mathbf{U}^n,i} &:= P_s(\widehat{X}_s^{n,\mathbf{U}^n,i}, \widehat{M}_s^{n,\mathbf{U}^n,i}, \pi_s(U_i^n, \widehat{X}_s^{n,\mathbf{U}^n,i})),\\
P_s^{n,i} &:= P_s(Y_s^{n,i}, \mathsf{W}\mu_s(U_i^n), \pi_s(U_i^n, Y_s^{n,i})).
\end{aligned}$$

More specifically, define the function $h' : [0,1] \times \mathbb{R}^d \times \mathcal{M}_+(\mathbb{R}^d)$ as follows:

$$h'(u,x,m) := \int_{(\mathbb{R}^d)^{n-1}} h\Big(U_i^n, \widehat{X}_t^{n,\mathbf{U}^n,i}, \frac{1}{n}\sum_{j=1}^{n}\xi_{ij}^n\delta_{x^j}\Big)\prod_{j\neq i}P_{t-1}(x, m, \pi_{t-1}(u,x))(dx^j).$$

Then

$$\mathrm{I} = \frac{1}{n}\sum_{i=1}^{n}\Big(h'(U_i^n, \widehat{X}_{t-1}^{n,\mathbf{U}^n,i}, \widehat{M}_{t-1}^{n,\mathbf{U}^n,i}) - h'(U_i^n, Y_{t-1}^{n,i}, \mathsf{W}\mu_{t-1}(U_i^n))\Big).$$

Similarly for II, note that $Y^{n,i}$ are independent; we have

$$\begin{aligned}
&\mathbb{E}[h(u, \widehat{X}_t^{n,\mathbf{U}^n,i}, M_t^{n,i}) - h(U_i^n, Y_t^{n,i}, M_t^{n,i}) \,|\, \mathcal{F}_{t-1}^n, Y_t^{n,j},\, j \neq i]\\
&= \int_{(\mathbb{R}^d)^{n-1}} h(U_i^n, x^i, M_t^{n,i})\widehat{P}_{t-1}^{n,\mathbf{U}^n,i}(dx^i) - h(U_i^n, y^i, M_t^{n,i})P_{t-1}^{n,i}(dy^i).
\end{aligned}$$

By defining the function $h^{''}:[0,1]\times\mathbb{R}^d\times\mathcal{M}_+(\mathbb{R}^d)$ as follows:

$$h^{''}(u,x,m):=\int_{(\mathbb{R}^d)^{n-1}}h(u,x^i,M_t^{n,i})P_{t-1}(x,m,\pi_{t-1}(u,x))(dx^i),$$

we get

$$\mathrm{II}\le\frac{1}{n}\sum_{i=1}^n\left(h^{''}(U_i^n,\widehat{X}_{t-1}^{n,\mathbf{U}^n,i},\widehat{M}_{t-1}^{n,\mathbf{U}^n,i})-h^{''}(U_i^n,Y_{t-1}^{n,i},\mathsf{W}\mu_{t-1}(U_i^n))\right).$$

Note that by the assumption that h and P are continuous, $h^{'}(t,\cdot,\cdot)$ and $h^{''}(t,\cdot,\cdot)$ are jointly continuous for every $t\in\mathbb{T}$. Combining I and II, by the tower property,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n\left(\mathbb{E}[h(U_i^n,\widehat{X}_t^{n,\mathbf{U}^n,i},\widehat{M}_t^{n,\mathbf{U}^n,i})]-\mathbb{E}[h(U_i^n,Y_t^{n,i},M_t^{n,i})]\right)\\
&\le\mathrm{I}+\mathrm{II}\\
&\le\underbrace{\frac{1}{n}\sum_{i=1}^n\mathbb{E}[(h^{'}+h^{''})(U_i^n,\widehat{X}_{t-1}^{n,\mathbf{U}^n,i},\widehat{M}_{t-1}^{n,\mathbf{U}^n,i})]-\frac{1}{n}\sum_{i=1}^n\mathbb{E}[(h^{'}+h^{''})(U_i^n,Y_{t-1}^{n,i},M_{t-1}^{n,i})]}_{\mathrm{I}'}\\
&\quad+\underbrace{\frac{1}{n}\sum_{i=1}^n\mathbb{E}[(h^{'}+h^{''})(U_i^n,Y_{t-1}^{n,i},M_{t-1}^{n,i})]-\frac{1}{n}\sum_{i=1}^n\mathbb{E}[(h^{'}+h^{''})(U_i^n,Y_{t-1}^{n,i},\mathsf{W}\mu_{t-1}(U_i^n))]}_{\mathrm{II}''}.
\end{aligned}$$

By our assumption on time $t-1$, $\mathrm{I}'\to 0$. In Step i we proved that $\mathsf{W}_{\xi^n}\mu^n(U)\to\mathsf{W}\mu(U)$. It is straightforward to show $\mathsf{W}_{\xi^n}\mu^n(U_i^n)\to\mathsf{W}\mu(U_i^n)$ with the same line of reasoning. Rewrite $W_{\xi^n}\mu^n(U_i^n)=M^{n,i}$, we actually have $M^{n,i}\to\mathsf{W}\mu(U_i^n)$ in probability. Combined with the boundedness of integrand, the convergences is in L^1 and thus $\mathrm{II}'\to 0$.

Step iii. Finally, we aim to show,

$$\frac{1}{n}\sum_{i=1}^n\mathbb{E}h(U_i^n,Y^{n,i},M^{n,i})\to\mathbb{E}[h(U,X,\mathsf{W}\mu(U))].$$

This is justified with similar argument as in (Lacker & Soret, 2023, Theorem 6.1), and this concludes the theorem. □

G.2. Proof of Theorem 4.9

Recall the definition of $\boldsymbol{\epsilon}^n(\mathbf{u}^n)$ in Definition 3.1; we have

$$\begin{aligned}
\epsilon_i^n(\mathbf{u}^n)&:=\sup_{\beta\in\mathcal{A}_n}J_i(\pi^{n,\mathbf{u}^n,1},\dots,\pi^{n,\mathbf{u}^n,i-1},\beta,\pi^{n,\mathbf{u}^n,i+1},\dots,\pi^{n,\mathbf{u}^n,n})-J_i(\boldsymbol{\pi}^{n,\mathbf{u}^n})\\
&\le\sup_{\beta\in\mathcal{A}_n}\Delta_1^{n,i}(\beta,\mathbf{u}^n)+\sup_{\beta\in\mathcal{A}_n}\Delta_2^{n,i}(\beta,\mathbf{u}^n)+\sup_{\beta\in\mathcal{A}_n}\Delta_3^{n,i}(\beta,\mathbf{u}^n)+\Delta_4^{n,i}(\mathbf{u}^n)+\Delta_5^{n,i}(\mathbf{u}^n),
\end{aligned}$$

where

$$\Delta_1^{n,i}(\beta,\mathbf{u}^n) := \mathbb{E}\left[\sum_{t\in\mathbb{T}} f^i(t,\widehat{X}_t^{n,\mathbf{u}^n,\beta,i},\widehat{M}_t^{n,\mathbf{u}^n,(\beta,i),i},\beta_t) + g(\widehat{X}_T^{n,\mathbf{u}^n,\beta,i},\widehat{M}_T^{n,\mathbf{u}^n,(\beta,i),i})\right]$$
$$- \mathbb{E}\left[\sum_{t\in\mathbb{T}} f^i(t,\widehat{X}_t^{n,\mathbf{u}^n,\beta,i},\mathsf{W}\mu_t(u_i^n),\beta_t) + g(\widehat{X}_T^{n,\mathbf{u}^n,\beta,i},\mathsf{W}\mu_T(u_i^n))\right],$$
$$\Delta_2^{n,i}(\beta,\mathbf{u}^n) := \mathbb{E}\left[\sum_{t\in\mathbb{T}} f^i(t,\widehat{X}_t^{n,\mathbf{u}^n,\beta,i},\mathsf{W}\mu_t(u_i^n),\beta_t) + g(\widehat{X}_T^{n,\mathbf{u}^n,\beta,i},\mathsf{W}\mu_T(u_i^n))\right]$$
$$- \mathbb{E}\left[\sum_{t\in\mathbb{T}} f^i(t,X_t^{\beta,u_i^n},\mathsf{W}\mu_t(u_i^n),\beta_t) + g(X_T^{\beta,u_i^n},\mathsf{W}\mu_T(u_i^n))\right],$$
$$\Delta_3^{n,i}(\beta,\mathbf{u}^n) := \mathbb{E}\left[\sum_{t\in\mathbb{T}} f^i(t,X_t^{\beta,u_i^n},\mathsf{W}\mu_t(u_i^n),\beta_t) + g(X_T^{\beta,u_i^n},\mathsf{W}\mu_T(u_i^n))\right]$$
$$- \mathbb{E}\left[\sum_{t\in\mathbb{T}} f^i(t,X^{u_i^n},\mathsf{W}\mu_t(u_i^n),\pi_t(u_i^n,\widehat{X}_t^{n,\mathbf{u}^n,i})) + g(X^{u_i^n},\mathsf{W}\mu_T(u_i^n))\right],$$
$$\Delta_4^{n,i}(\mathbf{u}^n) := \mathbb{E}\left[\sum_{t\in\mathbb{T}} f^i(t,X^{u_i^n},\mathsf{W}\mu_t(u_i^n),\pi_t(u_i^n,\widehat{X}_t^{n,\mathbf{u}^n,i})) + g(X^{u_i^n},\mathsf{W}\mu_T(u_i^n))\right]$$
$$- \mathbb{E}\left[\sum_{t\in\mathbb{T}} f^i(t,\widehat{X}_t^{n,\mathbf{u}^n,i},\mathsf{W}\mu_t(u_i^n),\pi_t(u_i^n,\widehat{X}_t^{n,\mathbf{u}^n,i})) + g(\widehat{X}_T^{n,\mathbf{u}^n,i},\mathsf{W}\mu_T(u_i^n))\right],$$
$$\Delta_5^{n,i}(\mathbf{u}^n) := \mathbb{E}\left[\sum_{t\in\mathbb{T}} f^i(t,\widehat{X}_t^{n,\mathbf{u}^n,i},\mathsf{W}\mu_t(u_i^n),\pi_t(u_i^n,\widehat{X}_t^{n,\mathbf{u}^n,i})) + g(\widehat{X}_T^{n,\mathbf{u}^n,i},\mathsf{W}\mu_T(u_i^n))\right]$$
$$- \mathbb{E}\left[\sum_{t\in\mathbb{T}} f^i(t,\widehat{X}_t^{n,\mathbf{u}^n,i},\widehat{M}_t^{n,\mathbf{u}^n,i},\pi_t(u_i^n,\widehat{X}_t^{n,\mathbf{u}^n,i})) + g(\widehat{X}_T^{n,\mathbf{u}^n,i},\widehat{M}_T^{n,\mathbf{u}^n,i})\right].$$

β_t in these formulae is short for $\beta_t(\widehat{\mathbf{X}}_t^{n,\mathbf{u}^n,\beta,i})$, for a closed-loop control $\beta : \mathbb{T}\times\mathbb{R}^d \to \mathcal{P}(A)$.

