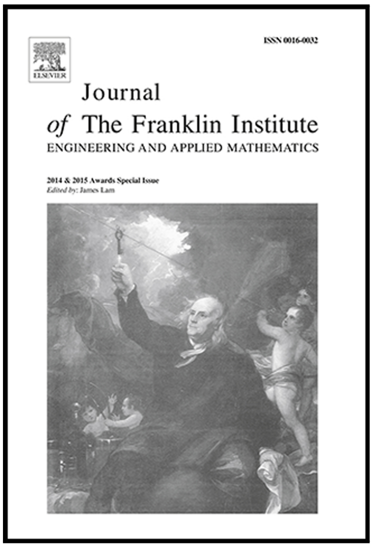
Journal Pre-proof



Graphon Mean-Field Control for Cooperative Multi-Agent Reinforcement Learning

Yuanquan Hu, Xiaoli Wei, Junji Yan, Hengxi Zhang

PII: S0016-0032(23)00548-3
DOI: https://doi.org/10.1016/j.jfranklin.2023.09.002
Reference: FI 6391

To appear in: *Journal of the Franklin Institute*

Received date: 8 September 2022
Revised date: 8 May 2023
Accepted date: 1 September 2023

Please cite this article as: Yuanquan Hu, Xiaoli Wei, Junji Yan, Hengxi Zhang, Graphon Mean-Field Control for Cooperative Multi-Agent Reinforcement Learning, *Journal of the Franklin Institute* (2023), doi: https://doi.org/10.1016/j.jfranklin.2023.09.002

Graphon Mean-Field Control for Cooperative Multi-Agent Reinforcement Learning

Yuanquan Hu, Xiaoli Wei*, Junji Yan, Hengxi Zhang

Abstract

The marriage between mean-field theory and reinforcement learning has shown a great capacity to solve large-scale control problems with homogeneous agents. To break the homogeneity restriction of mean-field theory, a recent interest is to introduce graphon theory to the mean-field paradigm. In this paper, we propose a graphon mean-field control (GMFC) framework to approximate cooperative heterogeneous multi-agent reinforcement learning (MARL) with nonuniform interactions and heterogeneous reward functions and state transition functions among agents and show that the approximate order is of $\mathcal{O}(\frac{1}{\sqrt{N}})$, with N the number of agents. By discretizing the graphon index of GMFC, we further introduce a smaller class of GMFC called block GMFC, which is shown to well approximate cooperative MARL in terms of the value function and the policy. Finally, we design a Proximal Policy Optimization based algorithm for block GMFC that converges to the optimal policy of cooperative MARL. Our empirical studies on several examples demonstrate that our GMFC approach is comparable with the state-of-art MARL algorithms while enjoying better scalability.

Keywords:
Cooperative Multi-Agent Reinforcement Learning, Graphon Theory, Graphon Mean-Field Control, Proximal Policy Optimization
2000 MSC: 60J20, 91A13

1. Introduction

Multi-agent reinforcement learning (MARL) has found various applications in the field of transportation and simulation [50, 1], stock price analysis and trading [32, 31], wireless communication networks [12, 11, 13], and learning behaviors in social dilemmas [33, 28, 34]. MARL, however, becomes intractable due to the complex interactions among agents as the number of agents increases.

A recent tractable approach is a mean-field approach by considering MARL in the regime with a large number of homogeneous agents under weak interactions [20]. According to the

*Corresponding author.
Email addresses: huyq21@mails.tsinghua.edu.cn (Yuanquan Hu), xiaoli.wei@hit.edu.cn (Xiaoli Wei), yan-jj21@mails.tsinghua.edu.cn (Junji Yan), zhanghx20@mails.tsinghua.edu.cn (Hengxi Zhang)

number of agents and learning goals, there are three subtle types of mean-field theories for MARL. The first one is called mean-field MARL (MF-MARL), which refers to the empirical average of the states or actions of a *finite* population. For example, [52] proposes to approximate interactions within the population of agents by averaging the actions of the overall population or neighboring agents. [35] proposes a mean-field proximal policy optimization algorithm for a class of MARL with permutation invariance. The second one is called mean-field game (MFG), which describes the asymptotic limit of non-cooperative stochastic games as the number of agents goes to infinity [30, 27, 8]. Recently, a rapidly growing literature studies MFG for noncooperative MARL either in a model-based way [53, 6, 26] or by a model-free approach [25, 48, 18, 14, 44]. The third one is called mean-field control (MFC), which is closely related to MFG yet different from MFG in terms of learning goals. For cooperative MFC, the Bellman equation for the value function is defined on an enlarged space of probability measures, and MFC is always reformulated as a new Markov decision process (MDP) with continuous state-action space. [9] shows the existence of optimal policies for MFC in the form of mean-field MDP and adapts classical reinforcement learning (RL) methods to the mean-field setups. [23] approximates MARL by a MFC approach, and proposes a model-free kernel-based Q-learning algorithm (MFC-K-Q) that enjoys a linear convergence rate and is independent of the number of agents. [44] presents a model-based RL algorithm M3-UCRL for MFC with a general regret bound. [2] proposes a unified two-timescale learning framework for MFG and MFC by tuning the ratio of learning rates of Q function and the population state distribution. Under the framework of MFC, [41] proposes locally executable policies such that the resulting discounted sum of average rewards well approximates the optimal value function over all policies with theoretical guarantee.

One restriction of the mean-field theory is that it eliminates the difference among agents and interactions between agents are assumed to be uniform. However, in many real world scenarios, strategic interactions between agents are not always uniform and rely on the relative positions of agents. To develop scalable learning algorithms for multi-agent systems with heterogeneous agents, one approach is to exploit the local network structure of agents [45, 37]. Another approach is to consider mean-field systems on large graphs and their asymptotic limits, which leads to graphon mean-field theory [39]. So far, most existing works on graphon mean-field theory consider either diffusion processes without learning in continuous time or non-cooperative graphon mean-field game (GMFG) in discrete time. [3] considers uncontrolled graphon mean-field systems in continuous time. [17] studies MFG on an Erdös-Rényi graph. [19] studies the convergence of weighted empirical measures described by stochastic differential equations. [4] studies propagation of chaos of weakly interacting particles on general graph sequences. [5] considers general GMFG and studies ε-Nash equilibria of the multi-agent system by a PDE approach in continuous time. [29] studies stochastic games on large graphs and their graphon limits. It shows that GMFG is viewed as a special case of MFG by viewing the label of agents as a component of the state process. [21, 22] study continuous-time cooperative graphon mean-field systems with linear dynamics. On the other hand, [7] studies static finite-agent network games and their associated graphon games. [49] provides a sequential decomposition algorithm to find Nash equilibria of discrete-time GMFG. [15] constructs a discrete-time learning GMFG framework to analyze approximate Nash equilibria for MARL with nonuniform interactions. However, little is focused on learning cooperative graphon mean-field systems in discrete time, except for [42, 43] on particular forms of nonuniform interactions among agents. [43] proves that when the reward is affine in the state distribution and action distribution, MARL with nonuniform interactions can still be approximated by classic MFC. [42] considers multi-class MARL, where agents belonging to the same class are homogeneous. In contrast, we consider a general discrete-time GMFC framework under which agents are allowed to be fully heterogeneous and interact non-uniformly on any network captured by a graphon.

Our Work. In this work, we propose a general discrete-time GMFC framework to approximate cooperative heterogeneous MARL on large graphs by combining classical MFC and network games. Theoretically, we first show that GMFC can be reformulated as a new MDP with deterministic dynamics and infinite-dimensional state-action space, hence the Bellman equation for Q function is established on the space of probability measure ensembles. It shows that GMFC approximates cooperative MARL well in terms of both value function and optimal policies. The approximation error is at order $\mathcal{O}(1/\sqrt{N})$, where N is the number of agents. Furthermore, instead of learning infinite-dimensional GMFC directly, we introduce a smaller class called block GMFC by discretizing the graphon index, which can be recast as a new MDP with deterministic dynamic and finite-dimensional continuous state-action space. We show that the optimal policy ensemble learned from block GMFC is near optimal for cooperative MARL. Using the approach in [38], we develop a proximal policy optimization (PPO) based algorithm for block GMFC, which, together with approximation result between block GMFC and cooperative MARL, shows that the proposed PPO algorithm converges to the optimal policy of MARL with the sample complexity guarantee. Empirically, our experiments in Section 5 demonstrate that when the number of agents becomes large, the mean episode reward of MARL becomes increasingly close to that of block GMFC, which verifies our theoretical findings. Furthermore, our block GMFC approach achieves comparable performances with other popular existing MARL algorithms in the finite-agent setting.

Outline. The rest of the paper is organized as follows. Section 2 recalls basic notations of graphons and introduces the setup of cooperative MARL with nonuniform interactions and its asymptotic limit called GMFC. Section 3 connects cooperative MARL and GMFC, introduces block GMFC for efficient algorithm design, and builds its connection with cooperative MARL. The main theoretical proofs are presented in Section 4. Section 5 tests the performance of block GMFC experimentally.

2. Mean-Field MARL on Dense Graphs

2.1. Preliminary: Graphon Theory

In the following, we consider a cooperative multi-agent system and its associated mean-field limit. In this system, each agent is affected by all others, with different agents exerting different effects on her. This multi-agent system with N agents can be described by a weighted graph $G_N = (\mathcal{V}_N, \mathcal{E}_N)$, where the vertex set $\mathcal{V}_N = \{1, \ldots, N\}$ and the edge set $\mathcal{E}_N$

represent agents and the interactions between agents, respectively. The adjacency matrix of G_N is represented as $\{\xi_{i,j}^N\}_{1\le i,j\le N}$. To study the limit of the multi-agent system as N goes to infinity, we adopt the graphon theory introduced in [39] used to characterize the limit behavior of dense graph sequences. Therefore, throughout the paper, we assume the graph G_N is dense and leave sparse graphs for future study.

In general, a graphon is represented by a bounded symmetric measurable function $W : \mathcal{I}\times\mathcal{I}\to\mathcal{I}$, with $\mathcal{I}=[0,1]$. We denote by $\mathcal{W}$ the space of all graphons and equip the space $\mathcal{W}$ with the cut norm $\|\cdot\|_\square$

$$\|W\|_\square = \sup_{S,T\subset\mathcal{I}}\left|\int_{S\times T} W(\alpha,\beta)d\alpha d\beta\right|.$$

For each weighted graph $G_N=(\mathcal{V}_N,\mathcal{E}_N)$, we consider the correspondence between the adjacency matrix $\{\xi_{i,j}^N\}$ and a function on $\mathcal{I}\times\mathcal{I}$ with constant value $\xi_{i,j}^N$ on each block $(\frac{i-1}{N},\frac{i}{N}]\times(\frac{j-1}{N},\frac{j}{N}]$. We make the following condition on the strength of interaction $\xi_{i,j}^N$ between agents i and j and the associated W_N.

Condition on W_N and $\xi_{i,j}^N$

1) W_N is a step graphon, that is, $0\le W_N\le 1$ and W_N is a constant on each block $(\frac{i-1}{N},\frac{i}{N}]\times(\frac{j-1}{N},\frac{j}{N}]$:

$$W_N(\alpha,\beta) = W_N\Big(\frac{i}{N},\frac{j}{N}\Big),\ \text{if}\ \alpha\in(\frac{i-1}{N},\frac{i}{N}],\ \beta\in(\frac{j-1}{N},\frac{j}{N}]. \tag{2.1}$$

2) $\xi_{i,j}^N$ is taken as either

$$\xi_{i,j}^N = W_N(\frac{i}{N},\frac{j}{N}) \tag{C1}$$

or

$$\xi_{i,j}^N \sim \text{Bernoulli}\big(W_N(\frac{i}{N},\frac{j}{N})\big). \tag{C2}$$

We further assume that the sequence of W_N converges to a graphon W in cut norm as the number of agents N goes to infinity, which is crucial for the convergence analysis of cooperative MARL in Section 3.

Assumption 2.1 *The sequence $(W_N)_{N\in\mathbb{N}}$ converges in cut norm to some graphon $W\in\mathcal{W}$ such that*

$$\|W_N - W\|_\square \to 0.$$

Some common examples of graphons include

1) Erdős Rényi: $W(\alpha,\beta)=p$, $0\le p\le 1$, $\alpha,\beta\in\mathcal{I}$;

2) Stochastic block model:

$$W(\alpha,\beta)=\begin{cases} p & \text{if } 0\leqslant \alpha,\beta\leqslant 0.5 \text{ or } 0.5\leqslant\alpha,\beta\leqslant 1,\\ q & \text{otherwise,}\end{cases}$$

where p represents the intra-community interaction and q the inter-community interaction;

3) Random geometric graphon: $W(\alpha,\beta)=f(\min(|\beta-\alpha|,1-|\beta-\alpha|))$, where $f:[0,0.5]\to[0,1]$ is a non-increasing function.

2.2. Cooperative Heterogeneous MARL

In this section, we facilitate the analysis of MARL by considering a particular class of MARL with nonuniform interactions, where each agent interacts with all other agents via the aggregated weighted mean-field effect of the population of all agents.

Recall that we use the weighted graph $G_N=(\mathcal{V}_N,\mathcal{E}_N)$ to represent the multi-agent system, in which agents are cooperative and coordinated by a central controller. They share a finite state space $\mathcal{S}$ and take actions from a finite action space $\mathcal{A}$. We denote by $\mathcal{P}(\mathcal{S})$ and $\mathcal{P}(\mathcal{A})$ the space of all probability measures on $\mathcal{S}$ and $\mathcal{A}$, respectively. Furthermore, denote by $\mathcal{B}(\mathcal{S})$ the space of all Borel measures on $\mathcal{S}$.