Lemma G.2 ((Lacker & Soret, 2023, Lemma 5.1)). *Fix $\mu \in \mathcal{P}_{\text{unif}}([0,1]\times\mathcal{C})$, $u\in[0,1]$. For any policy $\pi\in\mathcal{A}_1$ and $m\in\mathcal{P}(\mathbb{R}^d)$, define*

$$X_{t+1}^{m,\pi} \sim P(t,X_t^{m,\pi},\mathsf{W}\mu_t(u),\pi(t,X_t^{m,\pi})), \qquad X_0^{m,\pi}\sim m,$$

and

$$J_W^{u,m}(\mu,\pi) := \mathbb{E}\left[\sum_{t\in\mathbb{T}} f(t,X_t^{m,\pi},\mathsf{W}\mu_t(u),\pi(t,X_t^{m,\pi})) + g(X_T^{m,\pi},\mathsf{W}\mu_T(u))\right].$$

If $\pi\in\mathcal{A}_U$ is an optimal policy, in the sense that $J_W(\mu,\pi)\geq J_W(\mu,\beta)$ for all $\beta\in\mathcal{A}_U$, then

$$J_W^{u,\lambda_u}(\mu,\pi_u) = \sup_{\beta\in\mathcal{A}_1} J_W^{u,\lambda_u}(\mu,\beta), \tag{24}$$

where $\pi_u(t,x) := \pi(t,u,x)$.

Remark G.3. With a similar notation as in the proof for equilibrium existence (Appendix E.2), we may denote

$$\mathcal{R}_{u,\lambda_u}(\mu) := \{R\in\mathcal{R}(\mu)\subset\mathcal{P}(\mathcal{V}\times[0,1]\times\mathcal{C}) : R\circ U^{-1} = \delta_{u_i^n},\, R\circ X_0^{-1} = \lambda_u\}.$$

The joint law $\mathcal{L}(\pi, u, X^{\lambda_u,\pi}) \in \mathcal{R}_{u,\lambda_u}(\mu)$, and for any $\beta \in \mathcal{V}_U$, $\mathcal{L}(\beta, u, X^{\lambda_u,\beta}) \in \mathcal{R}_{u,\lambda_u}(\mu)$; note that β can be any open-loop policy. Thus, (24) can be rewritten as[5]

$$\langle \mathcal{L}(\pi, u, X^{\lambda_u,\pi}), \Xi^\mu \rangle \geq \langle R, \Xi^\mu \rangle, \qquad \forall R \in \mathcal{R}_{u,\lambda_u}(\mu),$$

where Ξ^μ is defined in (17). This view simplifies the analysis of the following lemma.

Lemma G.4. $\sup_{\beta\in\mathcal{A}_n} \Delta_3^{n,i}(\beta, \boldsymbol{u}^n) \leq 0$ *for a.e.* $\boldsymbol{u}^n \in [0,1]^n$ *and all* $i \in [n]$.

Proof. By construction, $\mathcal{L}(X^{u_i^n}) = \mathcal{L}(X^{\lambda_{u_i^n},\pi_{u_i^n}})$, then the second term of $\Delta_3^{n,i}(\beta, \mathbf{u}^n)$ is actually $J_W^{u_i^n,\lambda_{u_i^n}}(\mu, \pi_u)$. On the other hand, the joint law $\mathcal{L}(\beta, u_i^n, X^{\beta,u_i^n}) \in \mathcal{R}_{u_i^n,\lambda_{u_i^n}}(\mu)$. Thus, by Remark G.3, we deduce

$$\sup_{\beta\in\mathcal{A}_n} \Delta_3^{n,i}(\beta, \mathbf{u}^n) \leq \sup_{\beta\in\mathcal{A}_1} J_W^{u_i^n,\lambda_{u_i^n}}(\mu, \pi_u^*) - J_W^{u_i^n,\lambda_{u_i^n}}(\mu, \pi_u^*).$$

Following (24) in Lemma G.2, the right-hand side is ≤ 0 for a.e. $\mathbf{u}^n \in [0,1]^n$ and all $i \in [n]$. □

Taking average, we have

$$\frac{1}{n}\sum_{i=1}^n \epsilon_i^n(\mathbf{u}^n) \leq \frac{1}{n}\sum_{i=1}^n \sup_{\beta\in\mathcal{A}_n} \Delta_1^{n,i}(\beta, \mathbf{u}^n) + \frac{1}{n}\sum_{i=1}^n \sup_{\beta\in\mathcal{A}_n} \Delta_2^{n,i}(\beta, \mathbf{u}^n) + \frac{1}{n}\sum_{i=1}^n \Delta_4^{n,i}(\mathbf{u}^n) + \frac{1}{n}\sum_{i=1}^n \Delta_5^{n,i}(\mathbf{u}^n). \tag{25}$$

By Assumption 4.4, it's straightforward to see that $\{\mathcal{L}(\widehat{X}^{n,\mathbf{u}^n,\beta,i}) : (n, \mathbf{u}^n, \beta, i) \in \mathbb{N}_+ \times \mathbf{I}^n \times \mathcal{V} \times [n]\}$ is a tight collection of measures in $\mathcal{P}(\mathcal{C})$. Let $K \subset \mathcal{C}$ be a compact subset such that $\sup_{n,\mathbf{u}^n,\beta,i} \mathbb{P}(\widehat{X}^{n,\mathbf{u}^n,\beta,i} \notin K) \leq \eta$ for some fixed $\eta > 0$. Define function $h_1 : [0,1] \times \mathcal{M}_+(\mathcal{C}) \to \mathbb{R}$ by

$$h_1(u, m) := \sum_{t\in\mathbb{T}} \sup_{a\in A} \sup_{z\in K} \Big(|f(t, z_t, \mathsf{W}\mu_t(u), a) - f(t, z_t, m_t, a)| + |g(z_T, \mathsf{W}\mu_T(u)) - g(z_T, m_T)| \Big).$$

Similarly, define

$$h_2(u, x) := \sum_{t\in\mathbb{T}} \sup_{a\in A} \sup_{z\in K} \Big(\big|\mathbb{E} f^i(t, z_t, \mathsf{W}\mu_t(u), a) - f^i(t, x_t, \mathsf{W}\mu_t(u), a)\big| - |\mathbb{E} g(z_T, \mathsf{W}\mu_T(u)) - g(x_T, \mathsf{W}\mu_T(u))| \Big).$$

Function h_1 and h_2 are bounded measurable since f and g are bounded continuous (Aliprantis & Border, 2006, Theorem 18.19). Moreover, it follows from the compactness of A and K that $h_1(u, \cdot)$ is continuous on $\mathcal{M}_+(\mathcal{C})$ for a.e. u, and $h_2(u, \cdot)$ is continuous on E for a.e. u. Note that $(h_1, h_2)(U, X_t, \mathsf{W}\mu(U)) = 0$. We may thus use h_1 to bound Δ_1 and Δ_5, use h_2 to bound Δ_2 and Δ_4. To address the region outside K, let C be a constant such that $\max(|f|, |g|) \leq C$, and (25) becomes

$$\frac{1}{n}\sum_{i=1}^n \epsilon_i^n(\mathbf{u}^n) \leq \frac{2}{n}\sum_{i=1}^n \mathbb{E}\left[h_1(u_i^n, \widehat{M}_t^{n,\mathbf{u}^n,i})\right] + \frac{2}{n}\sum_{i=1}^n \mathbb{E}\left[h_2(u_i^n, \widehat{X}_t^{n,\mathbf{u}^n,i})\right] + 8\eta C.$$

By Proposition G.1,

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \epsilon_i^n(\mathbf{U}^n)\right] &\leq \frac{2}{n}\sum_{i=1}^n \mathbb{E}\left[h_1(U_i^n, \widehat{M}_t^{n,\mathbf{U}^n,i})\right] + \frac{2}{n}\sum_{i=1}^n \mathbb{E}\left[h_2(U_i^n, \widehat{X}_t^{n,\mathbf{U}^n,i})\right] + 8\eta C \\
&\longrightarrow 2\mathbb{E}\left[h_1(U, \mathsf{W}\mu_t(U))\right] + 2\mathbb{E}\left[h_2(U, X_t)\right] + 8\eta C \\
&= 8\eta C.
\end{aligned}$$

The proof for theorem 4.9 is concluded by letting $\eta \to 0$.

[5] Indeed, it might not be directly obvious why $\langle \mathcal{L}(\pi, u, X^{\lambda_u,\pi}), \Xi^\mu \rangle \geq \langle R, \Xi^\mu \rangle$ holds for all $R \in \mathcal{R}_{u,\lambda_u}(\mu)$, since R might induce open-loop policies while the supremum in (24) is over $\mathcal{A}_1$. This actually can be showed rigorously, however, and the reader may refer to (Lacker & Soret, 2023, Lemma 5.1) for a proof.

H. Proof for Online Learning Sample Complexity

H.1. A Concrete Algorithm Realization

For clarity, we present Algorithm 1 cobined with subroutine (15) in Algorithm 2.

Algorithm 2 Oracle-free Learning for GMFG

Initialize $Q^{0,0} = \{Q_d^{0,0}\}_{d=1}^D$ and $M^{0,0} = \{M_d^{0,0}\}_{d=1}^D$
for $k \leftarrow 0$ to $K-1$ **do**
  for $d \leftarrow 1$ to D **do**
    Sample initial state $x_0 \sim M_d^{k,0}$, action $a_0 \sim \Gamma_\pi(Q_d^{k,0})$
    for $\tau \leftarrow 0$ to $H-1$ **do**
      Sample reward $r_\tau = f(x_\tau, \mathbf{W\Pi}_D M^{k,0}(u_d), a_\tau)$, next state $x_{\tau+1} \sim P(x_\tau, \mathbf{W\Pi}_D M^{k,0}(u_d), a_\tau)$, and action $a_{\tau+1} \sim \Gamma_\pi(Q_d^{k,\tau})$
      Update Q-function:
      $Q_d^{k,\tau+1}(x_\tau, a_\tau) \leftarrow (1-\alpha_\tau) Q_d^{k,\tau}(x_\tau, a_\tau) + \alpha_\tau \left(r_\tau + \gamma Q_d^{k,\tau}(x_{\tau+1}, a_{\tau+1})\right)$
      Update population measure:
      $M_d^{k,\tau+1} \leftarrow (1-\beta_\tau) M_d^{k,\tau} + \beta_\tau \delta_{x_{\tau+1}}$
    end for
    Let $Q_d^{k+1,0} = Q_d^{k,H}$ and $M_d^{k+1,0} = M_d^{k,H}$
  end for
end for
Return policy $\pi^{(K)} := \Gamma_\pi(\mathbf{\Pi}_D Q^{K,0})$ and population measure $\mu^{(K)} := \mathbf{\Pi}_D M^{K,0}$, where $Q^{K,0} = \{Q_d^{K,0}\}_{d=1}^D$ and $M^{K,0} = \{M_d^{K,0}\}_{d=1}^D$

Algorithm 3 is adapted from Algorithm 2 to solve GMFGs with finite time horizons.