For each agent i, the *neighborhood empirical measure* is given by

$$\mu_t^{i,W_N}(\cdot):=\frac{1}{N}\sum_{j\in\mathcal{V}_N}\xi_{i,j}^N\delta_{s_t^j}(\cdot), \tag{2.2}$$

where $\delta_{s_t^j}$ denotes Dirac measure at s_t^j, and (See [15] for more details).

At each step $t=0,1,\cdots$, if agent i, $i\in[N]$ at state $s_t^i\in\mathcal{S}$ takes an action $a_t^i\in\mathcal{A}$, then she will receive a reward

$$r^i\Big(s_t^i,\ \mu_t^{i,W_N},\ a_t^i\Big),\quad i\in[N], \tag{2.3}$$

where $r^i:\mathcal{S}\times\mathcal{B}(\mathcal{S})\times\mathcal{A}\to\mathbb{R}$, $i\in[N]$, and she will change to a new state s_{t+1}^i according to a transition probability such that

$$s_{t+1}^i\sim P^i\left(\cdot\,\Big|\, s_t^i,\ \mu_t^{i,W_N},\ a_t^i\right),\quad i\in[N],\ s_0^i\sim\mu\in\mathcal{P}(\mathcal{S}), \tag{2.4}$$

where $P^i:\mathcal{S}\times\mathcal{B}(\mathcal{S})\times\mathcal{A}\to\mathcal{P}(\mathcal{S})$, $i\in[N]$.

(2.3)-(2.4) indicate that the reward and the transition probability of agent i at time t depend on both her individual information (s_t^i,a_t^i) and neighborhood empirical measure μ_t^{i,W_N}.

Furthermore, the policy is assumed to be stationary for simplicity and takes the Markovian form

$$a_t^i\sim\pi^i\left(\cdot|s_t^i,\mu_t^{i,W_N}\right)\in\mathcal{P}(\mathcal{A}),\quad i\in[N], \tag{2.5}$$

which maps agent i's state to a randomized action. (2.5) is called global policy since the policy of agent i depends on both her own state and the aggregate information of the whole population. For each agent i, the space of all global policies is denoted as Π.

Remark 2.2 *It is computationally expensive to collect the aggregate information of the whole population in many practical scenarios. Considering the costly collection of the aggregation information of the whole population, one can restrict the policy to be in a local manner, that is, the policy that the agent i can execute depends solely on her own state information:*

$$a_t^i \sim \pi^i\left(\cdot|s_t^i\right) \in \mathcal{P}(\mathcal{A}), \quad i \in [N].$$

This has been studied in [41] for standard MFC. Precisely, [41] designs locally executable policies such that the resulting discounted sum of average rewards well approximates the optimal value function over all policies. We expect that a similar result holds for GMFC.

Remark 2.3 *When $\xi_{ij}^N \equiv 1$, $r^i \equiv r$, $P^i \equiv P$, $i, j \in [N]$, it corresponds to classical mean-field theory with uniform interactions [9, 23]. Furthermore, our framework is flexible enough to include the nonuniform interactions of actions via $\nu_t^{i,W_N} = \frac{1}{N}\sum_{j\in\mathcal{V}_N} \xi_{i,j}^N \delta_{a_t^j}(\cdot)$.*

The expected discounted accumulated reward of agent i is

$$J_{N,i}(\mu, \pi^1, \ldots, \pi^N) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r^i\left(s_t^i,\ \mu_t^{i,W_N},\ a_t^i\right) \middle| s_0^i \sim \mu, a_t^i \sim \pi^i(\cdot|s_t^i, \mu_t^{i,W_N})\right], \quad (2.6)$$

subject to (2.2)-(2.5) with a discount factor $\gamma \in (0, 1)$.

The objective of this cooperative multi-agent system (2.2)-(2.5) is to find Pareto optimality given in the Definition 2.4 below.

Definition 2.4 (Pareto Optimality) $(\pi^{1,*}, \ldots, \pi^{N,*}) \in \Pi^N$ *is called Pareto optimality for the multi-agent system* (2.2)-(2.5) *if there does not exist* $(\pi^1, \ldots, \pi^N) \in \Pi^N$ *such that*

$$\begin{aligned} &\forall\, 1 \le i \le N,\ J_{N,i}(\mu, \pi^1, \ldots, \pi^N) \ge J_{N,i}(\mu, \pi^{1,*}, \ldots, \pi^{N,*}),\\ &\exists\, 1 \le i \le N,\ J_{N,i}(\mu, \pi^1, \ldots, \pi^N) > J_{N,i}(\mu, \pi^{1,*}, \ldots, \pi^{N,*}). \end{aligned}$$

To study Pareto optimality, we introduce the expected discounted accumulated reward averaged over all agents, i.e.,

$$\begin{aligned} V_N(\mu) &= \sup_{(\pi^1,\ldots,\pi^N)\in\Pi^N} J_N(\mu, \pi^1, \ldots, \pi^N) \quad (2.7)\\ &:= \sup_{(\pi^1,\ldots,\pi^N)\in\Pi^N} \frac{1}{N}\sum_{i=1}^{N} J_{N,i}(\mu, \pi^1, \ldots, \pi^N), \end{aligned}$$

subject to (2.2)-(2.5). Let $(\pi^{1,*}, \ldots, \pi^{N,*}) \in \underset{(\pi^1,\ldots,\pi^N)\in\Pi^N}{\arg\max} J_N(\mu, \pi^1, \ldots, \pi^N)$, then $(\pi^{1,*}, \ldots, \pi^{N,*})$ is shown to be a Pareto optimality in Definition 2.4. Therefore, searching for Pareto optimality of cooperative MARL amounts to solving the optimal policy of (2.7). However, it is always difficult to exactly obtain the optimal policy of cooperative MARL. We consider a weak notion of ε-Pareto optimality.

Definition 2.5 (ε-Pareto Optimality) *$(\pi_\varepsilon^{1,*}, \dots, \pi_\varepsilon^{N,*}) \in \Pi^N$ is called ε-Pareto optimality for the multi-agent system (2.2)-(2.5) if there does not exist $(\pi^1, \dots, \pi^N) \in \Pi^N$ such that*

$$\begin{aligned} &\forall\, 1 \le i \le N,\ J_{N,i}(\mu, \pi^1, \dots, \pi^N) \ge J_{N,i}(\mu, \pi_\varepsilon^{1,*}, \dots, \pi_\varepsilon^{N,*}) + \varepsilon, \\ &\exists\, 1 \le i \le N,\ J_{N,i}(\mu, \pi^1, \dots, \pi^N) > J_{N,i}(\mu, \pi_\varepsilon^{1,*}, \dots, \pi_\varepsilon^{N,*}) + \varepsilon. \end{aligned}$$

For any $\varepsilon > 0$, let $(\pi_\varepsilon^{1,*}, \dots, \pi_\varepsilon^{N,*}) \in \Pi^N$ such that

$$J_N(\mu, \pi_\varepsilon^{1,*}, \dots, \pi_\varepsilon^{N,*}) \ge \sup_{(\pi^1,\dots,\pi^N)\in\Pi^N} J_N(\mu, \pi^1, \dots, \pi^N) - \varepsilon, \tag{2.8}$$

then $(\pi_\varepsilon^{1,*}, \dots, \pi_\varepsilon^{N,*}) \in \Pi^N$ is an ε-Pareto Optimality in Definition 2.5.

2.3. Graphon Mean-Field Control

We expect the cooperative MARL (2.2)-(2.7) to become a GMFC problem as $N \to \infty$. In GMFC, there is a continuum of agents $\alpha \in \mathcal{I}$, and each agent with the index $\alpha \in \mathcal{I}$ follows

$$s_0^\alpha \sim \mu^\alpha,\ \ a_t^\alpha \sim \pi^\alpha(\cdot|s_t^\alpha, \mu_t^{\alpha,W}),\ \ s_{t+1}^\alpha \sim P^\alpha(\cdot|s_t^\alpha, \mu_t^{\alpha,W}, a_t^\alpha), \tag{2.9}$$

where $\mu_t^\alpha = \mathcal{L}(s_t^\alpha)$, $\alpha \in \mathcal{I}$ denotes the probability distribution of s_t^α, and $\mu_t^{\alpha,W}$ is defined as the *neighborhood mean-field measure* of agent α:

$$\mu_t^{\alpha,W} = \int_{\mathcal{I}} W(\alpha,\beta)\mu_t^\beta d\beta \in \mathcal{B}(\mathcal{S}), \tag{2.10}$$

with the graphon W given in Assumption 2.1.

To ease the sequel analysis, define the space of state distribution ensembles $\boldsymbol{\mathcal{M}} := \mathcal{P}(\mathcal{S})^{\mathcal{I}} := \{f : \mathcal{I} \to \mathcal{P}(\mathcal{S})\}$ and the space of policy ensembles $\boldsymbol{\Pi} := \mathcal{P}(\mathcal{A})^{\mathcal{S}\times\mathcal{I}}$. Then $\boldsymbol{\mu} := (\mu^\alpha)_{\alpha\in\mathcal{I}}$ and $\boldsymbol{\pi} := (\pi^\alpha)_{\alpha\in\mathcal{I}}$ are elements in $\boldsymbol{\mathcal{M}}$ and $\boldsymbol{\Pi}$, respectively.

The objective of GMFC is to maximize the expected discounted accumulated reward averaged over all agents $\alpha \in \mathcal{I}$

$$\begin{aligned} V(\boldsymbol{\mu}) :&= \sup_{\boldsymbol{\pi}\in\boldsymbol{\Pi}} J(\boldsymbol{\mu}, \boldsymbol{\pi}) \qquad (2.11) \\ &= \sup_{\boldsymbol{\pi}\in\boldsymbol{\Pi}} \int_{\mathcal{I}} \mathbb{E}\left[\sum_{t=0}^\infty \gamma^t r^\alpha(s_t^\alpha,\ \mu_t^{\alpha,W},\ a_t^\alpha) \,\middle|\, s_0^\alpha \sim \mu^\alpha, a_t^\alpha \sim \pi^\alpha(\cdot|s_t^\alpha, \mu_t^{\alpha,W})\right] d\alpha. \end{aligned}$$

3. Main Results

3.1. Reformulation of GMFC

In this section, we show that GMFC (2.9)-(2.11) can be reformulated as a MDP with deterministic dynamics and continuous state-action space $\boldsymbol{\mathcal{M}} \times \boldsymbol{\Pi}$.

Theorem 3.1 *GMFC (2.9)-(2.11) can be reformulated as*

$$V(\boldsymbol{\mu}) = \sup_{\boldsymbol{\pi}\in\boldsymbol{\Pi}} \sum_{t=0}^{\infty} \gamma^t R(\boldsymbol{\mu}_t, \boldsymbol{\pi}(\boldsymbol{\mu}_t)), \tag{3.1}$$

subject to

$$\mu^{\alpha}_{t+1}(\cdot) = \boldsymbol{\Phi}^{\alpha}(\boldsymbol{\mu}_t, \boldsymbol{\pi}(\boldsymbol{\mu}_t))(\cdot),\ t \in \mathbb{N},\ \mu^{\alpha}_0 = \mu^{\alpha},\ \alpha \in \mathcal{I}, \tag{3.2}$$

where the aggregated reward $R : \boldsymbol{\mathcal{M}} \times \boldsymbol{\Pi} \to \mathbb{R}$ *and the aggregated transition dynamics* $\boldsymbol{\Phi} : \boldsymbol{\mathcal{M}} \times \boldsymbol{\Pi} \to \boldsymbol{\mathcal{M}}$ *are given by*

$$R(\boldsymbol{\mu}, \boldsymbol{\pi}(\boldsymbol{\mu})) = \int_{\mathcal{I}} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} r^{\alpha}(s, a, \mu^{\alpha,W}) \pi^{\alpha}(a|s, \mu^{\alpha,W}) \mu^{\alpha}(s) d\alpha, \tag{3.3}$$

$$\boldsymbol{\Phi}^{\alpha}(\boldsymbol{\mu}, \boldsymbol{\pi}(\boldsymbol{\mu}))(\cdot) = \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} P^{\alpha}(\cdot|s, \mu^{\alpha,W}, a) \pi^{\alpha}(a|s, \mu^{\alpha,W}) \mu^{\alpha}(s). \tag{3.4}$$

The proof of Theorem 3.1 is similar to the proof of Lemma 2.2 in [24]. So we omit it here.

(3.4) and (3.2) indicate the evolution of the state distribution ensemble $\boldsymbol{\mu}_t$ over time. That is, under the fixed policy ensemble $\boldsymbol{\pi}$, the state distribution μ^{α}_{t+1} of agent α at time $t+1$ is fully determined by the policy ensemble $\boldsymbol{\pi}$ and the state distribution ensemble $\boldsymbol{\mu}_t$ at time t. Note that the change of *population state distribution ensemble* will affect *neighborhood mean-field measure*. In turn, the change of *neighborhood mean-field measure* will have an influence on *population state distribution ensemble*.

With the reformulation in Theorem 3.1, the associated Q function starting from $(\boldsymbol{\mu}, \boldsymbol{\pi}) \in \boldsymbol{\mathcal{M}} \times \boldsymbol{\Pi}$ is defined as

$$Q(\boldsymbol{\mu}, \boldsymbol{\pi}) \;=\; R(\boldsymbol{\mu}, \boldsymbol{\pi}(\boldsymbol{\mu})) + \sup_{\boldsymbol{\pi}'\in\boldsymbol{\Pi}} \Big[\sum_{t=1}^{\infty} \gamma^t R(\boldsymbol{\mu}_t, \boldsymbol{\pi}'(\boldsymbol{\mu}_t)) \;\Big|\; s^{\alpha}_0 \sim \mu^{\alpha}, a^{\alpha}_0 \sim \pi^{\alpha}(\cdot|s^{\alpha}_0, \mu^{\alpha,W}) \Big] \tag{3.5}$$

Hence its Bellman equation is given by

$$Q(\boldsymbol{\mu}, \boldsymbol{\pi}) = R(\boldsymbol{\mu}, \boldsymbol{\pi}(\boldsymbol{\mu})) + \gamma \sup_{\boldsymbol{\pi}'\in\boldsymbol{\Pi}} Q(\boldsymbol{\Phi}(\boldsymbol{\mu}, \boldsymbol{\pi}(\boldsymbol{\mu})), \boldsymbol{\pi}'). \tag{3.6}$$

Remark 3.2 *(Label-state formulation) GMFC (2.9)-(2.11) can be viewed as a classical MFC with extended state space* $\mathcal{S}\times\mathcal{I}$*, action space* $\mathcal{A}$*, policy* $\tilde{\pi} \in \mathcal{P}(\mathcal{A})^{\mathcal{S}\times\mathcal{I}}$*, mean-field information* $\tilde{\mu} \in \mathcal{P}(\mathcal{S}\times\mathcal{I})$*, reward* $\tilde{r}((s,\alpha), \tilde{\mu}, a) := r((s,\alpha), \int_{\mathcal{I}} W(\alpha,\beta)\tilde{\mu}(\cdot,\beta)d\beta, a)$*, transition dynamics of* $(\tilde{s}_t, \alpha_t)$ *such that*

$$\tilde{s}_{t+1} \sim P(\cdot|(\tilde{s}_t, \alpha_t), \tilde{a}_t, \int_{\mathcal{I}} W(\alpha_t,\beta)\tilde{\mu}_t(\cdot,\beta)d\beta),\ \alpha_{t+1} = \alpha_t,\ \tilde{a}_t \sim \tilde{\pi}(\cdot|\tilde{s}_t, \alpha_t, \int_{\mathcal{I}} W(\alpha_t,\beta)\tilde{\mu}_t(\cdot,\beta)d\beta),$$

with the initial condition $\tilde{s}_0 \sim \mu_0$, $\tilde{\alpha}_0 \sim Unif(0,1)$*. It is worth pointing out such a label-state formulation has also been studied in GMFG [29, 15].*

3.2. Approximation

In this section, we show that GMFC (2.9)-(2.11) provides a good approximation for the cooperative multi-agent system (2.2)-(2.7) in terms of the value function and the optimal policy ensemble. To do this, the following assumptions on W, P, r, and $\boldsymbol{\pi}$ are needed.

Assumption 3.3 (graphon W) *There exists $L_W > 0$ such that for all $\alpha, \alpha', \beta, \beta' \in \mathcal{I}$*

$$|W(\alpha,\beta) - W(\alpha',\beta')| \leq L_W \cdot \Big(|\alpha - \alpha'| + |\beta - \beta'|\Big).$$

Assumption 3.3 is common in graphon mean-field theory [21, 15, 29]. Indeed, the Lipschitz continuity assumption on W in Assumption 3.3 can be relaxed to piecewise Lipschitz continuity on W.

Assumption 3.4 (transition probability P) *There exists $L_P > 0$ and $\tilde{L}_P > 0$ such that for any $\alpha, \beta \in \mathcal{I}$, all $s \in \mathcal{S}, a \in \mathcal{A}$, $\mu_1, \mu_2 \in \mathcal{B}(\mathcal{S})$*

$$\|P^\alpha(\cdot|s,\mu_1,a) - P^\beta(\cdot|s,\mu_2,a)\|_1 \leq L_P \cdot \|\mu_1 - \mu_2\|_1 + \tilde{L}_P \cdot |\alpha - \beta|,$$

where $\|\cdot\|_1$ denotes L^1 norm here and throughout the paper.

Assumption 3.5 (reward r) *There exist $M_r > 0$, $L_r > 0$ and $\tilde{L}_r > 0$ such that for all $s \in \mathcal{S}, a \in \mathcal{A}$, $\mu_1, \mu_2 \in \mathcal{B}(\mathcal{S})$,*

$$|r^\alpha(s,\mu,a)| \leq M_r,\ |r^\alpha(s,\mu_1,a) - r^\beta(s,\mu_2,a)| \leq L_r \cdot ||\mu_1 - \mu_2||_1 + \tilde{L}_r \cdot |\alpha - \beta|.$$

Assumption 3.6 (policy $\boldsymbol{\pi}$) *There exists $L_{\boldsymbol{\Pi}} > 0$ and $\tilde{L}_{\boldsymbol{\Pi}} > 0$ such that for any policy ensemble $\boldsymbol{\pi} := (\pi^\alpha)_{\alpha \in \mathcal{I}} \in \boldsymbol{\Pi}$ is Lipschitz continuous, that is, for any $\alpha, \beta \in \mathcal{I}$ and $\mu_1, \mu_2 \in \mathcal{B}(\mathcal{S})$,*

$$\max_{s \in \mathcal{S}} \|\pi^\alpha(\cdot|s,\mu_1) - \pi^\beta(\cdot|s,\mu_2)\|_1 \leq L_{\boldsymbol{\Pi}} \cdot ||\mu_1 - \mu_2||_1 + \tilde{L}_{\boldsymbol{\Pi}}| \cdot \alpha - \beta|.$$

Assumptions 3.3-3.6 state that W, P, r and $\boldsymbol{\pi}$ are Lipschitz continuous with respect to both the index of the agent and the neighborhood mean-field measure. The distance between indexes $|\alpha - \beta|$ measures the similarity of agents. If P, r and $\boldsymbol{\pi}$ are identical, Assumptions 3.4-3.6 are commonly used to bridge the multi-agent system and classical mean-field theory [23, 41, 42, 43].

To show approximation properties of GMFC in the large-scale multi-agent system, we need to relate policy ensembles of GMFC to policies of the multi-agent system. On one hand, one can see that any $\boldsymbol{\pi} \in \boldsymbol{\Pi}$ leads to a N-agent policy tuple $(\pi^1, \dots, \pi^N) \in \Pi^N$ with

$$\Gamma^N : \boldsymbol{\Pi} \ni \boldsymbol{\pi} \mapsto (\pi^1, \dots, \pi^N) \in \Pi^N, \quad \text{with } \pi^i := \boldsymbol{\pi}^{\frac{i}{N}}. \tag{3.7}$$

On the other hand, any N-agent policy tuple $(\pi^1, \dots, \pi^N) \in \Pi^N$ can be seen as a step policy ensemble $\boldsymbol{\pi}^N$ in $\boldsymbol{\Pi}$:

$$\boldsymbol{\pi}^{N,\alpha} := \sum_{i=1}^N \pi^i \mathbf{1}_{\alpha \in (\frac{i-1}{N}, \frac{i}{N}]} \in \boldsymbol{\Pi}. \tag{3.8}$$

Similarly, any N-agent reward tuple $(r^1, \dots, r^N)$ can be regarded as a step reward function of GMFC:

$$r^{N,\alpha} := \sum_{i=1}^{N} r^i \mathbf{1}_{\alpha \in (\frac{i-1}{N}, \frac{i}{N}]}. \tag{3.9}$$

Theorem 3.7 (Approximate Pareto Property) *Assume Assumptions 2.1, 3.3, 3.4, 3.5 and 3.6. Then under either the condition (C1) or (C2), we have for any initial distribution* $\mu \in \mathcal{P}(\mathcal{S})$

$$|V_N(\mu) - V(\mu)| \to 0, \quad \text{as } N \to \infty. \tag{3.10}$$

Moreover, if the graphon convergence in Assumption 2.1 is at rate $\mathcal{O}(\frac{1}{\sqrt{N}})$*, then* $|V_N(\mu) - V(\mu)| = \mathcal{O}(\frac{1}{\sqrt{N}})$*. As a consequence, for any* $\varepsilon > 0$*, there exists an integer* N_ε *such that when* $N \geq N_\varepsilon$*, the optimal policy ensemble of GMFC denoted as* $\boldsymbol{\pi}^*$ *(if it exists) provides an* ε*-Pareto optimality* $(\pi^{1,*}, \dots, \pi^{N,*}) := \Gamma^N(\boldsymbol{\pi}^*)$ *for the multi-agent system (2.7), with* Γ^N *defined in (3.7).*

Theorem 3.7 implies that if we could compute an algorithm to learn the optimal policy ensemble of GMFC, then the learned optimal policy ensemble is close to the optimal policy of MARL. Directly learning the optimal policy of GMFC, however, will lead to high complexity due to the infinite-dimensional feature of $\boldsymbol{\mu}$ and $\boldsymbol{\pi}$. Instead, we will introduce a smaller class of GMFC with a lower dimension in the next section, which enables a scalable algorithm.

3.3. Algorithm Design and Convergence Analysis

This section will show that discretizing the graphon index $\alpha \in \mathcal{I}$ of GMFC enables to approximate Q function in (3.6) by an approximated Q function in (3.11) below defined on a smaller space, which is critical for designing efficient learning algorithms.

Precisely, we choose uniform grids $\alpha_m \in \mathcal{I}_M := \{\frac{m}{M}, 0 \leq m \leq M\}$ for simplicity, and approximate each agent $\alpha \in \mathcal{I}$ by the nearest $\alpha_m \in \mathcal{I}_M$ close to it. Introduce $\widetilde{\boldsymbol{\mathcal{M}}}_M := \mathcal{P}(\mathcal{S})^{\mathcal{I}_M}$, $\widetilde{\boldsymbol{\Pi}}_M := \mathcal{P}(\mathcal{A})^{\mathcal{S} \times \mathcal{I}_M}$. Meanwhile, $\tilde{\boldsymbol{\mu}} := (\tilde{\mu}^{\alpha_m})_{m \in [M]} \in \widetilde{\boldsymbol{\mathcal{M}}}_M$ and $\tilde{\boldsymbol{\pi}} := (\tilde{\pi}^{\alpha_m})_{m \in [M]} \in \widetilde{\boldsymbol{\Pi}}_M$ can be viewed as a piecewise constant state distribution ensemble in $\boldsymbol{\mathcal{M}}$ and a piecewise constant policy ensemble in $\boldsymbol{\Pi}$, respectively. Our arguments can be easily generalized to nonuniform grids.

Consequently, instead of performing algorithms according to (3.6) with a continuum of graphon labels directly, we work with GMFC with M blocks called **block GMFC**, in which agents in the same block are homogeneous. The Bellman equation for Q function of block GMFC is given by

$$\widetilde{Q}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\pi}}) = \widetilde{R}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\pi}}(\tilde{\boldsymbol{\mu}})) + \gamma \sup_{\tilde{\boldsymbol{\pi}}' \in \widetilde{\boldsymbol{\Pi}}_M} \widetilde{Q}(\widetilde{\boldsymbol{\Phi}}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\pi}}(\tilde{\boldsymbol{\mu}})), \tilde{\boldsymbol{\pi}}'), \tag{3.11}$$

where the neighborhood mean-field measure, the aggregated reward $\widetilde{R} : \widetilde{\boldsymbol{\mathcal{M}}}_M \times \widetilde{\boldsymbol{\Pi}}_M \to \mathbb{R}$ and the aggregated transition dynamics $\widetilde{\boldsymbol{\Phi}} : \widetilde{\boldsymbol{\mathcal{M}}}_M \times \widetilde{\boldsymbol{\Pi}}_M \to \widetilde{\boldsymbol{\mathcal{M}}}_M$ are given by

$$\tilde{\mu}^{\alpha_m, W} = \frac{1}{M} \sum_{m'=0}^{M-1} W(\alpha_m, \alpha_{m'}) \tilde{\mu}^{\alpha_{m'}}, m \in [M], \tag{3.12}$$

$$\widetilde{R}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\pi}}(\tilde{\boldsymbol{\mu}})) = \frac{1}{M} \sum_{m=0}^{M-1} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} r^{\alpha_m}(s, a, \tilde{\mu}^{\alpha_m, W}) \tilde{\mu}^{\alpha_m}(s) \tilde{\pi}^{\alpha_m}(a|s, \tilde{\mu}^{\alpha_m, W}), \tag{3.13}$$

$$\widetilde{\boldsymbol{\Phi}}^{\alpha_m}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\pi}}(\tilde{\boldsymbol{\mu}}))(\cdot) = \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} P^{\alpha_m}(\cdot|s, a, \tilde{\mu}^{\alpha_m, W}) \tilde{\mu}^{\alpha_m}(s) \tilde{\pi}^{\alpha_m}(a|s, \tilde{\mu}^{\alpha_m, W}). \tag{3.14}$$

We see from (3.11) that block GMFC is a MDP with deterministic dynamics $\widetilde{\boldsymbol{\Phi}}$ and continuous state-action space $\widetilde{\boldsymbol{\mathcal{M}}}_M \times \widetilde{\boldsymbol{\Pi}}_M$. The following Theorem shows that there exists an optimal policy ensemble of block GMFC in $\widetilde{\boldsymbol{\Pi}}_M$.

Theorem 3.8 (Existence of Optimal Policy Ensemble) *Given Assumptions 3.4, 3.5, assume $\gamma \cdot (1 + L_P + L_\Pi) < \infty$, then for any fixed integer $M > 0$, there exists an $\tilde{\boldsymbol{\pi}}^* \in \widetilde{\boldsymbol{\Pi}}_M$ that maximize $\widetilde{Q}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\pi}})$ in (3.11) for any $\tilde{\boldsymbol{\mu}} \in \widetilde{\boldsymbol{\mathcal{M}}}_M$.*

Furthermore, we show that with sufficiently fine partitions of the graphon index $\mathcal{I}$, i.e., M is sufficiently large, block GMFC (3.11)-(3.14) well approximates the multi-agent system in Section 2.2.