Algorithm 3 Oracle-free Learning for Finite Horizon GMFG

Initialize $Q^{0,0:T} = \{Q_d^{0,0:T}\}_{d=1}^D$ and $M^{0,0:T} = \{M_d^{0,0:T}\}_{d=1}^D$ where T is the time horizon
for $k \leftarrow 0$ to $K-1$ **do**
  for $d \leftarrow 1$ to D **do**
    Sample initial state $x_0 \sim M_d^{k,0}$ and action $a_0 \sim \Gamma_\pi(Q_d^{k,0})$
    for $t \leftarrow 0$ to $T-1$ **do**
      Sample reward $r_t = f(x_t, \mathbf{W\Pi}_D M^{k,t}(u_d), a_t)$, next state $x_{t+1} \sim P(x_t, \mathbf{W\Pi}_D M^{k,t}(u_d), a_t)$, and action $a_{t+1} \sim \Gamma_\pi(Q_d^{k,t+1})$
      Update population measure:
      $M_d^{k,t+1} \leftarrow (1-\beta_k) M_d^{k,t} + \beta_k \delta_{x_{t+1}}$
      Update Q-function:
      $Q_d^{k,t}(x_t, a_t) \leftarrow (1-\alpha_k) Q_d^{k,t}(x_t, a_t) + \alpha_k \left(r_t + \gamma Q_d^{k,t+1}(x_{t+1}, a_{t+1})\right)$
    end for
  end for
end for
Return policy $\pi^{(K)} := \Gamma_\pi(\mathbf{\Pi}_D Q^{K,0:T})$ and population measure $\mu^{(K)} := \mathbf{\Pi}_D M^{K,0:T}$, where $Q^{K,0:T} = \{Q_d^{K,0:T}\}_{d=1}^D$ and $M^{K,0:T} = \{M_d^{K,0:T}\}_{d=1}^D$

The difference between two algorithms lies in the learning rate. In Algorithm 3, the learning rate has to capture each time step t in the time horizon T. Therefore, we have $\beta_k = \frac{1}{1+\#(t,k)}$ and $\alpha_k = \frac{1}{1+\#(x,a,t,k)}$, where $\#(t,k)$ counts the number of visits to time step t up to epoch k. $\#(x,a,t,k)$ counts the number of visits to tuple (x,a,t) up to epoch k.

H.2. Discretization of Label Space

Recall that for the sample complexity analysis, we consider a finite state space $\mathcal{X}$ and action space A. Also recall that $\mathcal{U} \coloneqq \{u_1, \dots, u_D\}$ is the discretization of label space, and $\Pi_D : [0,1] \to \mathcal{U}$ is the projection mapping. For notation consistency, we use a tilde above any function (operator, measure, set, etc.) defined on $[0,1]$ to denote their counterparts defined on $\mathcal{U}$, and a hat over an operator to denote the algorithmic approximation of it. We define the operator $\mathbf{\Pi}_D$ which maps operators defined on $\mathcal{U}$ to operators defined on $[0,1]$: for any operator $\widetilde{\phi}$ defined on $\mathcal{U}$,

$$\mathbf{\Pi}_D \widetilde{\phi}(u) \coloneqq \sum_{d=1}^{D} \widetilde{\phi}(u_d) 1_{\{u \in I_{u_d}\}}.$$

In particular, for $\widetilde{\mu} = \{\widetilde{\mu}^{u_d}\}_{d=1}^{D} \in \mathcal{M}(\mathcal{X})^{\mathcal{U}}$, we regard $\mathbf{\Pi}_D \widetilde{\mu}$ as both the kernel $\nu : [0,1] \to \mathcal{M}(\mathcal{X})$ given by

$$\nu(u) \coloneqq \sum_{d=1}^{D} \widetilde{\mu}^{u_d} 1_{\{u \in I_{u_d}\}},$$

and also a measure in $\mathcal{M}_{\text{unif}}([0,1] \times \mathcal{X})$, constructed by $\text{Leb} \otimes \nu$. Here we denote $\mathcal{M}(\mathcal{X})$ the collection of all Borel measures with finite variation on $\mathcal{X}$, and $\mathcal{M}_{\text{unif}}([0,1] \times \mathcal{X})$ the collection of all Borel measures with finite variation on $[0,1] \times \mathcal{X}$ with uniform first marginal.

In addition to Π_D, we define a set value mapping $\Pi_D^\dagger : \mathcal{U} \to 2^{[0,1]}$ by $\Pi_D^\dagger(u_d) = I_d$ for any $u_d \in \mathcal{U}$. The operator $\mathbf{\Pi}_D^\dagger$ maps operators defined on $[0,1]$ to operators defined on $\mathcal{U}$. For any operator ϕ defined on $[0,1]$,

$$\mathbf{\Pi}_D^\dagger \phi(u_d) \coloneqq \phi(u_d), \qquad u_d \in \mathcal{U}.$$

Note that $\mathbf{\Pi}_D^\dagger \mathbf{\Pi}_D = \text{Id}_{\mathcal{U}}$, while the inverse is not necessarily true.

Lemma H.1. *The operator norm of $\mathbf{\Pi}_D : \mathcal{M}(\mathcal{X})^{\mathcal{U}} \to \mathcal{M}_{\text{unif}}([0,1] \times \mathcal{X})$ is bounded by 1, where we equip the product space $\mathcal{M}(\mathcal{X})^{\mathcal{U}}$ with the norm $\|\widetilde{\mu}\| = \sup_{u_d \in \mathcal{U}} \|\widetilde{\mu}^{u_d}\|_{\text{TV}}$.*

Proof. It holds for any $\widetilde{\mu} \in \mathcal{M}(\mathcal{X})^{\mathcal{U}}$ that

$$\begin{aligned}
\|\mathbf{\Pi}_D \widetilde{\mu}\|_{\text{TV}} &= \sup_{\|\phi\|_\infty \le 1} \Big| \int_{[0,1]\times\mathcal{X}} \phi(u,x) \mathbf{\Pi}_D \widetilde{\mu}(du, dx) \Big| \\
&\le \sum_{d=1}^{D} \sup_{\|\phi\|_\infty \le 1} \int_{I_{u_d}} \Big| \int_{\mathcal{X}} \phi(u,x) \widetilde{\mu}^{u_d}(dx) \Big| du \\
&\le \sum_{d=1}^{D} \int_{I_{u_d}} \|\widetilde{\mu}^{u_d}\|_{\text{TV}} du \le \sup_{u_d \in \mathcal{U}} \|\widetilde{\mu}^{u_d}\|_{\text{TV}} = \|\widetilde{\mu}\|.
\end{aligned}$$

□

The following lemma ensures that $\mathcal{U}$ is a good approximation of the label space.

Lemma H.2. *Suppose Assumption 5.1(2) holds. For any $\mu \in \mathcal{P}_{\text{unif}}([0,1] \times \mathcal{X})$, we have*

$$\sup_{u \in [0.1]} \|\mathsf{W}\mu(u) - \mathsf{W}\mu(\Pi_D(u))\|_{\text{TV}} \le \frac{L_d}{D}.$$

Proof. Recall the definition of total variation norm,

$$\begin{aligned}
\sup_{u \in [0,1]} \|\mathsf{W}\mu(u) - \mathsf{W}\mu(\Pi_D(u))\|_{\text{TV}} &= \sup_u \sup_{\|\phi\|_\infty \le 1} \left| \int_{[0,1]\times\mathcal{X}} (W(u,v) - W(\Pi_D(u),v)) \phi(x) \mu(dv, dx) \right| \\
&\le \sup_u \int_{[0,1]} \Big| (W(u,v) - W(\Pi_D(u),v)) \Big| dv \\
&\le \frac{L_d}{D},
\end{aligned}$$

□

where the last inequality follows from Assumption 5.1(2).

H.3. Best Response and Induced Population Operator

Recall $\mathcal{Q}$ is the collection of all $[0,1]\times\mathcal{X}\times A\to\mathbb{R}$ functions, for any $\mu\in\mathcal{P}_{\text{unif}}([0,1]\times\mathcal{X})$, the Bellman (optimality) operator $\mathcal{T}_\mu:\mathcal{Q}\to\mathcal{Q}$ is defined by

$$\mathcal{T}_\mu q(u,x,a)=f(x,\mathsf{W}\mu(u),a)+\gamma\langle P(x,\mathsf{W}\mu(u),a),\sup_{a\in A}q(u,\cdot,a)\rangle,$$

for any $q\in\mathcal{Q}$. It is known that $\mathcal{T}_\mu$ is a γ-contraction mapping, thus a unique fixed point exists, denoted as Q^μ. The state value function is $v^\mu(u,x):=\sup_{a\in A}Q^\mu(u,x,a)$.

BR and IP operator. The FPI Γ, given by $\Gamma(\mu)=\Gamma_2(\Gamma_1(\mu),\mu)$, can be alternatively decomposed into the *best response* (BR) w.r.t. the current population and the *induced population* (IP) w.r.t. the current policy. Define the BR operator $\Gamma_{\text{BR}}:\mathcal{P}_{\text{unif}}([0,1]\times\mathcal{X})\to\mathcal{Q}$ by $\Gamma_{\text{BR}}(\mu)=Q^\mu$ where Q^μ is the fixed point of $\mathcal{T}_\mu$.

The IP operator $\Gamma_{\text{IP}}:\mathcal{Q}\times\mathcal{P}_{\text{unif}}([0,1]\times\mathcal{X})\to\mathcal{P}_{\text{unif}}([0,1]\times\mathcal{X})$ is defined by $\Gamma_{\text{IP}}(Q,\mu)=\mathcal{L}(U,X)$ where X follows the Markov transition with population measure μ, and under policy $\Gamma_\pi(Q)$.

Actually, $\Gamma_\pi\circ\Gamma_{\text{BR}}(\mu)=\Gamma_1(\mu)$, and $\Gamma_{\text{IP}}(Q,\mu)=\Gamma_2(\Gamma_\pi(Q),\mu)$, and we have $\Gamma(\mu)=\Gamma_2(\Gamma_1(\mu),\mu)=\Gamma_{\text{IP}}(\Gamma_{\text{BR}}(\mu),\mu)$. However, both Γ_{BR} and Γ_{IP} are defined in terms of $\mathcal{Q}$, where the label space $[0,1]$ is continuous. Thus, we define the following operators with Q-functions on $\mathcal{U}$.

Discretized BR and IP operator. Let $\widetilde{\mathcal{Q}}$ be the collection of all L_2-integrable $\mathcal{U}\times\mathcal{X}\times A\to\mathbb{R}$ functions, we define the discretized BR operator $\widetilde{\Gamma}_{\text{BR}}:\mathcal{P}_{\text{unif}}([0,1]\times\mathcal{X})\to\widetilde{\mathcal{Q}}$ by $\widetilde{\Gamma}_{\text{BR}}(\mu)=\widetilde{Q}^\mu$ which solves the equation

$$\widetilde{Q}^\mu(u_d,x,a)=f(x,\mathsf{W}\mu(u_d),a)+\gamma\langle P(x,\mathsf{W}\mu(u_d),a),\sup_{a\in A}\widetilde{Q}^\mu(u_d,\cdot,a)\rangle,\qquad\forall u_d\in\mathcal{U}.$$

$\widetilde{\Gamma}_{\text{BR}}$ returns D best responses for labels in $\mathcal{U}$ w.r.t. population distribution μ. In particular, $\widetilde{Q}^\mu$ and Q^μ coincide at $\mathcal{U}\times\mathcal{X}\times A$.