Theorem 3.9 *Assume $\gamma \cdot (1 + L_P + L_\Pi) < \infty$ and Assumptions 2.1, 3.3, 3.4, 3.5 and 3.6. Under either (C1) or (C2), for any $\varepsilon > 0$, there exists N_ε, M_ε such that for $N \geq N_\varepsilon$, the optimal policy ensemble $\tilde{\boldsymbol{\pi}}^*$ of block GMFC (3.11) with M_ε blocks provides an ε-Pareto optimality $(\tilde{\pi}^{1,*}, \ldots, \tilde{\pi}^{N,*}) := \Gamma^N(\tilde{\boldsymbol{\pi}}^*)$ for the multi-agent system (2.7) with N agents.*

Theorem 3.9 shows that the optimal policy ensemble of block GMFC is near-optimal for *all* sufficiently large multi-agent systems, meaning that block GMFC provides a good approximation for the multi-agent system. Therefore, If we could develop an algorithm for block GMFC to extract an optimal policy ensemble of block GMFC, then the extracted policy is near optimal for MARL.

When model parameters P^α, r^α and W are known, one can easily extract the optimal policy based on Bellman equation. If any of these model parameters P^α, r^α and W is unknown, we take a model-free approach. The key issue is to handle population state distribution ensemble $\tilde{\boldsymbol{\mu}}$, which is an input of $\widetilde{Q}$ function in (3.11). We assume that we have a block GMFC simulator $\mathcal{G}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\pi}}) = (\tilde{\boldsymbol{\mu}}', \widetilde{R})$. That is, for any pair of population state distribution ensemble and policy ensemble $(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\pi}})$, we can sample the aggregated reward $\tilde{R}$ and the next population state distribution ensemble $\tilde{\boldsymbol{\mu}}'$. To learn the optimal policy of block GMFC, one can adopt any existing techniques for standard MFC, such as a kernel-based Q learning method in [23] and a uniform discretization method in [9].

Remark 3.10 *If we can only observe the state of agent $\alpha_m \in \mathcal{I}_M$ and do not have access to population state distribution ensemble, we can estimate $\tilde{\mu}^{\alpha_m}$ following [2] or [42]. However,* *different from [2] and [42], we also need to estimate $\tilde{\mu}^{\alpha_m,W}$ due to the graphon structure W and leave it for future study.*

We choose to adapt DRL algorithm neural Proximal Policy Optimization (PPO) [47, 38] to block GMFC given in Algorithm 1. Following Corollary 4.11 in [38], together with Theorem 3.9, we can state the global convergence of neural PPO for block GMFC. Since assumptions that make the result hold are similar as [38], we do not state these assumptions here.

Algorithm 1 Neural PPO for block GMFC

Input Width of neural network M, radius of constraint R, number of SGD and TD iterations T, number of PPO iteration K, penalty parameter β
Initialize
for $k = 0$ to $K-1$ **do**
set temperature parameter $\tau_{k+1} \leftarrow \frac{\beta\sqrt{K}}{k+1}$ and penalty parameter $\beta_k \leftarrow \beta\sqrt{K}$.
Sample $(\tilde{\boldsymbol{\mu}}_t, \tilde{\boldsymbol{\pi}}_t, \tilde{R}_t, \tilde{\boldsymbol{\mu}}'_t, \tilde{\boldsymbol{\pi}}'_t)_{t=1}^T$ with $\tilde{\boldsymbol{\pi}}_0 \sim \Pi^0(\cdot|\tilde{\boldsymbol{\mu}})$, $\tilde{\boldsymbol{\mu}}'_t = \tilde{\boldsymbol{\Phi}}(\tilde{\boldsymbol{\mu}}_t, \tilde{\boldsymbol{\pi}}_t)$, $\tilde{\boldsymbol{\pi}}_t \sim \Pi^{\theta_k}(\cdot|\tilde{\boldsymbol{\mu}}_t)$.
Solve for Q function parameterized by neural network $Q_{\omega_k} = NN(\omega_k, M)$ using the TD update.
Solve for energy function parameterized by neural network $f_{\theta_{k+1}} = NN(\theta_{k+1}, M)$ using the SGD update.
Update policy: $\Pi^{\theta_k} \propto \exp(\tau_{k+1}^{-1} f_{\theta_{k+1}})$.
end for

Theorem 3.11 *Suppose that Assumptions 2.1, 3.3, 3.4, 3.5 and 3.6 hold. Further assume $\gamma \cdot (1 + L_\Pi + L_P) < 1$. Furthermore, suppose that the width of neural network is sufficiently large. For any $\varepsilon > 0$, there exists M_ε and N_ε such that for any $M \geq M_\varepsilon$ and $N \geq N_\varepsilon$, and the policy attained by Algorithm 1 denoted as $\boldsymbol{\pi}_{PPO}$*

$$|J_N(\mu; \pi^{1,*}, \dots, \pi^{N,*}) - \tilde{J}^M(\mu; \boldsymbol{\pi}_{PPO})| \leq \frac{C}{\sqrt{K}} + \bar{C}\varepsilon, \tag{3.15}$$

where J_N and $\tilde{J}^M$ are given in (2.7) and (4.7) respectively, K is the number of iteration, C and $\bar{C}$ are constants.

By setting $K = \frac{C}{\varepsilon^2}$, the optimal empirical value function of MARL is approximated by the value function of block GMFC under the learned policy in Algorithm 1 with the error $\mathcal{O}(\varepsilon)$. In other words, Theorem 3.11 states that, with a sample complexity of $\mathcal{O}(\frac{1}{\varepsilon^2})$, Algorithm 1 generates a $\mathcal{O}(\varepsilon)$-Pareto optimality of cooperative MARL.

To evaluate the performance of Algorithm 1 and to validate our theoretical findings, we describe the deployment of block GMFC in the multi-agent system in Algorithm 2, which we call it **N-agent GMFC**.

Algorithm 2 N-agent GMFC

Input Initial state distribution μ_0, number of agents N, episode length T, the learned policy $\tilde{\boldsymbol{\pi}} \in \widetilde{\boldsymbol{\Pi}}_M$ learned by PPO
Initialize $s_0^i \sim \mu_0$, $i \in [N]$
for $t = 1$ to T **do**
  for $i = 1$ to N **do**
    Choose $m(i) = \arg\min_{m \in [M]} |\frac{i}{N} - \frac{m}{M}|$
    Sample action $a_t^i \sim \tilde{\pi}^{\alpha_{m(i)}}(\cdot|s_t^i)$, observe reward r_t^i and new state s_{t+1}^i
  end for
end for

4. Proofs of Main Results

In this section, we will provide proofs of Theorems 3.7-3.9.

4.1. Proof of Theorem 3.7

To prove Theorem 3.7, we need the following two Lemmas. We start by defining the step state distribution $\boldsymbol{\mu}_t^N := (\mu_t^{N,\alpha})_{\alpha \in \mathcal{I}}$ for notational simplicity

$$\mu_t^{N,\alpha}(\cdot) = \sum_{i \in \mathcal{V}_N} \delta_{s_t^i}(\cdot) \mathbf{1}_{\alpha \in (\frac{i-1}{N}, \frac{i}{N}]}. \tag{4.1}$$

Lemma 4.1 shows the convergence of the neighborhood empirical measure to the neighborhood mean-field measure.

Lemma 4.1 *Assume Assumptions 2.1, 3.3, 3.4 and 3.6. Under either condition* (C1) *or* (C2)*, for any policy ensemble* $\boldsymbol{\pi} \in \boldsymbol{\Pi}$*, we have*

$$\sum_{i=1}^{N} \int_{(\frac{i-1}{N}, \frac{i}{N}]} \mathbb{E}\big[\|\mu_t^{i,W_N} - \mu_t^{\alpha,W}\|_1\big] d\alpha \to 0, \quad \text{as } N \to \infty, \tag{4.2}$$

where $\mu_t^i = \mu_t^\alpha \equiv \mu \in \mathcal{P}(\mathcal{S})$.

Moreover, if the graphon convergence in Assumption 2.1 is at rate $\mathcal{O}(\frac{1}{\sqrt{N}})$, then

$$\sum_{i=1}^{N} \int_{(\frac{i-1}{N}, \frac{i}{N}]} \mathbb{E}\big[\|\mu_t^{i,W_N} - \mu_t^{\alpha,W}\|_1\big] d\alpha = \mathcal{O}(\frac{1}{\sqrt{N}}).$$

Proof of Lemma 4.1 We first prove (4.2) under the condition (C1) and then show (4.2) also holds under the condition (C2).

Case 1: $\xi^N_{i,j} = W_N(\frac{i}{N}, \frac{j}{N})$. Note that under the condition (C1), $\mu_t^{i,W_N} = \int_{\mathcal{I}} W_N(\frac{i}{N}, \beta)\mu_t^{N,\beta} d\beta$ by the definition of $\mu_t^{N,\alpha}$ in (4.1). Then

$$
\begin{aligned}
& \sum_{i=1}^{N} \int_{(\frac{i-1}{N},\frac{i}{N}]} \mathbb{E}\big[\|\mu_t^{i,W_N} - \mu_t^{\alpha,W}\|_1\big] d\alpha \\
= \; & \sum_{i=1}^{N} \int_{(\frac{i-1}{N},\frac{i}{N}]} \mathbb{E}\Big[\Big\| \int_{\mathcal{I}} W_N(\frac{i}{N},\beta)\mu_t^{N,\beta} d\beta - \int_{\mathcal{I}} W(\alpha,\beta)\mu_t^{\beta} d\beta \Big\|_1\Big] d\alpha \\
\leq \; & \sum_{i=1}^{N} \int_{(\frac{i-1}{N},\frac{i}{N}]} \mathbb{E}\Big[\Big\| \int_{\mathcal{I}} W_N(\frac{i}{N},\beta)\mu_t^{N,\beta} d\beta - \int_{\mathcal{I}} W_N(\frac{i}{N},\beta)\mu_t^{\beta} d\beta \Big\|_1\Big] d\alpha \\
& \quad + \sum_{i=1}^{N} \int_{(\frac{i-1}{N},\frac{i}{N}]} \mathbb{E}\Big[\Big\| \int_{\mathcal{I}} W_N(\frac{i}{N},\beta)\mu_t^{\beta} d\beta - \int_{\mathcal{I}} W(\alpha,\beta)\mu_t^{\beta} d\beta \Big\|_1\Big] d\alpha \\
= \; & : I_1 + I_2.
\end{aligned}
$$

For the term I_1, we adapt Theorem 2 that works with local policy in [15] to our setting of global policy and have that under the policy ensemble π and N-agent policy $(\pi^1, \ldots, \pi^N) := \Gamma_N(\boldsymbol{\pi})$, with Γ_N defined in (3.7)

$$
I_1 = \mathbb{E}\Big[\Big\| \int_{\mathcal{I}} W_N(\frac{i}{N},\beta)\mu_t^{N,\beta} d\beta - \int_{\mathcal{I}} W_N(\frac{i}{N},\beta)\mu_t^{\beta} d\beta \Big\|_1\Big] \to 0, \text{ as } N \to \infty.
$$

Moreover, if the graphon convergence in Assumption 2.1 is at rate $\mathcal{O}(\frac{1}{\sqrt{N}})$, then the term I_1 is also at rate $\mathcal{O}(\frac{1}{\sqrt{N}})$.

By noting that $W_N(\alpha,\beta) = W_N(\frac{\lceil N\alpha \rceil}{N}, \frac{\lceil N\beta \rceil}{N})$,

$$
\begin{aligned}
I_2 \; & = \; \sum_{i=1}^{N} \int_{(\frac{i-1}{N},\frac{i}{N}]} \Big\| \int_{\mathcal{I}} W_N(\frac{\lceil N\alpha \rceil}{N},\beta)\mu_t^{\beta} d\beta - \int_{\mathcal{I}} W(\alpha,\beta)\mu_t^{\beta} d\beta \Big\|_1 d\alpha \\
& = \; \sum_{i=1}^{N} \int_{(\frac{i-1}{N},\frac{i}{N}]} \Big\| \int_{\mathcal{I}} W_N(\alpha,\beta)\mu_t^{\beta} d\beta - \int_{\mathcal{I}} W(\alpha,\beta)\mu_t^{\beta} d\beta \Big\|_1 d\alpha \\
& = \; \int_{\mathcal{I}} \Big\| \int_{\mathcal{I}} W_N(\alpha,\beta)\mu_t^{\beta} d\beta - \int_{\mathcal{I}} W(\alpha,\beta)\mu_t^{\beta} d\beta \Big\|_1 d\alpha \\
& = \; \sum_{s\in\mathcal{S}} \int_{\mathcal{I}} \Big| \int_{\mathcal{I}} W_N(\alpha,\beta)\mu_t^{\beta}(s) d\beta - \int_{\mathcal{I}} W(\alpha,\beta)\mu_t^{\beta}(s) d\beta \Big| d\alpha \\
& \to \; 0,
\end{aligned}
$$

where the last inequality is from the fact in [39] that the convergence of $\|W_N - W\|_\square \to 0$ is equivalent to the convergence of

$$
\|W_N - W\|_{L_\infty \to L_1} := \sup_{\|g\|_\infty \leq 1} \int_{\mathcal{I}} \Big| \int_{\mathcal{I}} \big(W_N(\alpha,\beta) - W(\alpha,\beta)\big) g(\beta) d\beta \Big| d\alpha \to 0.
$$

Combining I_1 and I_2, we prove (4.2) under the condition (C1).