The discretized IP operator $\widetilde{\Gamma}_{\text{IP}}:\widetilde{\mathcal{Q}}\times\mathcal{P}_{\text{unif}}([0,1]\times\mathcal{X})\to\mathcal{P}(\mathcal{X})^{\mathcal{U}}$ is defined by $\widetilde{\Gamma}_{\text{IP}}(\widetilde{Q},\mu)=\mathcal{L}(U,X)$ where X follows the Markov transition with population measure μ, and under policy $\Gamma_\pi(\widetilde{Q})$, conditional on $U\in\mathcal{U}$. In other words, it is the induced state distribution on $\mathcal{P}(\mathcal{X})^{\mathcal{U}}$ for the D classes.

For notation simplicity, we denote $\Gamma_{\text{IP}}\Gamma_{\text{BR}}(\mu)=\Gamma_{\text{IP}}(\Gamma_{\text{BR}}(\mu),\mu)$, similarly for $\widetilde{\Gamma}_{\text{IP}}\widetilde{\Gamma}_{\text{BR}}$.

Algorithm operator. Finally, the algorithm operator $\widehat{\Gamma}:\mathcal{P}(\mathcal{X})^{\mathcal{U}}\to\mathcal{P}(\mathcal{X})^{\mathcal{U}}$ is defined by

$$\widehat{\Gamma}:\{M_d^{k,0}\}_{d=1}^D\mapsto\{M_d^{k,H}\}_{d=1}^D.$$

It returns the updated D-class population measure after an outer iteration of Algorithm 2, consisting of H online stochastic updates to the D-class Q- and M-value functions.

Given the initial D-class population estimate $M_0\coloneqq\{M_d^{0,0}\}_{d=1}^D\in\mathcal{P}(\mathcal{X})^{\mathcal{U}}$, we can express Algorithm 2 as

$$\mathbf{\Pi}_D\widehat{\Gamma}^K M_0=\mathbf{\Pi}_D\left(\widehat{\Gamma}\mathbf{\Pi}_D^\dagger\mathbf{\Pi}_D\right)^K M_0=\left(\mathbf{\Pi}_D\widehat{\Gamma}\mathbf{\Pi}_D^\dagger\right)^K\mathbf{\Pi}_D M_0. \tag{26}$$

H.4. Sample Complexity Analysis

Recall that in Assumptions 5.1 and 5.3, L_P, L_f are the Lipschitz constants of transition kernel and reward function, L_d is the constant controlling graphon discretization error, κ is the contraction factor of one step FPI, and c_1, c_2 are the constants associated with the ergodicity. We now give a paraphrase of Theorem 5.4 which includes the dependence of sample complexity on these assumed constants.

Theorem H.3. *Let $\widehat{\mu}$ be the stationary equilibrium measure of the infinite horizon GMFG. Suppose Assumptions 5.1 and 5.3 hold. For any initial estimate $M^{0,0}\in\mathcal{P}(\mathcal{X})^{\mathcal{U}}$, the sample complexity of Algorithm 2 is given by*

$$\begin{aligned}\mathbb{E}\left\|\mathbf{\Pi}_D M^{K,H}-\widehat{\mu}\right\|^2\le O\Bigg(&\exp(-\kappa K)\mathbb{E}\left\|\mathcal{M}^{0,0}-\widehat{\mu}\right\|^2\\&+\frac{1}{\kappa^2}\left(\frac{|\mathcal{X}||A|L_\pi^2L_f^2L_d^2\sigma^2\left(1-\gamma+|f|_\infty\right)^2}{(1-\gamma)^4D^2}+\frac{L_P^2L_d^2\sigma^2}{D^2}+\frac{D|\mathcal{X}|^2|A||f|_\infty^2L_\pi^2\sigma^2\log H}{\theta^2(1-\gamma)^4H}\right)\Bigg),\end{aligned} \tag{27}$$

where $\sigma \coloneqq \widehat{n} + c_1 c_2^{\widehat{n}}/(1-c_2)$, $\widehat{n} = \left\lceil \log_{c_2} c_1^{-1} \right\rceil$, *and*

$$\theta := \inf_{(u,x,a)\in[0,1]\times\mathcal{X}\times A} \inf_{q\in\mathcal{Q}} \mu_q(u,x)\Gamma_\pi(q^u)[a\,|\,x] > 0$$

is the lower bound of the probability of visiting any label-state-action tuple under the steady distribution $\mu_q \in \mathcal{P}_{\mathrm{unif}}([0,1]\times\mathcal{X})$ *induced by any value function* q.

Our analysis follows the following illustration:

$$\underbrace{\boldsymbol{\Pi}_D \widehat{\Gamma}^K M_0}_{\text{Algorithm 2}} \xrightarrow{\text{approximates}} \underbrace{\left(\boldsymbol{\Pi}_D \widetilde{\Gamma}_{\mathrm{IP}} \widetilde{\Gamma}_{\mathrm{BR}}\right)^K \boldsymbol{\Pi}_D M_0}_{\text{Finite-label FPI}} \xrightarrow{\text{approximates}} \underbrace{\left(\Gamma_{\mathrm{IP}}\Gamma_{\mathrm{BR}}\right)^K \boldsymbol{\Pi}_D M_0}_{\text{FPI}},$$

and we first give the one-step approximation error of Algorithm 2.

Proposition H.4 (One-step approximation error)**.** *For any* $\nu \in \boldsymbol{\Pi}_D \mathcal{P}(\mathcal{X})^{\mathcal{U}}$, *we have*

$$\mathbb{E}\left\|\left(\Gamma_{\mathrm{IP}}\Gamma_{\mathrm{BR}} - \boldsymbol{\Pi}_D \widehat{\Gamma} \boldsymbol{\Pi}_D^\dagger\right)\nu\right\|_{\mathrm{TV}}^2 = O\left(\frac{D\log H}{H} + \frac{1}{D^2}\right).$$

Proof. Consider the decomposition

$$\begin{aligned}
\mathbb{E}\left\|\left(\Gamma_{\mathrm{IP}}\Gamma_{\mathrm{BR}} - \boldsymbol{\Pi}_D \widehat{\Gamma} \boldsymbol{\Pi}_D^\dagger\right)\nu\right\|_{\mathrm{TV}}^2 \le & 3\mathbb{E}\underbrace{\left\|\Gamma_{\mathrm{IP}}\left(\Gamma_{\mathrm{BR}} - \boldsymbol{\Pi}_D\widetilde{\Gamma}_{\mathrm{BR}}\right)\nu\right\|_{\mathrm{TV}}^2}_{G_1} \\
& + 3\mathbb{E}\underbrace{\left\|\left(\Gamma_{\mathrm{IP}}\boldsymbol{\Pi}_D - \boldsymbol{\Pi}_D\widetilde{\Gamma}_{\mathrm{IP}}\right)\widetilde{\Gamma}_{\mathrm{BR}}\nu\right\|_{\mathrm{TV}}^2}_{G_2} \\
& + 3\underbrace{\mathbb{E}\left\|\boldsymbol{\Pi}_D\left(\widetilde{\Gamma}_{\mathrm{IP}}\widetilde{\Gamma}_{\mathrm{BR}} - \widehat{\Gamma}\boldsymbol{\Pi}_D^\dagger\right)\nu\right\|_{\mathrm{TV}}^2}_{G_3}.
\end{aligned}$$

Note that the kernel resulting from disintegration is only Lebesgue a.e. defined. However, we only consider those $\nu \in \boldsymbol{\Pi}_D\mathcal{P}(\mathcal{X})^{\mathcal{U}}$, i.e., there exists some $M \in \mathcal{P}(\mathcal{X})^{\mathcal{U}}$ such that $\nu = \boldsymbol{\Pi}_D M$, and thus $\boldsymbol{\Pi}_D^\dagger \nu = M$ is unique without ambiguity.

Let $q \coloneqq \Gamma_{\mathrm{BR}}\nu$ and $\mu \coloneqq \Gamma_{\mathrm{IP}}(q,\nu) = \Gamma_{\mathrm{IP}}\Gamma_{\mathrm{BR}}\nu$. Similarly, let $\widetilde{q} \coloneqq \widetilde{\Gamma}_{\mathrm{BR}}\nu$ and $\widetilde{\mu} \coloneqq \Gamma_{\mathrm{IP}}(\boldsymbol{\Pi}_D\widetilde{q},\nu) = \Gamma_{\mathrm{IP}}(\boldsymbol{\Pi}_D\widetilde{\Gamma}_{\mathrm{BR}}\nu,\nu)$. To distinguish, $\mu, \widetilde{\mu} \in \mathcal{P}_{\mathrm{unif}}([0,1]\times\mathcal{X})$, $q \in \mathcal{Q}$, $\widetilde{q} \in \widetilde{\mathcal{Q}}$. Then, we have

$$\begin{aligned}
\sqrt{G_1} = \|\mu - \widetilde{\mu}\|_{\mathrm{TV}} &\le \sup_{\|\phi\|_\infty \le 1} \int_{[0,1]} \left| \int_{\mathcal{X}} \phi(u,x)(\mu_u - \widetilde{\mu}_u)(dx)\right| du \\
&\le \int_{[0,1]} \|\mu_u - \widetilde{\mu}_u\|_{\mathrm{TV}} du \\
&\le \sup_{u\in[0,1]} \|\mu_u - \widetilde{\mu}_u\|_{\mathrm{TV}}.
\end{aligned}$$

Since μ_u and $\widetilde{\mu}_u$ are the law of process $X|U=u$ with the same transition kernel, by (Zhang et al., 2024a, Lemma 4), we have for almost every u,

$$\|\mu_u - \widetilde{\mu}_u\|_{\mathrm{TV}} \le L_\pi \sigma \left\|q(u,\cdot) - \widetilde{q}(\Pi_D(u),\cdot)\right\|_2 \le L_\pi\sigma\sqrt{|\mathcal{X}||A|}\left\|q(u,\cdot) - \widetilde{q}(\Pi_D(u),\cdot)\right\|_\infty,$$

which gives

$$\sup_{u\in[0,1]} \|\mu_u - \widetilde{\mu}_u\|_{\mathrm{TV}} \le L_\pi\sigma\sqrt{|\mathcal{X}||A|}\left\|q - \boldsymbol{\Pi}_D\widetilde{q}\right\|_\infty.$$

Therefore, by Lemma H.7, we get

$$G_1 \leq \frac{|\mathcal{X}||A|L_\pi^2 L_d^2 \sigma^2((1-\gamma)L_f + \gamma\|f\|_\infty L_P)^2}{(1-\gamma)^4 D^2}. \tag{28}$$

For G_2, by Lemma H.6, we have

$$G_2 \leq \frac{L_P^2 L_D^2 \sigma^2}{D^2}. \tag{29}$$

And Lemma H.5 gives

$$G_3 = O\left(\frac{D|\mathcal{X}|^2|A|\|f\|_\infty^2 L_\pi^2 \sigma^2 \log H}{\theta^2(1-\gamma)^4 H}\right). \tag{30}$$

Plugging the above bounds on G_1, G_2, and G_3 into gives the desired result. □

Combining Proposition H.4 and the contraction assumption of FPI (Assumption 5.1(3)), we are able to show Theorem H.3 recursively.