Case 2: $\xi_{i,j}^N$ are random variables with Bernoulli($W_N(\frac{i}{N},\frac{j}{N})$).

$$\begin{aligned}
&\sum_{i=1}^N \int_{(\frac{i-1}{N},\frac{i}{N}]} \mathbb{E}\|\mu_t^{i,W_N} - \mu_t^{\alpha,W}\|_1 d\alpha \\
=\ & \sum_{i=1}^N \int_{(\frac{i-1}{N},\frac{i}{N}]} \mathbb{E}\Big\|\frac{1}{N}\sum_{j=1}^N \xi_{ij}^N \delta_{s_t^j} - \int_{\mathcal{I}} W(\alpha,\beta)\mu_t^\beta d\beta\Big\|_1 d\alpha \\
\le\ & \sum_{i=1}^N \int_{(\frac{i-1}{N},\frac{i}{N}]} \mathbb{E}\Big\|\frac{1}{N}\sum_{j=1}^N \xi_{ij}^N \delta_{s_t^j} - \int_{\mathcal{I}} W_N(\frac{i}{N},\beta)\mu_t^{N,\beta} d\beta\Big\|_1 d\alpha \\
&\quad + \sum_{i=1}^N \int_{(\frac{i-1}{N},\frac{i}{N}]} \mathbb{E}\Big\|\int_{\mathcal{I}} W_N(\frac{i}{N},\beta)\mu_t^{N,\beta} d\beta - \int_{\mathcal{I}} W(\alpha,\beta)\mu_t^\beta d\beta\Big\|_1 d\alpha \\
=:\ & I_1 + I_2.
\end{aligned}$$

Note from **Case 1** that $I_2 \to 0$ as $N\to\infty$ and $I_2 = \mathcal{O}(\frac{1}{\sqrt{N}})$ if the graphon convergence in Assumption 2.1 is at rate $\mathcal{O}(\frac{1}{\sqrt{N}})$. Therefore, it is enough to estimate I_1.

$$\begin{aligned}
I_1 &= \mathbb{E}\Big\|\frac{1}{N}\sum_{j=1}^N \xi_{ij}^N \delta_{s_t^j} - \int_{\mathcal{I}} W_N(\frac{i}{N},\beta)\mu_t^{N,\beta} d\beta\Big\|_1 \\
&\le \mathbb{E}\Big[\mathbb{E}\Big[\sup_{f:\mathcal{S}\to\{-1,1\}} \Big\{\frac{1}{N}\sum_{j=1}^N \xi_{ij}^N f(s_t^j) - \frac{1}{N}\sum_{j=1}^N W_N(\frac{i}{N},\frac{j}{N}) f(s_t^j)\Big\}\Big| s_t^1,\dots,s_t^N\Big]\Big].
\end{aligned}$$

We proceed the same argument as in the proof of Theorem 6.3 in [23]. Precisely, conditioned on $s_t^1,\dots,s_t^N$, $\Big\{\xi_{ij}^N f(s_t^j) - W_N(\frac{i}{N},\frac{j}{N}) f(s_t^j)\Big\}_{j=1}^N$ is a sequence of independent mean-zero random variables bounded in $[-1,1]$ due to $\mathbb{E}[\xi_{i,j}^N] = W_N(\frac{i}{N},\frac{j}{N})$. This implies that each $\xi_{ij}^N f(s_t^j) - W_N(\frac{i}{N},\frac{j}{N}) f(s_t^j)$ is a sub-Gaussian with variance bounded by 4. As a result, conditioned on $s_t^1,\dots,s_t^N$, $\Big\{\frac{1}{N}\sum_{j=1}^N \xi_{ij}^N f(s_t^j) - \frac{1}{N}\sum_{j=1}^N W_N(\frac{i}{N},\frac{j}{N}) f(s_t^j)\Big\}_{i=1}^N$ is a mean-zero sub-Gaussian random variable with variance $\frac{4}{N}$. By the equation (2.66) in [51], we have

$$\begin{aligned}
I_1 &\le \mathbb{E}\Big[\mathbb{E}\Big[\sup_{f:\mathcal{S}\to\{-1,1\}} \Big\{\frac{1}{N}\sum_{j=1}^N \xi_{ij}^N f(s_t^j) - \frac{1}{N}\sum_{j=1}^N W_N(\frac{i}{N},\frac{j}{N}) f(s_t^j)\Big\}\Big| s_t^1,\dots,s_t^N\Big]\Big] \\
&\le \frac{\sqrt{8\ln(2)|\mathcal{S}|}}{\sqrt{N}}.
\end{aligned}$$

Therefore, combining I_1 and I_2 in **Case 2**, we show that when $\xi_{i,j}^N$ are random variables with Bernoulli($W_N(\frac{i}{N},\frac{j}{N})$), (4.2) holds under the condition (C2). □

Lemma 4.2 shows the convergence of the state distribution of N-agent game to the state distribution of GMFC.

Lemma 4.2 *Assume Assumptions 2.1, 3.3, 3.4 and 3.6. For any uniformly bounded family $\mathcal{G}$ of functions $g^\alpha : \mathcal{S} \to \mathbb{R}$, we have*

$$\sup_{\{g^\alpha\}_{\alpha\in\mathcal{I}}\in\mathcal{G}} \sum_{i=1}^{N} \int_{(\frac{i-1}{N},\frac{i}{N}]} \Big|\mathbb{E}[g^\alpha(s_t^i) - g^\alpha(s_t^\alpha)]\Big| d\alpha \to 0, \tag{4.3}$$

where $s_0^i \sim \mu_0$, $s_0^\alpha \sim \mu_0$. Moreover, if the graphon convergence in Assumption 2.1 is at rate $\mathcal{O}(\frac{1}{\sqrt{N}})$, then

$$\sup_{\{g^\alpha\}_{\alpha\in\mathcal{I}}\in\mathcal{G}} \sum_{i=1}^{N} \int_{(\frac{i-1}{N},\frac{i}{N}]} \Big|\mathbb{E}[g^\alpha(s_t^i) - g^\alpha(s_t^\alpha)]\Big| d\alpha = \mathcal{O}(\frac{1}{\sqrt{N}}).$$

Proof of Lemma 4.2 The proof is by induction as follows. To do this, first introduce

$$l_{g^\alpha}^{\beta}(s,\mu,\pi) := \sum_{a\in\mathcal{A}}\sum_{s'\in\mathcal{S}} g^\alpha(s')P^\beta(s'|s,\mu,a)\pi(a|s,\mu).$$

(4.3) holds obviously at $t = 0$. Suppose that (4.3) holds at t. Then for any uniformly bounded function g^α with $|g^\alpha| \le M_g$ at $t+1$

$$\begin{aligned}
& \sum_{i=1}^{N} \int_{(\frac{i-1}{N},\frac{i}{N}]} \Big|\mathbb{E}[g^\alpha(s_{t+1}^i) - g^\alpha(s_{t+1}^\alpha)]\Big| d\alpha \\
= & \sum_{i=1}^{N} \int_{(\frac{i-1}{N},\frac{i}{N}]} \Big|\mathbb{E}\big[l_{g^\alpha}^{\frac{i}{N}}(s_t^i,\mu_t^{i,W_N},\pi^i)\big] - \mathbb{E}\big[l_{g^\alpha}^{\alpha}(s_t^\alpha,\mu_t^{\alpha,W},\pi^\alpha)\big]\Big| d\alpha \\
\le & \sum_{i=1}^{N} \int_{(\frac{i-1}{N},\frac{i}{N}]} \Big|\mathbb{E}\big[l_{g^\alpha}^{\frac{i}{N}}(s_t^i,\mu_t^{i,W_N},\pi^i)\big] - \mathbb{E}\big[l_{g^\alpha}^{\alpha}(s_t^i,\mu_t^{\alpha,W},\pi^i)\big]\Big| d\alpha \\
& + \sum_{i=1}^{N} \int_{(\frac{i-1}{N},\frac{i}{N}]} \Big|\mathbb{E}\big[l_{g^\alpha}^{\alpha}(s_t^i,\mu_t^{\alpha,W},\pi^i)\big] - \mathbb{E}\big[l_{g^\alpha}^{\alpha}(s_t^\alpha,\mu_t^{\alpha,W},\pi^i)\big]\Big| d\alpha \\
& + \sum_{i=1}^{N} \int_{(\frac{i-1}{N},\frac{i}{N}]} \Big|\mathbb{E}\big[l_{g^\alpha}^{\alpha}(s_t^\alpha,\mu_t^{\alpha,W},\pi^i)\big] - \mathbb{E}\big[l_{g^\alpha}^{\alpha}(s_t^\alpha,\mu_t^{\alpha,W},\pi^\alpha)\big]\Big| d\alpha \\
= & : I + II + III,
\end{aligned} \tag{4.4}$$

where the first equality is by the law of total expectation.

First term of (4.4).

$$\begin{aligned}
I &= \sum_{i=1}^{N} \int_{(\frac{i-1}{N},\frac{i}{N}]} \Big|\mathbb{E}\big[l_{g^\alpha}^{\frac{i}{N}}(s_t^i,\mu_t^{i,W_N},\pi^i)\big] - \mathbb{E}\big[l_{g^\alpha}^{\alpha}(s_t^i,\mu_t^{\alpha,W},\pi^i)\big]\Big| d\alpha \\
&\le M_g\Big(L_P \sum_{i=1}^{N} \int_{(\frac{i-1}{N},\frac{i}{N}]} \mathbb{E}\big[\|\mu_t^{i,W_N} - \mu_t^{\alpha,W}\|_1\big] d\alpha + \tilde{L}_P \sum_{i=1}^{N} \int_{(\frac{i-1}{N},\frac{i}{N}]} |\alpha - \frac{i}{N}| d\alpha\Big) \\
&\to 0, \quad \text{as } N \to \infty
\end{aligned}$$

where the second inequality is from the continuity of P, and the last inequality is from Lemma 4.1.

Second term of (4.4). One can view $l^{\alpha}_{g^{\alpha}}(s, \mu^{\alpha,W}_t, \pi^i)$ as a function of $s \in \mathcal{S}$ for any fixed $\mu^{\alpha,W}_t$ and π^i, $\alpha \in \mathcal{I}$. Note that $|l^{\alpha}_{g^{\alpha}}(s, \mu^{\alpha,W}_t, \pi^i)| \leq M_g$, where M_g is a constant independent of $\mu^{\alpha,W}_t$, π^i. Since (4.3) holds at t, then

$$\begin{aligned} II &= \sum_{i=1}^{N} \int_{(\frac{i-1}{N}, \frac{i}{N}]} \Big| \mathbb{E}\big[l^{\alpha}_{g^{\alpha}}(s^i_t, \mu^{\alpha,W}_t, \pi^i)\big] - \mathbb{E}\big[l^{\alpha}_{g^{\alpha}}(s^{\alpha}_t, \mu^{\alpha,W}_t, \pi^i)\big] \Big| d\alpha \\ &\to 0, \quad \text{as } N \to \infty. \end{aligned}$$

Third term of (4.4).

$$\begin{aligned} III &= \sum_{i=1}^{N} \int_{(\frac{i-1}{N}, \frac{i}{N}]} \Big| \mathbb{E}\big[l^{\alpha}_{g^{\alpha}}(s^{\alpha}_t, \mu^{\alpha,W}_t, \pi^i)\big] - \mathbb{E}\big[l^{\alpha}_{g^{\alpha}}(s^{\alpha}_t, \mu^{\alpha,W}_t, \pi^{\alpha})\big] \Big| d\alpha \\ &\leq M_g \sum_{i=1}^{N} \int_{(\frac{i-1}{N}, \frac{i}{N}]} \mathbb{E}\big[\|\pi^i(s^{\alpha}_t) - \pi^{\alpha}(s^{\alpha}_t)\|_1\big] d\alpha \\ &\leq M_g L_{\Pi} \sum_{i=1}^{N} \int_{(\frac{i-1}{N}, \frac{i}{N}]} \max_{\alpha \in (\frac{i-1}{N}, \frac{i}{N}]} \Big|\frac{i}{N} - \alpha\Big| d\alpha \\ &= \mathcal{O}(\frac{1}{N}), \end{aligned}$$

where the second inequality is by the uniform boundedness of g and the third inequality is from Assumption 3.6. □

Now we are ready to prove Theorem 3.7. We start by defining $\widehat{r}^{\alpha}$ the aggregated reward over all possible actions under the policy π

$$\widehat{r}^{\alpha}(s, \mu, \pi) := \sum_{a \in \mathcal{A}} r^{\alpha}(s, \mu, a)\pi(a|s, \mu).$$

Proof of Theorem 3.7

$$
\begin{aligned}
&|V_N(\mu) - V(\mu)| \\
&= \Big|\sup_{\Pi^N} \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\Big[\sum_{t=0}^{\infty}\gamma^t r^i(s_t^i,\ \mu_t^{i,W_N},\ a_t^i)\Big] - \sup_{\boldsymbol{\pi}\in\boldsymbol{\Pi}}\int_{\mathcal{I}}\mathbb{E}\Big[\sum_{t=0}^{\infty}\gamma^t r^\alpha(s_t^\alpha,\ \mu_t^{\alpha,W},\ a_t^\alpha)\Big]d\alpha\Big| \\
&\le \sup_{\boldsymbol{\pi}\in\boldsymbol{\Pi}}\Big|\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\Big[\sum_{t=0}^{\infty}\gamma^t r^i(s_t^i,\ \mu_t^{i,W_N},\ a_t^i)\Big] - \int_{\mathcal{I}}\mathbb{E}\Big[\sum_{t=0}^{\infty}\gamma^t r^\alpha(s_t^\alpha,\ \mu_t^{\alpha,W},\ a_t^\alpha)\Big]d\alpha\Big| \\
&= \sup_{\boldsymbol{\pi}\in\boldsymbol{\Pi}}\Big|\sum_{t=0}^{\infty}\gamma^t\sum_{i=1}^{N}\int_{(\frac{i-1}{N},\frac{i}{N}]}\Big(\mathbb{E}\big[\widehat{r}^i(s_t^i,\ \mu_t^{i,W_N},\ \pi^i)\big] - \mathbb{E}\big[\widehat{r}^\alpha(s_t^\alpha,\ \mu_t^{\alpha,W},\ \pi^\alpha)\big]\Big)d\alpha\Big| \\
&\le \sup_{\boldsymbol{\pi}\in\boldsymbol{\Pi}}\Big|\sum_{t=0}^{\infty}\gamma^t\sum_{i=1}^{N}\int_{(\frac{i-1}{N},\frac{i}{N}]}\Big(\mathbb{E}\big[\widehat{r}^i(s_t^i,\ \mu_t^{i,W_N},\ \pi^i)\big] - \mathbb{E}\big[\widehat{r}^\alpha(s_t^i,\ \mu_t^{\alpha,W},\ \pi^i)\big]\Big)d\alpha\Big| \\
&\quad + \sup_{\boldsymbol{\pi}\in\boldsymbol{\Pi}}\Big|\sum_{t=0}^{\infty}\gamma^t\sum_{i=1}^{N}\int_{(\frac{i-1}{N},\frac{i}{N}]}\Big(\mathbb{E}\big[\widehat{r}^\alpha(s_t^i,\ \mu_t^{\alpha,W},\ \pi^i)\big] - \mathbb{E}\big[\widehat{r}^\alpha(s_t^\alpha,\ \mu_t^{\alpha,W},\ \pi^i)\big]\Big)d\alpha\Big| \\
&\quad + \sup_{\boldsymbol{\pi}\in\boldsymbol{\Pi}}\Big|\sum_{t=0}^{\infty}\gamma^t\sum_{i=1}^{N}\int_{(\frac{i-1}{N},\frac{i}{N}]}\Big(\mathbb{E}\big[\widehat{r}^\alpha(s_t^\alpha,\ \mu_t^{\alpha,W},\ \pi^i)\big] - \mathbb{E}\big[\widehat{r}^\alpha(s_t^\alpha,\ \mu_t^{\alpha,W},\ \pi^\alpha)\big]\Big)d\alpha\Big| \\
&:= I + II + III,
\end{aligned}
\tag{4.5}
$$

360 where we use (3.8) in the second inequality.

First term of (4.5).

$$
\begin{aligned}
I &= \sup_{\boldsymbol{\pi}\in\boldsymbol{\Pi}}\Big|\sum_{t=0}^{\infty}\gamma^t\sum_{i=1}^{N}\int_{(\frac{i-1}{N},\frac{i}{N}]}\Big(\mathbb{E}\big[\widehat{r}^i(s_t^i,\ \mu_t^{i,W_N},\ \pi^i)\big] - \mathbb{E}\big[\widehat{r}^\alpha(s_t^i,\ \mu_t^{\alpha,W},\ \pi^i)\big]\Big)d\alpha\Big| \\
&= \sup_{\boldsymbol{\pi}\in\boldsymbol{\Pi}}\Big|\sum_{t=0}^{\infty}\gamma^t\sum_{i=1}^{N}\int_{(\frac{i-1}{N},\frac{i}{N}]}\Big(\mathbb{E}\big[\widehat{r}^i(s_t^i,\ \mu_t^{i,W_N},\ \pi^i)\big] - \mathbb{E}\big[\widehat{r}^i(s_t^i,\ \mu_t^{\alpha,W},\ \pi^i)\big]\Big)d\alpha\Big| \\
&\quad + \sup_{\boldsymbol{\pi}\in\boldsymbol{\Pi}}\Big|\sum_{t=0}^{\infty}\gamma^t\sum_{i=1}^{N}\int_{(\frac{i-1}{N},\frac{i}{N}]}\Big(\mathbb{E}\big[\widehat{r}^i(s_t^i,\ \mu_t^{\alpha,W},\ \pi^i)\big] - \mathbb{E}\big[\widehat{r}^\alpha(s_t^i,\ \mu_t^{\alpha,W},\ \pi^i)\big]\Big)d\alpha\Big| \\
&\le \sup_{\boldsymbol{\pi}} L_r\sum_{t=0}^{\infty}\gamma^t\sum_{i=1}^{N}\int_{(\frac{i-1}{N},\frac{i}{N}]}\mathbb{E}\|\mu_t^{i,W_N} - \mu_t^{\alpha,W}\|_1 d\alpha + \sup_{\boldsymbol{\pi}}\tilde{L}_r\sum_{t=0}^{\infty}\gamma^t\sum_{i=1}^{N}\int_{(\frac{i-1}{N},\frac{i}{N}]}\Big|\frac{i}{N} - \alpha\Big|d\alpha \\
&= \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big),
\end{aligned}
\tag{4.6}
$$

361 where the last equality is from Lemma 4.1 when the convergence in Assumption 2.1 is at
362 rate $\mathcal{O}(1/\sqrt{N})$.

363 *Second term of* (4.5). From Lemma 4.2, we have $II = \mathcal{O}(\frac{1}{\sqrt{N}})$.

Third term of (4.5).

$$
\begin{aligned}
III &\leq \sup_{\boldsymbol{\pi}} L_r \sum_{t=0}^{\infty} \gamma^t \sum_{i=1}^{N} \int_{(\frac{i-1}{N}, \frac{i}{N}]} \max_{s \in \mathcal{S}} \|\pi^i(s) - \pi^\alpha(s)\|_1 d\alpha \\
&\leq L_r \tilde{L}_\Pi \sup_{\boldsymbol{\pi}} \sum_{t=0}^{\infty} \gamma^t \sum_{i=1}^{N} \int_{(\frac{i-1}{N}, \frac{i}{N}]} |\frac{i}{N} - \alpha| d\alpha \\
&= \mathcal{O}(\frac{1}{N}).
\end{aligned}
$$

Therefore, combining I, II and III yields the desired result. □

4.2. Proof of Theorem 3.8

First, we see that (3.11) corresponds to the following optimal control problem

$$
\begin{aligned}
\widetilde{V}^M(\tilde{\boldsymbol{\mu}}) &:= \sup_{\tilde{\boldsymbol{\pi}} \in \widetilde{\boldsymbol{\Pi}}_M} \tilde{J}^M(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\pi}}) \\
&= \sup_{\tilde{\boldsymbol{\pi}} \in \widetilde{\boldsymbol{\Pi}}_M} \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r\big(\tilde{s}_t^{\alpha_m}, \tilde{\mu}_t^{\alpha_m, W}, \tilde{a}_t^{\alpha_m}\big) \middle| \tilde{s}_0^{\alpha_m} \sim \tilde{\mu}^{\alpha_m}, \tilde{a}_t^{\alpha_m} \sim \tilde{\pi}^{\alpha_m}(\cdot | \tilde{s}_t^{\alpha_m}) \right]
\end{aligned} \tag{4.7}
$$

The associated Q function of (4.7) is defined as

$$
\begin{aligned}
\tilde{Q}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\pi}}) &= \sup_{\tilde{\boldsymbol{\pi}}'} \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r\big(\tilde{s}_t^{\alpha_m}, \tilde{\mu}_t^{\alpha_m, W}, \tilde{a}_t^{\alpha_m}\big) \middle| \tilde{s}_0^{\alpha_m} \sim \tilde{\mu}^{\alpha_m}, \tilde{a}_0^{\alpha_m} \sim \tilde{\pi}^{\alpha_m}(\cdot | \tilde{s}_t^{\alpha_m}) \right] \\
&= R(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\pi}}) + \sup_{\tilde{\boldsymbol{\pi}}' \in \widetilde{\boldsymbol{\Pi}}_M} \sum_{t=1}^{\infty} \gamma^t \tilde{R}(\tilde{\boldsymbol{\mu}}_t, \tilde{\boldsymbol{\pi}}'),
\end{aligned} \tag{4.8}
$$

subject to $\tilde{\boldsymbol{\mu}}_{t+1} = \widetilde{\boldsymbol{\Phi}}(\tilde{\boldsymbol{\mu}}_t, \tilde{\boldsymbol{\pi}})$, $\tilde{\boldsymbol{\mu}}_0 = \tilde{\boldsymbol{\mu}}$.

We first show the verification result and then prove the continuity property of $\tilde{Q}$ in (4.8), which thus leads to Theorem 3.8.

Lemma 4.3 (Verification) *Assume Assumption 3.5. Then $\tilde{Q}$ in (4.8) is the unique function satisfying the Bellman equation (3.11). Furthermore, if there exists $\tilde{\boldsymbol{\pi}}^* \in \arg\max_{\widetilde{\boldsymbol{\Pi}}_M} \tilde{Q}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\pi}})$ for each $\tilde{\boldsymbol{\mu}} \in \widetilde{\boldsymbol{\mathcal{M}}}_M$, then $\tilde{\boldsymbol{\pi}}^* \in \widetilde{\boldsymbol{\Pi}}_M$ is an optimal stationary policy ensemble.*

The proof of Lemma 4.3 is standard and very similar to the proof of Proposition 3.3 in [23].

Proof of Lemma 4.3 First, define $\frac{M_r}{1-\gamma}$-bounded function space $\mathcal{Q} := \{f : \widetilde{\boldsymbol{\mathcal{M}}}_M \times \widetilde{\boldsymbol{\Pi}}_M \to [-\frac{M_r}{1-\gamma}, \frac{M_r}{1-\gamma}]\}$. Then we define a Bellman operator $B : \mathcal{Q} \to \mathcal{Q}$

$$
(Bq)(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\pi}}) := \widetilde{R}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\pi}}) + \gamma \sup_{\tilde{\boldsymbol{\pi}}' \in \widetilde{\boldsymbol{\Pi}}_M} q(\widetilde{\boldsymbol{\Phi}}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\pi}}), \tilde{\boldsymbol{\pi}}'),
$$

One can show that B is a contraction operator with the module-γ. By Banach fixed point theorem, B admits a unique fixed point. As $\tilde{Q}$ function of (4.8) satisfies $B\tilde{Q} = \tilde{Q}$, $\tilde{Q}$ is unique solution of (3.11).

We next define $B^{\tilde{\boldsymbol{\pi}}'} : \mathcal{Q} \to \mathcal{Q}$ under the policy ensemble $\tilde{\boldsymbol{\pi}}' \in \widetilde{\boldsymbol{\Pi}}_M$ with

$$(B^{\tilde{\boldsymbol{\pi}}'} q)(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\pi}}) := \widetilde{R}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\pi}}) + \gamma q(\widetilde{\boldsymbol{\Phi}}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\pi}}), \tilde{\boldsymbol{\pi}}').$$

Similarly, we can show that $B^{\tilde{\boldsymbol{\pi}}'}$ is a contraction map with the module-γ and thus admits a unique fixed point, which is denoted as $\tilde{Q}^{\tilde{\boldsymbol{\pi}}'}$. From this, we have

$$\begin{aligned} \tilde{Q}^{\tilde{\boldsymbol{\pi}}^*}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\pi}}) &= \widetilde{R}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\pi}}) + \gamma \tilde{Q}^{\tilde{\boldsymbol{\pi}}^*}(\tilde{\boldsymbol{\Phi}}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\pi}}), \tilde{\boldsymbol{\pi}}^*) \\ &= \widetilde{R}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\pi}}) + \gamma \sup_{\tilde{\boldsymbol{\pi}}' \in \widetilde{\boldsymbol{\Pi}}_M} \tilde{Q}(\widetilde{\boldsymbol{\Phi}}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\pi}}), \tilde{\boldsymbol{\pi}}') = \tilde{Q}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\pi}}), \end{aligned}$$

which implies $\tilde{\boldsymbol{\pi}}^*$ is an optimal policy ensemble. □

Lemma 4.4 *Let Assumptions 3.4, 3.5 hold. Assume further $\gamma \cdot (1 + L_P + L_\Pi) < 1$. Then $\tilde{Q}$ in (4.8) is continuous.*

Proof of Lemma 4.4 We will show that as $\tilde{\boldsymbol{\mu}}_n \to \tilde{\boldsymbol{\mu}}$, $\tilde{\pi}_n \to \tilde{\boldsymbol{\pi}}$ in the sense that $\frac{1}{M}\sum_{m=0}^{M-1} \|\tilde{\mu}^{\alpha_m} - \tilde{\mu}_n^{\alpha_m}\|_1 + \frac{1}{M}\sum_{m=0}^{M-1} \max_{s\in\mathcal{S}} \|\tilde{\pi}^{\alpha_m}(\tilde{\mu}^{\alpha_m,W}) - \tilde{\pi}_n^{\alpha_m}(\tilde{\mu}_n^{\alpha_m,W})\|_1 \to 0$,

$$\tilde{Q}(\tilde{\boldsymbol{\mu}}_n, \tilde{\boldsymbol{\pi}}_n(\tilde{\boldsymbol{\mu}}_n)) \to \tilde{Q}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\pi}}(\tilde{\boldsymbol{\mu}})).$$