Proof of Theorem H.3. In this proof, we omit the subscript of the total variation norm for simplicity. We denote $M_k = M^{k,0} = \{M_d^{k,0}\}_{d=1}^D$ and $\mu_k := \mathbf{\Pi}_D M_k$ for $k = 0, \ldots, K$. Note that $M_k = \widehat{\Gamma}^k M_0$. By (26) and the definition of the equilibrium population measure $\widehat{\mu}$, we have

$$\mathbb{E}\left\|\mu_K - \widehat{\mu}\right\|^2 = \mathbb{E}\left\|\mathbf{\Pi}_D \widehat{\Gamma}^K M_0 - \widehat{\mu}\right\|^2 = \mathbb{E}\left\|\mathbf{\Pi}_D \widehat{\Gamma} \mathbf{\Pi}_D^\dagger \mu_{K-1} - \Gamma_{\mathrm{IP}}\Gamma_{\mathrm{BR}}\widehat{\mu}\right\|^2, \tag{31}$$

Then, by Young's inequality, we have

$$\begin{aligned}
&\mathbb{E}\left\|\mathbf{\Pi}_D \widehat{\Gamma} \mathbf{\Pi}_D^\dagger \mu_{K-1} - \Gamma_{\mathrm{IP}}\Gamma_{\mathrm{BR}}\widehat{\mu}\right\|^2 \\
=&\mathbb{E}\left\|\left(\mathbf{\Pi}_D \widehat{\Gamma} \mathbf{\Pi}_D^\dagger - \Gamma_{\mathrm{IP}}\Gamma_{\mathrm{BR}}\right)\mu_{K-1} + \Gamma_{\mathrm{IP}}\Gamma_{\mathrm{BR}}(\mu_{K-1} - \widehat{\mu})\right\|^2 \\
\leq&(1+1/\kappa)\mathbb{E}\left\|\left(\mathbf{\Pi}_D \widehat{\Gamma} \mathbf{\Pi}_D^\dagger - \Gamma_{\mathrm{IP}}\Gamma_{\mathrm{BR}}\right)\mu_{K-1}\right\|^2 + (1+\kappa)\mathbb{E}\left\|\Gamma_{\mathrm{IP}}\Gamma_{\mathrm{BR}}(\mu_{K-1} - \widehat{\mu})\right\|^2.
\end{aligned} \tag{32}$$

Applying Proposition H.4 for the first term and the contracting FPI assumption for the second term in (32), we get

$$\begin{aligned}
\mathbb{E}\left\|\mathbf{\Pi}_D \widehat{\Gamma} \mathbf{\Pi}_D^\dagger \mu_{K-1} - \Gamma_{\mathrm{IP}}\Gamma_{\mathrm{BR}}\widehat{\mu}\right\|^2 \leq&(1+1/\kappa)\cdot O\left(\frac{D}{H} + \frac{1}{D^2}\right) + (1+\kappa)(1-\kappa)^2\mathbb{E}\left\|\mu_{K-1} - \widehat{\mu}\right\|^2 \\
\leq&\frac{1}{\kappa}\cdot O\left(\frac{D}{H} + \frac{1}{D^2}\right) + (1-\kappa)\mathbb{E}\left\|\mu_{K-1} - \widehat{\mu}\right\|^2.
\end{aligned}$$

Recursively applying the above inequality to Equation (31) gives

$$\begin{aligned}
\mathbb{E}\left\|\mu_K - \widehat{\mu}\right\|^2 \leq&(1-\kappa)^K\mathbb{E}\left\|\mu_0 - \widehat{\mu}\right\|^2 + \sum_{k=1}^K (1-\kappa)^k \frac{1}{\kappa}\cdot O\left(\frac{D}{H} + \frac{1}{D^2}\right) \\
=&O\left(\exp(-\kappa K)\mathbb{E}\left\|\mu_0 - \widehat{\mu}\right\|^2 + \frac{1}{\kappa^2}O\Big(\frac{D}{H} + \frac{1}{D^2}\Big)\right),
\end{aligned}$$

which indicates (27) by substituting the shorthand notation $O\left(\frac{D}{H} + \frac{1}{D^2}\right)$ with the explicit bounds of G_1, G_2, G_3 in (28), (29), (30) respectively. Therefore, to find an approximation equilibrium population measure μ_K such that $\mathbb{E}\|\mu_K - \widehat{\mu}\| \leq \epsilon$ for some error ϵ, we need at most

$$K = O(\kappa^{-1}\log\epsilon^{-1}), \quad D = O(\kappa^{-1}\epsilon^{-1}), \quad H = O(\kappa^{-3}\epsilon^{-3}\log\epsilon^{-1}).$$

□

H.5. Auxiliary Lemmas

The following lemmas address G_3, G_2, and G_1 in Proposition H.4 respectively.

Lemma H.5 (Online learning approximation error)**.** *Suppose Assumption 5.3 holds. With step sizes of $\alpha_\tau, \beta_\tau \asymp 1/\tau$, for any $M \in \mathcal{P}(\mathcal{X})^{\mathcal{U}}$, we have*

$$\mathbb{E}\left\|\mathbf{\Pi}_D\left(\widetilde{\Gamma}_{\mathrm{IP}}\widetilde{\Gamma}_{\mathrm{BR}}\mathbf{\Pi}_D - \widehat{\Gamma}\right)M\right\|_{\mathrm{TV}}^2 = O\left(\frac{D\|f\|_\infty^2 L_\pi^2\sigma^2|\mathcal{X}|^2|A|\log H}{\theta^2(1-\gamma)^4 H}\right).$$

Proof. We first denote $\widetilde{\mu} = \{\widetilde{\mu}^{u_d}\}_{d=1}^D \coloneqq \widetilde{\Gamma}_{\mathrm{IP}}\widetilde{\Gamma}_{\mathrm{BR}}\mathbf{\Pi}_D M$ and $\widetilde{M} = \{\widetilde{M}^{u_d}\} \coloneqq \widehat{\Gamma}M$. Then, we know that $\widetilde{\mu}^{u_d}$ is the stationary distribution of the MDP dynamic with population measure $\mathbf{\Pi}_D M$ and policy $\Gamma_\pi(\widetilde{\Gamma}_{\mathrm{BR}}\mathbf{\Pi}_D M)$, conditional on $U = u_d$ at time 0. In other words, the measure argument of reward function and transition kernel is $\mathsf{W}\mathbf{\Pi}_D M(u_d)$, and the process is controlled by policy $\Gamma_\pi(\widetilde{\Gamma}_{\mathrm{BR}}\mathbf{\Pi}_D M)(u_d,\cdot)$, which is the optimal policy.

This is the same MDP in Algorithm 2 for label u_d. Thus, by (Zhang et al., 2024a, Lemma 3), for any $u_d \in \mathcal{U}$, we have

$$\mathbb{E}\left\|\widetilde{\mu}^{u_d} - \widetilde{M}^{u_d}\right\|_2^2 = O\left(\frac{\|f\|_\infty^2 L_\pi^2\sigma^2|\mathcal{X}||A|\log H}{\theta^2(1-\gamma)^4 H}\right),$$

where $\sigma \coloneqq \widehat{n} + c_1 c_2^{\widehat{n}}/(1-c_2)$, $\widehat{n} = \left\lceil \log_{c_2} c_1^{-1}\right\rceil$, and

$$\theta \coloneqq \inf_{(u,x,a)\in[0,1]\times\mathcal{X}\times A}\inf_{q\in\mathcal{Q}} \mu_q(u,x)\Gamma_\pi(q^u)[a\,|\,x] > 0.$$

Therefore, by Lemma H.1, we have

$$\begin{aligned}
\mathbb{E}\left\|\mathbf{\Pi}_D\left(\widetilde{\mu} - \widetilde{M}\right)\right\|_{\mathrm{TV}}^2 \le \mathbb{E}\left\|\widetilde{\mu} - \widetilde{M}\right\|_{\mathrm{TV}}^2 &\le D\sup_{u_d\in\mathcal{U}}\mathbb{E}\left\|\widetilde{\mu}^{u_d} - \widetilde{M}^{u_d}\right\|_{\mathrm{TV}}^2 \\
&\le D|\mathcal{X}|\sup_{u_d\in\mathcal{U}}\mathbb{E}\left\|\widetilde{\mu}^{u_d} - \widetilde{M}^{u_d}\right\|_2^2 \\
&= O\left(\frac{D\|f\|_\infty^2 L^2\sigma^2|\mathcal{X}|^2|A|\log H}{\theta^2(1-\gamma)^4 H}\right),
\end{aligned}$$

where we recall the total variation of measure on finite space is equivalent to l_1 norm of the density vector. □

Lemma H.6 (Population discretization error)**.** *For any population distribution $\mu \in \mathcal{P}_{\mathrm{unif}}([0,1]\times\mathcal{X})$ and any D-class Q-value function $\widetilde{q} \in \widetilde{Q}$, we have*

$$\left\|\mathbf{\Pi}_D\widetilde{\Gamma}_{\mathrm{IP}}(\mu,\widetilde{q}) - \Gamma_{\mathrm{IP}}(\mu,\mathbf{\Pi}_D\widetilde{q})\right\|_{\mathrm{TV}} \le \frac{\sigma L_P L_d}{D}.$$

Proof. We first denote $\widetilde{\nu} \coloneqq \widetilde{\Gamma}_{\mathrm{IP}}(\mu,\widetilde{q})$ and $\nu \coloneqq \Gamma_{\mathrm{IP}}(\mu,\mathbf{\Pi}_D\widetilde{q})$.

Let ν admits disintegration $du\nu_u(dx)$. By construction, for a.e. $u \in I_d$, conditional on $U = u$, $\widetilde{\nu}^{u_d}$ and ν_u are the invariant measures of two Markov processes that follow the same policy $\Gamma_\pi(\widetilde{q}^{u_d})$, but w.r.t. different neighborhood measure. By (Mitrophanov, 2005, Corollary 3.1), for the same σ in Lemma H.5, we have for a.e. $u \in I_d$,

$$\begin{aligned}
\left\|\widetilde{\nu}^{u_d} - \nu_u\right\|_{\mathrm{TV}} &\le \sigma\sup_{x,a}\|P(x,\mathsf{W}\mu(u_d),a) - P(x,\mathsf{W}\mu(u),a)\|_{\mathrm{TV}} \\
&\le \sigma L_P\|\mathsf{W}\mu(u_d) - \mathsf{W}\mu(u)\|_{\mathrm{TV}} \le \frac{\sigma L_P L_d}{D},
\end{aligned}$$

Thus,

$$\begin{aligned}
\left\|\mathbf{\Pi}_D\widetilde{\nu} - \nu\right\|_{\mathrm{TV}} &= \sup_{\|\phi\|_\infty\le 1}\left|\int_{[0,1]\times\mathcal{X}}\phi(u,x)(\mathbf{\Pi}_D\widetilde{\nu} - \nu)(du,dx)\right| \\
&\le \sum_{d=1}^D\sup_{\|\phi\|_\infty\le 1}\int_{I_{u_d}}\left|\int_{\mathcal{X}}\phi(u,x)\big(\widetilde{\nu}^{u_d}(dx) - \nu_u(dx)\big)\right|du \\
&\le \sum_{d=1}^D\int_{I_{u_d}\times\mathcal{X}}\left\|\widetilde{\nu}^{u_d} - \nu_u\right\|_{\mathrm{TV}}du \le \frac{\sigma L_P L_d}{D}.
\end{aligned}$$