From (4.8) and (3.13),

$$\begin{aligned} & |\tilde{Q}(\tilde{\boldsymbol{\mu}}_n, \tilde{\boldsymbol{\pi}}_n) - \tilde{Q}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\pi}})| \\ \le & \left| \widetilde{R}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\pi}}(\tilde{\boldsymbol{\mu}})) + \sup_{\tilde{\boldsymbol{\pi}}' \in \widetilde{\boldsymbol{\Pi}}_M} \sum_{t=1}^{\infty} \gamma^t \widetilde{R}(\tilde{\boldsymbol{\mu}}_t, \tilde{\boldsymbol{\pi}}'(\tilde{\boldsymbol{\mu}}_t)) - \widetilde{R}(\tilde{\boldsymbol{\mu}}_n, \tilde{\boldsymbol{\pi}}_n(\tilde{\boldsymbol{\mu}}_n)) + \sup_{\tilde{\boldsymbol{\pi}}' \in \widetilde{\boldsymbol{\Pi}}_M} \sum_{t=1}^{\infty} \gamma^t \widetilde{R}(\tilde{\boldsymbol{\mu}}_{n,t}, \tilde{\boldsymbol{\pi}}'(\tilde{\boldsymbol{\mu}}_{n,t})) \right| \\ \le & \left| \widetilde{R}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\pi}}(\tilde{\boldsymbol{\mu}})) - \widetilde{R}(\tilde{\boldsymbol{\mu}}_n, \tilde{\pi}_n(\tilde{\boldsymbol{\mu}}_n)) \right| + \sup_{\tilde{\boldsymbol{\pi}}' \in \widetilde{\boldsymbol{\Pi}}_M} \sum_{t=1}^{\infty} \gamma^t \left| \widetilde{R}(\tilde{\boldsymbol{\mu}}_{n,t}, \tilde{\boldsymbol{\pi}}'(\tilde{\boldsymbol{\mu}}_{n,t})) - \widetilde{R}(\tilde{\boldsymbol{\mu}}_t, \tilde{\boldsymbol{\pi}}'(\tilde{\boldsymbol{\mu}}_t)) \right| \\ \le & L_r \cdot \frac{1}{M} \sum_{m=0}^{M-1} \|\tilde{\mu}^{\alpha_m,W} - \tilde{\mu}_n^{\alpha_m,W}\|_1 + M_r \cdot \frac{1}{M} \sum_{m=0}^{M-1} \|\tilde{\mu}^{\alpha_m} - \tilde{\mu}_n^{\alpha_m}\|_1 d\alpha \\ & + M_r \cdot \frac{1}{M} \sum_{m=0}^{M-1} \max_{s\in\mathcal{S}} \|\tilde{\pi}^{\alpha}(\tilde{\mu}^{\alpha_m,W}) - \tilde{\pi}_n^{\alpha}(\tilde{\mu}_n^{\alpha_m,W})\|_1 d\alpha \\ & + \sup_{\tilde{\boldsymbol{\pi}}' \in \widetilde{\boldsymbol{\Pi}}_M} \sum_{t=1}^{\infty} \gamma^t \cdot \Big((L_r + L_\Pi) \cdot \frac{1}{M} \sum_{m=0}^{M-1} \|\tilde{\mu}_t^{\alpha_m,W} - \tilde{\mu}_{n,t}^{\alpha_m,W}\|_1 + M_r \cdot \frac{1}{M} \sum_{m=0}^{M-1} \|\tilde{\mu}_t^{\alpha_m} - \tilde{\mu}_{n,t}^{\alpha_m}\|_1 \Big) \\ \le & (L_r + M_r) \cdot \frac{1}{M} \sum_{m=0}^{M-1} \|\tilde{\mu}^{\alpha_m} - \tilde{\mu}_n^{\alpha_m}\|_1 + M_r \cdot \frac{1}{M} \sum_{m=0}^{M-1} \max_{s\in\mathcal{S}} \|\tilde{\pi}^{\alpha}(\tilde{\mu}^{\alpha_m,W}) - \tilde{\pi}_n^{\alpha}(\tilde{\mu}_n^{\alpha_m,W})\|_1 d\alpha \\ & + \sup_{\tilde{\boldsymbol{\pi}}' \in \widetilde{\boldsymbol{\Pi}}_M} \sum_{t=1}^{\infty} \gamma^t \cdot (L_r + L_\Pi + M_r) \cdot \frac{1}{M} \sum_{m=0}^{M-1} \|\tilde{\mu}_t^{\alpha_m} - \tilde{\mu}_{n,t}^{\alpha_m}\|_1. \end{aligned}$$

By induction, we obtain

$$
\begin{aligned}
&\frac{1}{M}\sum_{m=0}^{M-1}\|\tilde{\mu}_t^{\alpha_m}-\tilde{\mu}_{n,t}^{\alpha_m}\|_1\\
=\;&\frac{1}{M}\sum_{m=0}^{M-1}\sum_{s'\in\mathcal{S}}\Big|\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}P^{\alpha_m}(s'|s,a,\tilde{\mu}_{t-1}^{\alpha_m,W})\tilde{\mu}_{t-1}^{\alpha_m}(s)\tilde{\pi}(\alpha|s,\tilde{\mu}_{t-1}^{\alpha_m,W})\\
&-\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}P^{\alpha_m}(s'|s,a,\tilde{\mu}_{n,t-1}^{\alpha_m,W})\tilde{\mu}_{n,t-1}^{\alpha_m}(s)\tilde{\pi}(\alpha|s,\tilde{\mu}_{n,t-1}^{\alpha_m,W})\Big|\\
\le\;&(L_P+L_\Pi+1)\cdot\frac{1}{M}\sum_{m=0}^{M-1}\|\tilde{\mu}_{t-1}^{\alpha}-\tilde{\mu}_{n,t-1}^{\alpha}\|_1\\
\le\;&\ldots\le(L_P+L_\Pi+1)^{(t-1)}\frac{1}{M}\sum_{m=0}^{M-1}\|\tilde{\mu}_1^{\alpha}-\tilde{\mu}_{n,1}^{\alpha}\|_1.
\end{aligned}
$$

Therefore, if $\gamma\cdot(1+L_P+L_\Pi)<1$, then

$$
|\tilde{Q}(\tilde{\boldsymbol{\mu}}_n,\tilde{\boldsymbol{\pi}}_n)-\tilde{Q}(\tilde{\boldsymbol{\mu}},\tilde{\boldsymbol{\pi}})|\le C\Big(\frac{1}{M}\sum_{m=0}^{M-1}\|\tilde{\mu}^{\alpha_m}-\tilde{\mu}_n^{\alpha_m}\|_1+\frac{1}{M}\sum_{m=0}^{M-1}\max_{s\in\mathcal{S}}\|\tilde{\pi}^{\alpha_m}(\tilde{\mu}^{\alpha_m,W})-\tilde{\pi}_n^{\alpha_m}(\tilde{\mu}_n^{\alpha_m,W})\|_1\Big).
$$

where C is a constant depending on L_r, M_r, L_P, L_Π. □

Now we prove Theorem 3.8.

Proof of Theorem 3.8 By Lemma 4.4, along with the compactness of $\widetilde{\boldsymbol{\Pi}}_M$, there exists $\tilde{\boldsymbol{\pi}}^*\in\widetilde{\boldsymbol{\Pi}}_M$ such that $\tilde{\boldsymbol{\pi}}^*\in\arg\max_{\tilde{\boldsymbol{\pi}}\in\widetilde{\boldsymbol{\Pi}}_M}Q(\tilde{\boldsymbol{\mu}},\tilde{\boldsymbol{\pi}})$. By Lemma 4.3, there exists an optimal policy ensemble $\tilde{\boldsymbol{\pi}}^*\in\widetilde{\boldsymbol{\Pi}}_M$. □

4.3. Proof of Theorem 3.9

We first prove the following Lemma, which shows that GMFC and block GMFC become increasingly close to each other as the number of blocks becomes larger.

Lemma 4.5 *Under Assumptions 3.3, 3.4 and 3.6, we have*

$$
\sum_{m=1}^{M}\int_{(\frac{m-1}{M},\frac{m}{M}]}\|\mu_t^{\alpha,W}-\tilde{\mu}_t^{\alpha_m,W}\|_1 d\alpha\le\Big[(1+L_P+L_\Pi)^t-1\Big]\frac{\tilde{L}_\Pi+\tilde{L}_P+2(L_P+L_\Pi)L_W}{M}+\frac{2L_W}{M},
$$

$$
\sum_{m=1}^{M}\int_{(\frac{m-1}{M},\frac{m}{M}]}\|\mu_t^{\alpha}-\tilde{\mu}_t^{\alpha_m}\|_1 d\alpha\le\Big[(1+L_P+L_\Pi)^t-1\Big]\frac{\tilde{L}_\Pi+\tilde{L}_P+2(L_P+L_\Pi)L_W}{M}.
$$

Proof of Lemma 4.5

$$\begin{aligned}
&\sum_{m=1}^{M}\int_{(\frac{m-1}{M},\frac{m}{M}]}\|\mu_t^{\alpha,W}-\tilde{\mu}_t^{\alpha_m,W}\|_1 d\alpha \\
\leq\ & \sum_{m=1}^{M}\int_{(\frac{m-1}{M},\frac{m}{M}]}\|\mu_t^{\alpha,W}-\mu_t^{\alpha_m,W}\|_1 d\alpha+\frac{1}{M}\sum_{m=1}^{M}\|\mu_t^{\alpha_m,W}-\bar{\mu}_t^{\alpha_m,W}\|_1 \\
&+\frac{1}{M}\sum_{m=1}^{M}\|\bar{\mu}_t^{\alpha_m,W}-\tilde{\mu}_t^{\alpha_m,W}\|_1,
\end{aligned} \tag{4.9}$$

where $\bar{\mu}^{\alpha_m,W}:=\frac{1}{M}\sum_{m'=1}^{M}W(\alpha_m,\alpha_{m'})\mu^{\alpha_{m'}}$.

By the definition of $\mu_t^{\alpha,W},\mu_t^{\alpha_m,W}$ in (2.10), $\tilde{\mu}_t^{\alpha_m,W}$ in (3.12) and $\bar{\mu}^{\alpha_m,W}$, together with the Lipschitz continuity of W in Assumption 3.3,

$$\sum_{m=1}^{M}\int_{(\frac{m-1}{M},\frac{m}{M}]}\|\mu_t^{\alpha,W}-\mu_t^{\alpha_m,W}\|_1 d\alpha \leq \sum_{m=1}^{M}\int_{(\frac{m-1}{M},\frac{m}{M}]}\|\mu_t^{\alpha}-\mu_t^{\alpha_m}\|_1 d\alpha+\frac{L_W}{M}, \tag{4.10}$$

$$\frac{1}{M}\sum_{m=1}^{M}\|\mu_t^{\alpha_m,W}-\bar{\mu}_t^{\alpha_m,W}\|_1 \leq \frac{L_W}{M}, \tag{4.11}$$

$$\frac{1}{M}\sum_{m=1}^{M}\|\bar{\mu}_t^{\alpha_m,W}-\tilde{\mu}_t^{\alpha_m,W}\|_1 \leq \frac{1}{M}\sum_{m=1}^{M}\|\mu_t^{\alpha_m}-\tilde{\mu}_t^{\alpha_m}\|_1. \tag{4.12}$$

Plugging these into (4.9),

$$\sum_{m=1}^{M}\int_{(\frac{m-1}{M},\frac{m}{M}]}\|\mu_t^{\alpha,W}-\tilde{\mu}_t^{\alpha_m,W}\|_1 d\alpha \leq A_t+\frac{2L_W}{M}, \tag{4.13}$$

where $A_t:=\sum_{m=1}^{M}\int_{(\frac{m-1}{M},\frac{m}{M}]}\|\mu_t^{\alpha}-\mu_t^{\alpha_m}\|_1 d\alpha+\frac{1}{M}\sum_{m=1}^{M}\|\mu_t^{\alpha_m}-\tilde{\mu}_t^{\alpha_m}\|_1$.

On the other hand,

$$\sum_{m=1}^{M}\int_{(\frac{m-1}{M},\frac{m}{M}]}\|\mu_t^{\alpha}-\tilde{\mu}_t^{\alpha_m}\|_1 d\alpha \tag{4.14}$$

$$\leq \sum_{m=1}^{M}\int_{(\frac{m-1}{M},\frac{m}{M}]}\|\mu_t^{\alpha}-\mu_t^{\alpha_m}\|_1 d\alpha+\frac{1}{M}\sum_{m=1}^{M}\|\mu_t^{\alpha_m}-\tilde{\mu}_t^{\alpha_m}\|_1 = A_t. \tag{4.15}$$

Therefore, it is enough to estimate A_t. We next estimate A_{t+1} by an inductive way. Note that $A_0 = 0$.

$$
\begin{aligned}
& A_{t+1} \\
= & \sum_{m=1}^{M} \int_{(\frac{m-1}{M}, \frac{m}{M}]} \|\mu_{t+1}^{\alpha} - \mu_{t+1}^{\alpha_m}\|_1 d\alpha + \frac{1}{M} \sum_{m=1}^{M} \|\mu_{t+1}^{\alpha_m} - \tilde{\mu}_{t+1}^{\alpha_m}\|_1 \\
= & \sum_{m=1}^{M} \int_{(\frac{m-1}{M}, \frac{m}{M}]} \Big\| \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \Big(P^{\alpha}(\cdot|s, \mu_t^{\alpha,W}, a) \pi^{\alpha}(a|s, \mu_t^{\alpha,W}) \mu_t^{\alpha}(s) \\
& - P^{\alpha_m}(\cdot|s, a, \mu_t^{\alpha_m,W}) \mu_t^{\alpha_m}(s) \pi^{\alpha_m}(a|s, \mu_t^{\alpha_m,W}) \Big) \Big\|_1 d\alpha \\
& + \frac{1}{M} \sum_{m=1}^{M} \Big\| \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \Big(P^{\alpha_m}(\cdot|s, \mu_t^{\alpha_m,W}, a) \pi^{\alpha_m}(a|s, \mu_t^{\alpha_m,W}) \mu_t^{\alpha_m}(s) \\
& - P^{\alpha_m}(\cdot|s, a, \tilde{\mu}_t^{\alpha_m,W}) \tilde{\mu}_t^{\alpha_m}(s) \tilde{\pi}^{\alpha_m}(a|s, \tilde{\mu}_t^{\alpha_m,W}) \Big) \Big\|_1 \\
\leq & \sum_{m=1}^{M} \int_{(\frac{m-1}{M}, \frac{m}{M}]} \Big((L_P + L_\Pi) \cdot \|\mu_t^{\alpha,W} - \mu^{\alpha_m,W}\|_1 + \frac{\tilde{L}_\Pi}{M} + \|\mu_t^{\alpha} - \mu_t^{\alpha_m}\|_1 \Big) d\alpha \\
& + \frac{1}{M} \sum_{m=1}^{M} \Big((L_P + L_\Pi) \cdot \|\mu_t^{\alpha_m,W} - \tilde{\mu}^{\alpha_m,W}\|_1 + \frac{\tilde{L}_P}{M} + \|\mu_t^{\alpha_m} - \tilde{\mu}_t^{\alpha_m}\|_1 \Big) \\
\leq & (1 + L_P + L_\Pi) A_t + (\tilde{L}_\Pi + \tilde{L}_P + 2(L_P + L_\Pi) L_W) \frac{1}{M},
\end{aligned}
$$

where the second equality is from (3.4) and (3.14), and we use Assumptions 3.3, 3.4 and 3.6 in the third inequality, and we use (4.10)-(4.12) in the last inequality.
By induction, we have

$$
A_{t+1} \leq \Big[(1 + L_P + L_\Pi)^t - 1 \Big] \frac{\tilde{L}_\Pi + \tilde{L}_P + 2(L_P + L_\Pi) L_W}{M}.
$$

□

Based on Lemma 4.5, we have the following Proposition.