□

Lemma H.7 (Population discretization error)**.** *For any population distribution $\mu \in \mathcal{P}_{\text{unif}}([0,1]\times\mathcal{X})$, let $q_* \coloneqq \Gamma_{\text{BR}}\mu$ and $\widetilde{q}_* \coloneqq \widetilde{\Gamma}_{\text{BR}}\mu$. We have*

$$\sup_{\mu} \|q_* - \mathbf{\Pi}_D \widetilde{q}_*\|_\infty \le \frac{L_d((1-\gamma)L_f + \gamma\|f\|_\infty L_P)}{(1-\gamma)^2 D}.$$

Proof. We defined a generalized state value function associated with a policy $\rho : [0,1]\times\mathcal{X} \to \mathcal{P}(A)$ by

$$v^{\rho_{u_0}}(u_1,u_2,x) = \mathbb{E}\bigg[\sum_{\tau\ge 0}\gamma^\tau f(X_\tau^{\rho_{u_0}}, \mathsf{W}\mu(u_2), \alpha_\tau^{\rho_{u_0}})\Big| X_0^{\rho_{u_0}} = x, U = u_1\bigg].$$

With a slight abuse of notation, we denote

$$q^{\rho_{u_0}}(u_1,u_2,x,a) := f(x,\mathsf{W}\mu(u_2),a) + \gamma\langle P(x,\mathsf{W}\mu(u_1),a), v^{\rho_{u_0}}(u_1,u_2,\cdot)\rangle,$$

where ρ_{u_0} is to fix u_0 as the first argument of ρ, i.e., $\rho_{u_0}(x) = \rho(u_0,x)$. In words, $v^{\rho_{u_0}}(u_1,u_2,x)$ and $q^{\rho_{u_0}}(u_1,u_2,x,a)$ are generalization of typical value function and Q functions, where the policy follows label u_0, state transition follows u_1, and the reward follows u_2.

Note that $q_* \in \mathcal{Q}$, and $\widetilde{q}_* \in \widetilde{\mathcal{Q}}$. Let $\pi = \Gamma_\pi(q_*)$. By definitions of Γ_{BR} and $\widetilde{\Gamma}_{\text{BR}}$, we know that

$$\begin{aligned} q_*(u,x,a) = q^{\pi_u}(u,u,x,a) &\iff v^{\pi_u}(u,u,x) = \sup_{a\in A} q_*(u,x,a), \\ \widetilde{q}_*(u_d,x,a) = q^{\pi_{u_d}}(u_d,u_d,x,a) &\iff v^{\pi_{u_d}}(u_d,u_d,x) = \sup_{a\in A}\widetilde{q}_*(u_d,x,a), \qquad 1\le d\le D. \end{aligned}$$

Note that q_* and $\widetilde{q}_*$ coincide on the space $\mathcal{U}\times\mathcal{X}\times A$ by definitions. On $([0,1]\backslash\mathcal{U})\times\mathcal{X}\times A$, the Q-function q_* is strictly larger than $\mathbf{\Pi}_D\widetilde{q}_*$ by its optimality. With this in mind, by the definition of the L_∞ norm, we have

$$\begin{aligned} \|q_* - \mathbf{\Pi}_D\widetilde{q}_*\|_\infty &= \sup_{x,a}\sup_{1\le d\le D}\sup_{u\in I_{u_d}}\Big(q_*(u,x,a) - \mathbf{\Pi}_D\widetilde{q}_*(u,x,a)\Big) \\ &= \sup_{x,a}\sup_{1\le d\le D}\sup_{u\in I_{u_d}}\Big(q^{\pi_u}(u,u,x,a) - q^{\pi_{u_d}}(u_d,u_d,x,a)\Big), \end{aligned}$$

where

$$\begin{aligned} q^{\pi_u}(u,u,x,a) - q^{\pi_{u_d}}(u_d,u_d,x,a) \le{}& \underbrace{\Big|q^{\pi_u}(u,u,x,a) - q^{\pi_u}(u,u_d,x,a)\Big|}_{\text{I}} \\ &+ \underbrace{\Big|q^{\pi_u}(u,u_d,x,a) - q^{\pi_u}(u_d,u_d,x,a)\Big|}_{\text{II}} \\ &+ \underbrace{\Big(q^{\pi_u}(u_d,u_d,x,a) - q^{\pi_{u_d}}(u_d,u_d,x,a)\Big)}_{\text{III}}. \end{aligned}$$

Term I. we use the Lipschitzness of the reward function, and obtain

$$\begin{aligned} \text{I} &\le \Big|f(x,\mathsf{W}\mu(u),a) - f(x,\mathsf{W}\mu(u_d),a)\Big| + \gamma\Big|\langle P(x,\mathsf{W}\mu(u),a), v^{\pi_u}(u,u,\cdot) - v^{\pi_u}(u,u_d,\cdot)\rangle\Big| \\ &\le L_f\|\mathsf{W}\mu(u) - \mathsf{W}\mu(u_d)\|_{\text{TV}} \\ &+ \gamma\bigg\langle P(x,\mathsf{W}\mu(u),a)\,,\, \mathbb{E}\bigg[\sum_{\tau\ge 0}\gamma^\tau\Big|f(X_\tau^{\pi_u},\mathsf{W}\mu(u),\alpha_\tau^{\pi_u}) - f(X_\tau^{\pi_u},\mathsf{W}\mu(u_d),\alpha_\tau^{\pi_u})\Big|\Big|X_0^{\pi_u} = \cdot, U = u\bigg]\bigg\rangle \\ &\le L_f\|\mathsf{W}\mu(u) - \mathsf{W}\mu(u_d)\|_{\text{TV}} + \gamma\bigg\langle P(x,\mathsf{W}\mu(u),a)\,,\, \mathbb{E}\bigg[\sum_{\tau\ge 0}\gamma^\tau L_f\|\mathsf{W}\mu(u) - \mathsf{W}\mu(u_d)\|_{\text{TV}}\Big|X_0^{\pi_u} = \cdot, U = u\bigg]\bigg\rangle \\ &\le \frac{L_f}{1-\gamma}\|\mathsf{W}\mu(u) - \mathsf{W}\mu(u_d)\|_{\text{TV}} \\ &\le \frac{L_f L_d}{(1-\gamma)D}. \end{aligned}$$

Term II. we first define iteratively the measure of state-action pair at time $t \geq 1$ under any policy $\rho_u : \mathcal{X} \to \mathcal{P}(A)$ as

$$\begin{aligned}\underline{P}_t^{\rho_u}(x_0, m, a_0) &:= \mathcal{L}(X_t^{\rho_u}, \alpha_t^{\rho_u} | X_0^{\rho_u} = x_0, \alpha_0^{\rho_u} = a_0) \\ &= \int_{\mathcal{X}^2 \times A^2} \Big[\delta_{x_t} \delta_{a_t} \rho_{u,x_t}(da_t) P(x_{t-1}, m, a_{t-1})(dx_t)\Big] \underline{P}_{t-1}^{\rho_u}(x_0, m, a_0)(dx_{t-1}, da_{t-1}) \\ &\in \mathcal{P}(\mathcal{X} \times A).\end{aligned}$$

We claim that for any $\rho_u : \mathcal{X} \to \mathcal{P}(A)$, any $(x_0, a_0) \in \mathcal{X} \times A$, any $m_1, m_2 \in \mathcal{P}(\mathcal{X})$ and any time $t \geq 1$,

$$\|\underline{P}_t^{\rho_u}(x_0, m_1, a_0) - \underline{P}_t^{\rho_u}(x_0, m_2, a_0)\|_{\mathrm{TV}} \leq t L_P \|m_1 - m_2\|_{\mathrm{TV}}. \tag{33}$$

It is trivial that

$$\underline{P}_1^{\rho_u}(x_0, m, a_0) = P(x_0, m, a_0)$$

is uniformly Lipschitz in measure argument under assumption Assumption 5.1. Assuming (33) holds for $t-1$, we now show it holds for t with the add-and-subtract trick again.

$$\begin{aligned}&\|\underline{P}_t^{\rho_u}(x_0, m_1, a_0) - \underline{P}_t^{\rho_u}(x_0, m_2, a_0)\|_{\mathrm{TV}} \\ &\leq \sup_{\|\phi\|_\infty \leq 1} \int_{A \times \mathcal{X}^2} \phi(a_t, x_t) \rho_{u,x_t}(da_t) \Big[P_{x_{t-1}, m_1, a_{t-1}}(dx_t) \underline{P}_{t-1}^{\rho_u}(x_0, m_1, a_0)(dx_{t-1}, da_{t-1}) \\ &- P_{x_{t-1}, m_2, a_{t-1}}(dx_t) \underline{P}_{t-1}^{\rho_u}(x_0, m_2, a_0)(dx_{t-1}, da_{t-1})\Big] \\ &\leq \sup_{\|\phi\|_\infty \leq 1} \int_{A \times \mathcal{X}^2} \phi(a_t, x_t) \rho_{u,x_t}(da_t) \Big[P_{x_{t-1}, m_1, a_{t-1}} - P_{x_{t-1}, m_2, a_{t-1}}\Big](dx_t) \underline{P}_{t-1}^{\rho_u}(x_0, m_1, a_0)(dx_{t-1}, da_{t-1}) \\ &+ \sup_{\|\phi\|_\infty \leq 1} \int_{A \times \mathcal{X}^2} \phi(a_t, x_t) \rho_{u,x_t}(da_t) P_{x_{t-1}, m_2, a_{t-1}}(dx_t) \Big[\underline{P}_{t-1}^{\rho_u}(x_0, m_1, a_0) - \underline{P}_{t-1}^{\rho_u}(x_0, m_2, a_0)\Big](dx_{t-1}, da_{t-1}) \\ &\leq (t-1) L_P \|m_1 - m_2\|_{\mathrm{TV}} + L_P \|m_1 - m_2\|_{\mathrm{TV}} \\ &= t L_P \|m_1 - m_2\|_{\mathrm{TV}}.\end{aligned}$$

With this claim, we have

$$\begin{aligned}\mathrm{II} &\leq \Big|q^{\pi_u}(u, u_d, x, a) - q^{\pi_u}(u_d, u_d, x, a)\Big| \\ &\leq \sum_{t \geq 0} \gamma^t \Big|\Big\langle \underline{P}_t^{\pi_u}(x, \mathsf{W}\mu(u), a) - \underline{P}_t^{\pi_u}(x, \mathsf{W}\mu(u_d), a)\,,\, f(\cdot, \mathsf{W}\mu(u_d), \cdot)\Big\rangle\Big| \\ &\leq \sum_{t \geq 0} \gamma^t \|f\|_\infty \|\underline{P}_t^{\pi_u}(x, \mathsf{W}\mu(u), a) - \underline{P}_t^{\pi_u}(x, \mathsf{W}\mu(u_d), a)\|_{\mathrm{TV}} \\ &\leq L_P \|f\|_\infty \|\mathsf{W}\mu(u) - \mathsf{W}\mu(u_d)\|_{\mathrm{TV}} \sum_{t \geq 0} t\gamma^t \\ &\leq \frac{\gamma \|f\|_\infty L_P L_d}{(1-\gamma)^2 D}.\end{aligned}$$

Term III. It is immediate that

$$\mathrm{III} = q^{\pi_u}(u_d, u_d, x, a) - q^{\pi_{u_d}}(u_d, u_d, x, a) \leq 0,$$

as π_{u_d} is the optimizer of $v^{\pi_{u_d}}(u_d, u_d, \cdot)$.

Finally, we conclude

$$\|q_* - \mathbf{\Pi}_D \widetilde{q}_*\|_\infty \leq \frac{L_d((1-\gamma)L_f + \gamma \|f\|_\infty L_P)}{(1-\gamma)^2 D}.$$

□

I. Experiment Setup

I.1. Experiment 1: Flocking-Graphon

The Flocking-Graphon game studies the flocking behavior, i.e., the phenomenon that agents gather together at some location as time goes by, in a large populations (of animals). Its modeling finds applications in psychology, animation, social science, or swarm robotics (Perrin et al., 2021). Each player in the game makes decisions regarding velocity control to avoid its own deviation from the centroid of the population, and the desirable outcome (i.e., equilibrium reached by the population) reveals how a consensus can be reached in a group without centralized decision-making.

We consider a flocking game (Lacker & Soret, 2023) on one-dimensional space $\mathcal{X} = \mathbb{R}$, and each agent is allowed to control its velocity in the compact action space $A \subset \mathbb{R}$. The transition dynamic is defined to be a continuous time state process, given by:

$$dx_t = \alpha_t dt + \sigma dB_t,$$

,where $x_t \in \mathcal{X}$. α_t is the velocity control at time t, and we usually consider it to be a closed loop control, i.e., $\alpha_t = \alpha_t(x)$ for function α, which represents the velocity at position x at time t. B_t is a one-dimensional Brownian motion. The player aims to optimize the following objective

$$J_W(\mu, \alpha) := -\mathbb{E}\Big[\int_0^T \alpha_t^2 dt + c\big|x_T - G^\mu(U)\big|^2\Big],$$

where $c > 0$ is a constant, and

$$G^\mu(u) := \langle \mathsf{W}\mu_T(u), \mathrm{Id}\rangle = \int_{[0,1]\times\mathbb{R}} W(u,v)x\mu_T(dv, dx),$$

with Id being the identity mapping. $G^\mu(u)$ is interpreted as the centroid of the population over the space domain $\mathcal{X}$. More specifically, $G^\mu(u)$ is the average of the state distribution of the population μ, weighted from the perspective of player with label u. Intuitively, the running cost arises from change in the velocity, and the terminal cost is associated with deviation from the centroid at terminal time.

I.2. Experiment 2: SIS-Graphon

(Cui & Koeppl, 2022) considers a game that models pandemic evolution. It admits state space $\mathcal{X} = \{x_S, x_I\}$ where x_S represents a safe state, and x_I represents an infection state. The action space is taken to be $A = \{a_U, a_D\}$, where a_U represents keeping interaction with others and a_D represents taking a quarantine. The terminal time is set to $T = 50$. The transition probability is

$$\begin{aligned}
\mathbb{P}(x_S|x_I, m, a) &= \frac{1}{2} \qquad \forall (m, a) \in \mathcal{M}_+(\mathcal{X}) \times A \\
\mathbb{P}(x_I|x_S, m, a_U) &= \frac{4}{5} m(x_I) \qquad \forall m \in \mathcal{M}_+(\mathcal{X}) \\
\mathbb{P}(x_I|x_S, m, a_D) &= 0 \qquad \forall m \in \mathcal{M}_+(\mathcal{X}).
\end{aligned}$$

An infected agent may turn safe with half probability each time step, regardless of the action. The probability a safe agent is infected is proportion to the infected individuals in her neighborhood when she keeps interaction with others, and is 0 when she takes a quarantine. The reward function is given by

$$f(x, m, a) = -2 \cdot \mathbf{1}_{x_I}(x) - 0.5 \cdot \mathbf{1}_{a_D}(a).$$

An agent takes cost from both being infected and taking quarantine action.

I.3. Experiment 3: Investment-Graphon

In the Investment-Graphon game (Cui & Koeppl, 2022), the terminal time is set to $T = 50$. Each agent is viewed as a firm, and let $\mathcal{X} = \{0, 1, \dots, 9\}$ be the quality of products this firm provides. With action space given by $A = \{a_I, a_O\}$, the transition kernel is defined by

$$\begin{aligned}
\mathbb{P}(x+1|x, m, a_I) &= \frac{9-x}{10} \qquad \forall m \in \mathcal{M}_+(\mathcal{X}) \\
\mathbb{P}(x|x, m, a_I) &= \frac{1+x}{10} \qquad \forall m \in \mathcal{M}_+(\mathcal{X}) \\
\mathbb{P}(x|x, m, a_O) &= 1 \qquad \forall m \in \mathcal{M}_+(\mathcal{X}).
\end{aligned}$$

We Interpret a_I as investment, and a_O as not investing. A firm may improve the product quality by investing, and the probability of a successful investment decrease as the current quality is already high. Initially, every firm starts from quality 0. The reward function is given by

$$f(x, m, a) = \frac{0.3x}{1 + \sum_{x' \in \mathcal{X}} x' m(x')} - 2 \cdot \mathbf{1}_{a_I}(a).$$

A firm's profit is proportion to the quality of product, and decrease with the average product quality within its neighborhood.

J. Experiment Results

In this section, we present detailed numerical results for three graphon games utilized in the main body. The experiment results include algorithm performance (convergence gap, W1-distance, exploitability) and GMFE.

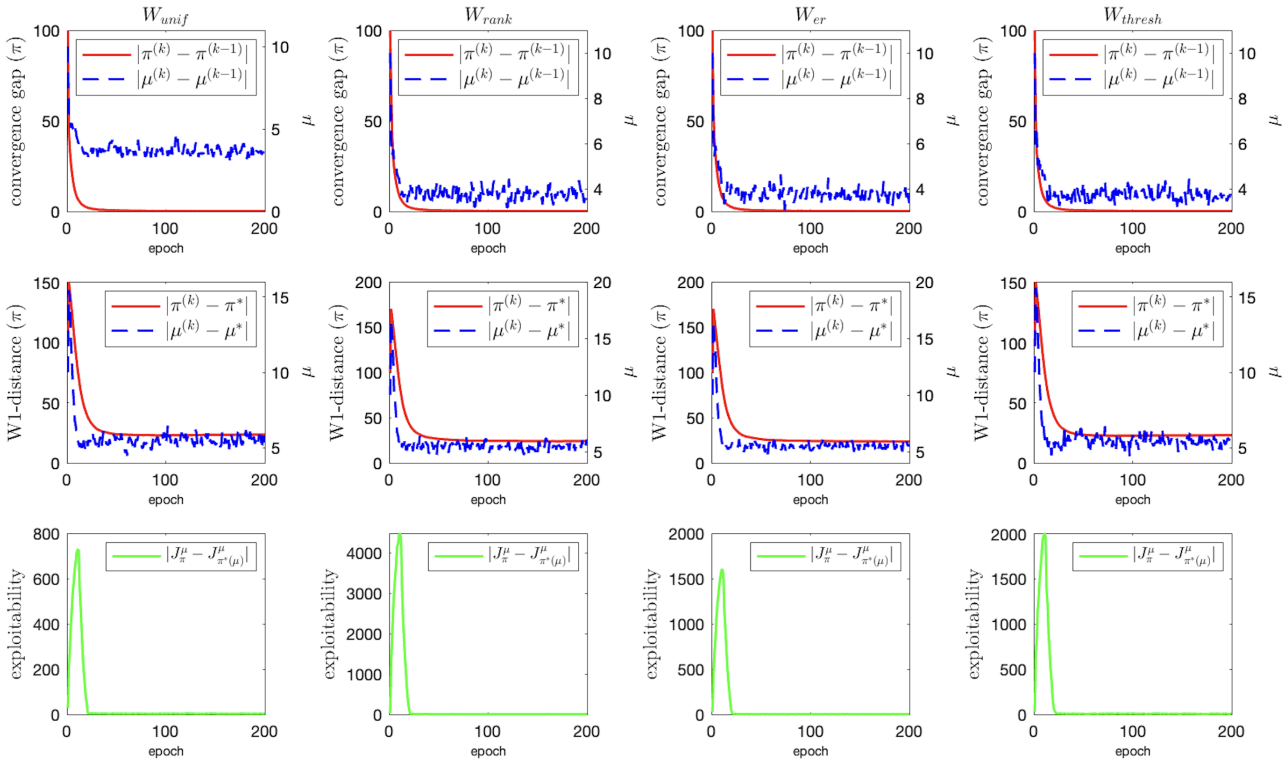

Figure 3. **Flocking-Graphon**: Algorithm performance. We demonstrate the convergence gap (top), W1-distance (middle) and exploitability (bottom) corresponding to four types of graphs. The exploitability indicates how an agent can improve be deviating from the policy used by the rest of the population. Mathematically, the exploitability is calculated as $|J^{\mu}_{\pi} - J^{\mu}_{\pi^*(\mu)}|$. It measures the gap between the policy adopted by the population and the best policy that an agent can achieve in response to the population state.

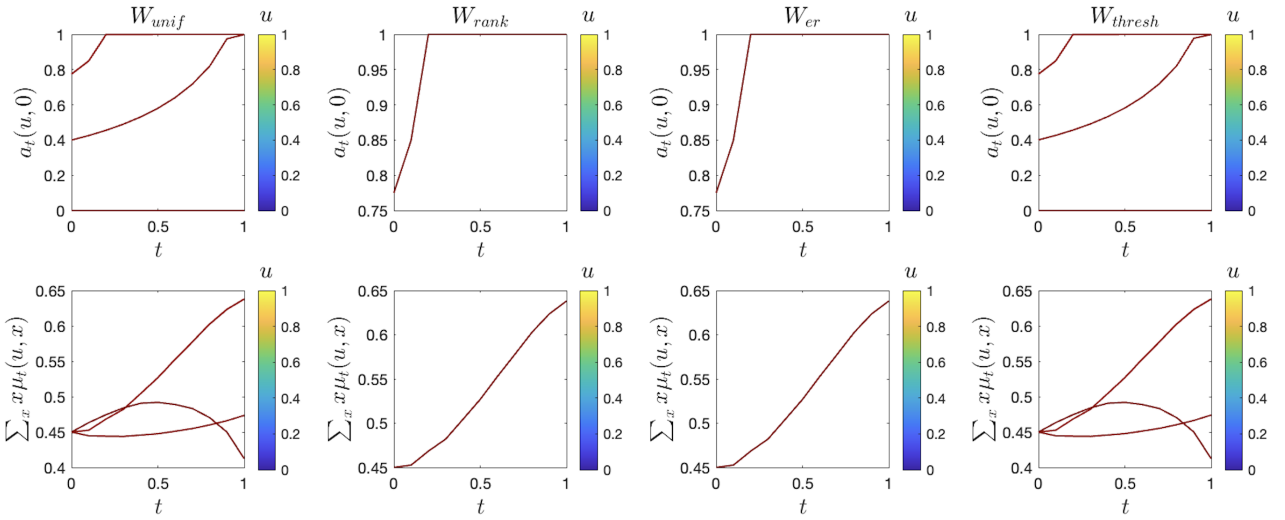

Figure 4. **Flocking-Graphon**: GMFE. Top: The velocity control at position $x = 0$. The x-axis denotes the time horizon and the y-axis denotes the velocity at equilibrium. The color bar denotes the label state. Bottom: The expected position x across the time. It can be regarded as the centroid of the population.

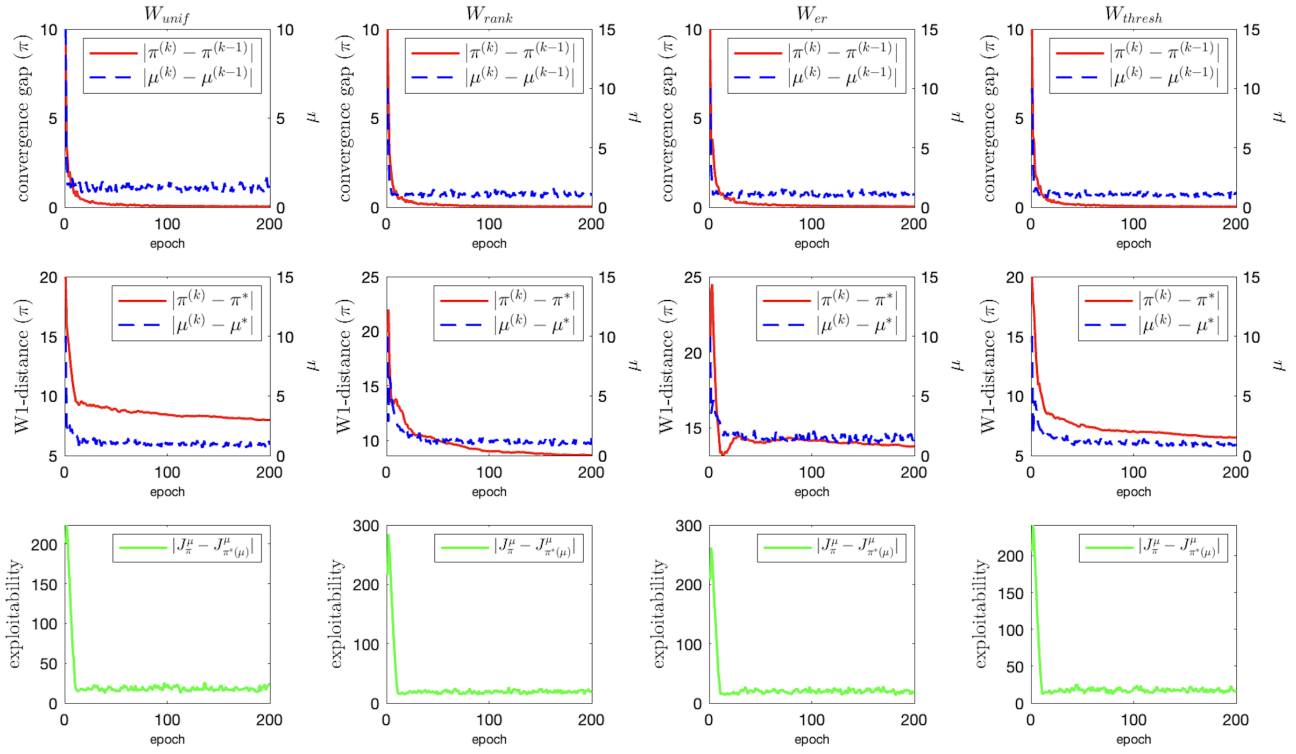

Figure 5. **SIS-Graphon**: Algorithm performance. We demonstrate the convergence gap (top), W1-distance (middle) and exploitability (bottom) corresponding to four types of graphs.

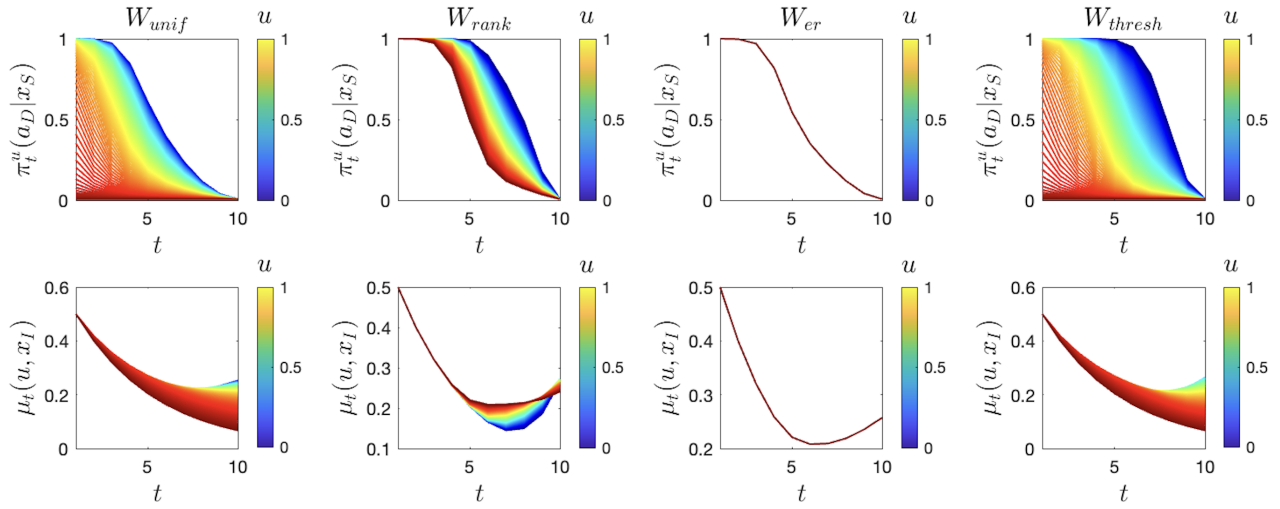

Figure 6. **SIS-Graphon**: GMFE. Top: The probability of taking precautions when healthy. The results for graphs W_{unif}, W_{rank} and W_{er} is consistent with (Cui & Koeppl, 2022). We add the results for graph W_{thresh}. It is shown that the GMFE with W_{thresh} is similar to W_{unif}. Bottom: The population being infected. Agents with a higher u have fewer connections with others. It means they are less likely infected by the population in a comparison to others. Thus, they take fewer precautions.

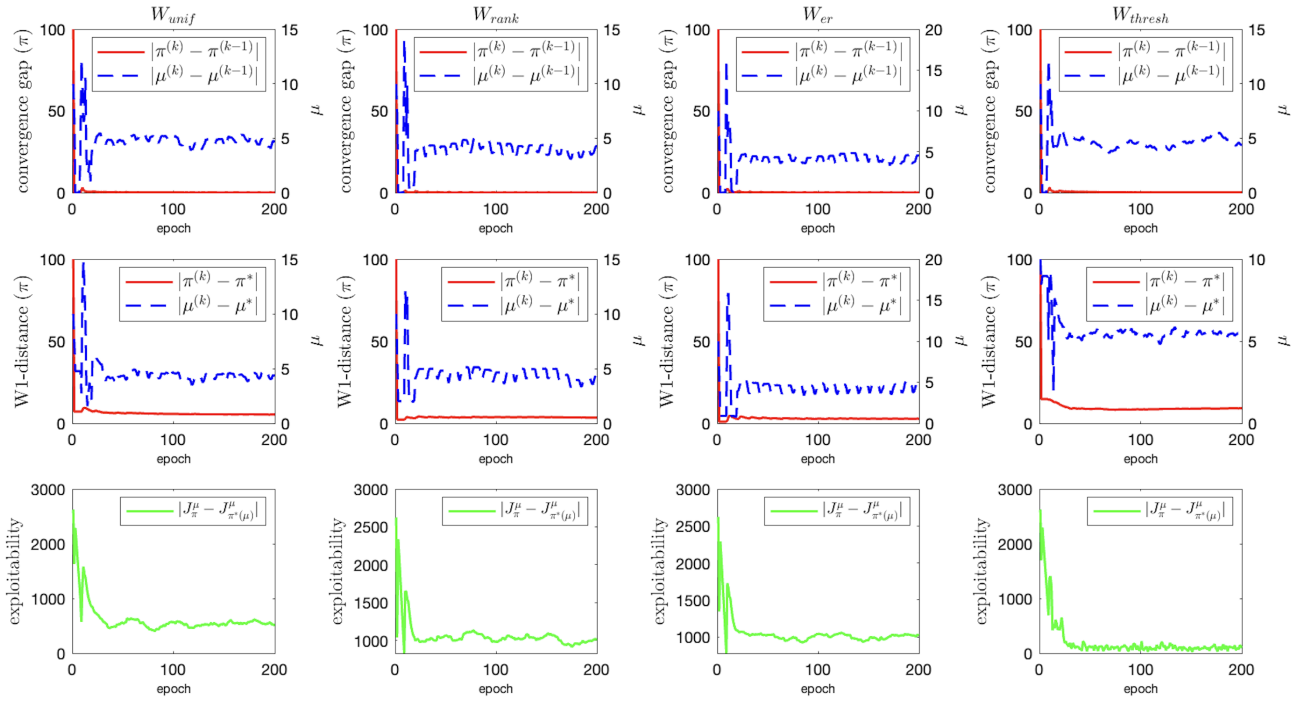

Figure 7. **Invest-Graphon**: Algorithm performance. We demonstrate the convergence gap (top), W1-distance (middle) and exploitability (bottom) corresponding to four types of graphs.

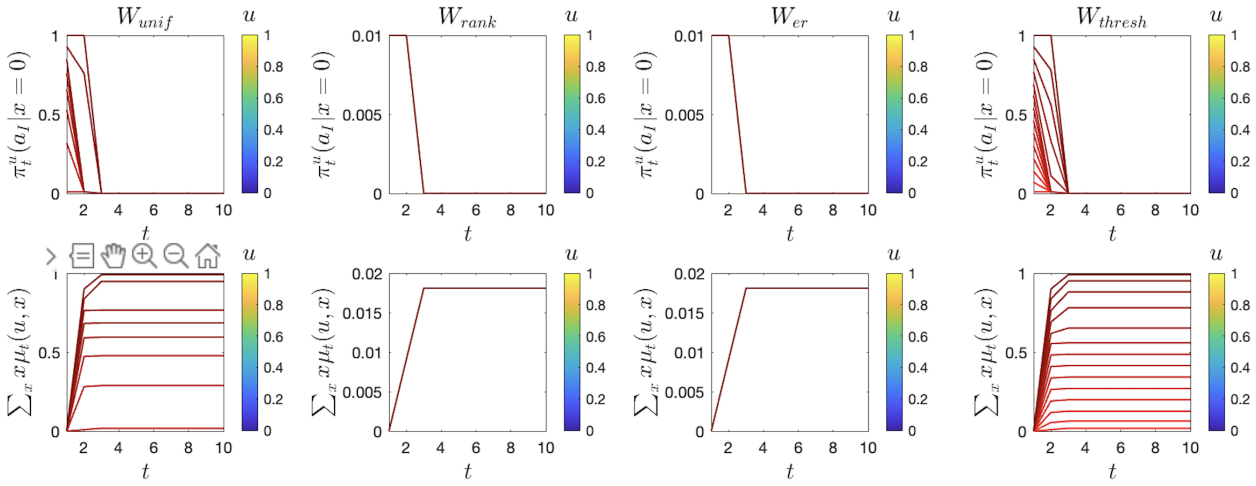

Figure 8. **Invest-Graphon**: GMFE. Top: the probability of investing on product quality when $x = 0$. Bottom: The expected product quality across the time.