Proposition 4.6 *Assume Assumptions 3.3, 3.4, 3.5, 3.6, and $\gamma \cdot (L_P + L_\Pi + 1) < 1$. Then we have for any $\mu \in \mathcal{P}(\mathcal{S})$*

$$
\sup_{\boldsymbol{\pi}\in\boldsymbol{\Pi}} \left| \tilde{J}^M(\mu, \boldsymbol{\pi}) - J(\mu, \boldsymbol{\pi}) \right| \to 0, \quad \text{as } M \to +\infty, \tag{4.16}
$$

where $\tilde{J}^M$ and J are given in (4.7) and (2.11), respectively.

Proof of Proposition 4.6 Recall from (3.12) that

$$
\tilde{J}^M(\mu, \tilde{\boldsymbol{\pi}}) = \sum_{t=0}^{\infty} \gamma^t \widetilde{R}(\tilde{\boldsymbol{\mu}}_t, \tilde{\boldsymbol{\pi}}(\tilde{\boldsymbol{\mu}}_t)),
$$

subject to $\tilde{\mu}^{\alpha_m}_{t+1} = \widetilde{\mathbf{\Phi}}^{\alpha_m}(\tilde{\mu}^{\alpha_m}_t, \tilde{\pi}^{\alpha_m})$, $t \in \mathbb{N}_+$, $\tilde{\mu}^{\alpha}_0 \equiv \mu$, and $\tilde{\mu}^{\alpha_m,W}_t$ given in (3.12).

$$J(\mu, \boldsymbol{\pi}) \quad = \quad \sum_{t=0}^{\infty} \gamma^t R(\boldsymbol{\mu}_t, \boldsymbol{\pi}(\boldsymbol{\mu}_t)),$$

subject to $\mu^{\alpha}_{t+1} = \mathbf{\Phi}^{\alpha}(\mu^{\alpha}_t, \pi^{\alpha}), t \in \mathbb{N}_+$, $\mu^{\alpha}_0 \equiv \mu$, and $\mu^{\alpha,W}_t$ given in (2.10). Since $\tilde{\boldsymbol{\pi}} := (\tilde{\pi}^{\alpha_m})_{m\in[M]} \in \widetilde{\mathbf{\Pi}}_M$ can be viewed as a piecewise-constant projection of $\boldsymbol{\pi} \in \mathbf{\Pi}$ onto $\widetilde{\mathbf{\Pi}}_M$. Then,

$$\begin{aligned}
& \sup_{\boldsymbol{\pi}\in\mathbf{\Pi}} \left| \tilde{J}^M(\mu, \boldsymbol{\pi}) - J(\mu, \boldsymbol{\pi}) \right| \\
\leq \;& \sup_{\boldsymbol{\pi}\in\mathbf{\Pi}} \sum_{t=0}^{\infty} \gamma^t \left| \tilde{R}(\tilde{\boldsymbol{\mu}}_t, \tilde{\boldsymbol{\pi}}(\tilde{\boldsymbol{\mu}}_t)) - R(\boldsymbol{\mu}_t, \boldsymbol{\pi}(\boldsymbol{\mu}_t)) \right| \\
\leq \;& \sup_{\boldsymbol{\pi}\in\mathbf{\Pi}} \sum_{t=0}^{\infty} \gamma^t \left| \tilde{R}(\tilde{\boldsymbol{\mu}}_t, \tilde{\boldsymbol{\pi}}(\tilde{\boldsymbol{\mu}}_t)) - R(\boldsymbol{\mu}_t, \tilde{\boldsymbol{\pi}}(\boldsymbol{\mu}_t)) \right| + \sup_{\boldsymbol{\pi}\in\mathbf{\Pi}} \sum_{t=0}^{\infty} \gamma^t \left| R(\boldsymbol{\mu}_t, \tilde{\boldsymbol{\pi}}(\boldsymbol{\mu}_t)) - R(\boldsymbol{\mu}_t, \boldsymbol{\pi}(\boldsymbol{\mu}_t)) \right| \\
:= \;& I + II.
\end{aligned}$$

In terms of the term I, we first estimate $\left| \tilde{R}(\tilde{\boldsymbol{\mu}}_t, \tilde{\boldsymbol{\pi}}) - R(\boldsymbol{\mu}_t, \tilde{\boldsymbol{\pi}}) \right|$:

$$\begin{aligned}
& \left| \tilde{R}(\tilde{\boldsymbol{\mu}}_t, \tilde{\boldsymbol{\pi}}(\tilde{\boldsymbol{\mu}}_t)) - R(\boldsymbol{\mu}_t, \tilde{\boldsymbol{\pi}}(\boldsymbol{\mu}_t)) \right| \\
= \;& \Bigg| \sum_{m=1}^{M} \int_{(\frac{m-1}{M}, \frac{m}{M}]} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} r^{\alpha_m}(s, a, \tilde{\mu}^{\alpha_m,W}_t) \tilde{\mu}^{\alpha_m}_t(s) \tilde{\pi}^{\alpha_m}(a|s, \tilde{\mu}^{\alpha_m,W}_t) d\alpha \\
& - \sum_{m=1}^{M} \int_{(\frac{m-1}{M}, \frac{m}{M}]} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} r^{\alpha}(s, a, \mu^{\alpha,W}_t) \mu^{\alpha}_t(s) \tilde{\pi}^{\alpha_m}(a|s, \mu^{\alpha,W}_t) d\alpha \Bigg| \\
\leq \;& \Bigg| \sum_{m=1}^{M} \int_{(\frac{m-1}{M}, \frac{m}{M}]} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \big(r^{\alpha_m}(s, a, \tilde{\mu}^{\alpha_m,W}_t) - r^{\alpha}(s, a, \mu^{\alpha,W}_t) \big) \tilde{\mu}^{\alpha_m}_t(s) \tilde{\pi}^{\alpha_m}(a|s, \tilde{\mu}^{\alpha_m,W}_t) d\alpha \\
& + \Bigg| \sum_{m=1}^{M} \int_{(\frac{m-1}{M}, \frac{m}{M}]} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} r^{\alpha}(s, a, \mu^{\alpha,W}_t) \big(\tilde{\mu}^{\alpha_m}_t(s) - \mu^{\alpha}_t(s) \big) \tilde{\pi}^{\alpha_m}(a|s, \tilde{\mu}^{\alpha_m,W}_t) d\alpha \Bigg| \\
& + \Bigg| \sum_{m=1}^{M} \int_{(\frac{m-1}{M}, \frac{m}{M}]} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} r^{\alpha}(s, a, \mu^{\alpha,W}_t) \tilde{\mu}^{\alpha}_t(s) \big(\tilde{\pi}^{\alpha_m}(a|s, \tilde{\mu}^{\alpha_m,W}_t) - \tilde{\pi}^{\alpha_m}(a|s, \mu^{\alpha,W}_t) \big) d\alpha \Bigg| \\
\leq \;& L_r \cdot \sum_{m=1}^{M} \int_{(\frac{m-1}{M}, \frac{m}{M}]} \|\mu^{\alpha,W}_t - \tilde{\mu}^{\alpha_m,W}_t\|_1 d\alpha + \frac{\tilde{L}_r}{M} \\
& + M_r \cdot \sum_{m=1}^{M} \int_{(\frac{m-1}{M}, \frac{m}{M}]} \|\mu^{\alpha}_t - \tilde{\mu}^{\alpha_m}_t\|_1 d\alpha + M_r L_{\Pi} \cdot \sum_{m=1}^{M} \int_{(\frac{m-1}{M}, \frac{m}{M}]} \|\mu^{\alpha,W}_t - \tilde{\mu}^{\alpha_m,W}_t\|_1 d\alpha.
\end{aligned}$$

By Lemma 4.5,

$$I \leq \frac{C(\gamma, L_{\Pi}, L_P, L_W, L_r, M_r)}{M}.$$

For the term II,

$$
\begin{aligned}
& \sup_{\boldsymbol{\pi}\in\boldsymbol{\Pi}} \sum_{t=0}^{\infty} \gamma^t \Big| R(\boldsymbol{\mu}_t, \tilde{\boldsymbol{\pi}}(\boldsymbol{\mu}_t)) - R(\boldsymbol{\mu}_t, \boldsymbol{\pi}(\boldsymbol{\mu}_t)) \Big| \\
\leq \; & \sup_{\boldsymbol{\pi}\in\boldsymbol{\Pi}} \sum_{t=0}^{\infty} \gamma^t M_r \sum_{m=1}^{M} \int_{(\frac{m-1}{M}, \frac{m}{M}]} \max_{s\in\mathcal{S}} \|\pi^{\alpha} - \pi^{\alpha^m}\|_1 d\alpha \\
& + \sup_{\boldsymbol{\pi}\in\boldsymbol{\Pi}} \sum_{t=0}^{\infty} \gamma^t M_r \sum_{m=1}^{M} \int_{(\frac{m-1}{M}, \frac{m}{M}]} \|\mu_t^{\alpha,W} - \tilde{\mu}_t^{\alpha_m,W}\|_1 d\alpha \\
\leq \; & \frac{C(\gamma, L_{\Pi}, L_P, L_W, L_r, M_r)}{M}.
\end{aligned}
$$

□

Proof of Theorem 3.9 Suppose that $\tilde{\boldsymbol{\pi}}^* \in \widetilde{\boldsymbol{\Pi}}_M \subset \boldsymbol{\Pi}$ and $(\pi^{1,*}, \dots, \pi^{N,*}) \in \Pi^N$ are optimal policies of the problems (4.7) and (2.7), respectively. From Proposition 4.6, for any $\varepsilon > 0$, there exists sufficiently large $M_\varepsilon > 0$

$$
|\tilde{J}^{M_\varepsilon}(\mu, \tilde{\boldsymbol{\pi}}^*) - J(\mu, \tilde{\boldsymbol{\pi}}^*)| \leq \frac{\varepsilon}{3},
$$

where by (3.8), $\boldsymbol{\pi}^{N,*} := \sum_{i=1}^{N} \pi^{i,*} \mathbf{1}_{\alpha\in(\frac{i-1}{N}, \frac{i}{N}]}$.
From Theorem 3.7, for any $\varepsilon > 0$, there exists N_ε such that for all $N \geq N_\varepsilon$

$$
|J_N(\mu, \tilde{\pi}^{1,*}, \dots, \tilde{\pi}^{N,*}) - J(\mu, \tilde{\boldsymbol{\pi}}^*)| \leq \frac{\varepsilon}{3}, \quad |J_N(\mu, \pi^{1,*}, \dots, \pi^{N,*}) - J(\mu, \boldsymbol{\pi}^{N,*})| \leq \frac{\varepsilon}{3}.
$$

Then we have

$$
\begin{aligned}
& J_N(\mu, \tilde{\pi}^{1,*}, \dots, \tilde{\pi}^{N,*}) - J_N(\mu, \pi^{1,*}, \dots, \pi^{N,*}) \\
\geq \; & \underbrace{J_N(\mu, \tilde{\pi}^{1,*}, \dots, \tilde{\pi}^{N,*}) - J(\mu, \tilde{\boldsymbol{\pi}}^*)}_{I_1} + \underbrace{J(\mu, \tilde{\boldsymbol{\pi}}^*) - \tilde{J}_{M_\varepsilon}(\mu, \tilde{\boldsymbol{\pi}}^*)}_{I_2} \\
& + \underbrace{\tilde{J}^{M_\varepsilon}(\mu, \tilde{\boldsymbol{\pi}}^*) - \tilde{J}^{M_\varepsilon}(\mu, \boldsymbol{\pi}^{N,*})}_{I_3} + \underbrace{\tilde{J}^{M_\varepsilon}(\mu, \boldsymbol{\pi}^{N,*}) - J_N(\mu, \pi^{1,*}, \dots, \pi^{N,*})}_{I_4} \\
\geq \; & -\frac{\varepsilon}{3} - \frac{\varepsilon}{3} - \frac{\varepsilon}{3} = -\varepsilon.
\end{aligned}
$$

where $I_3 \geq 0$ due to the optimality of $\tilde{\boldsymbol{\pi}}^*$ for $\tilde{V}^{M_\varepsilon}$. This means that the optimal policy of block GMFC provides an ε-optimal policy for the multi-agent system with $(\tilde{\pi}_1^*, \dots, \tilde{\pi}_N^*) := \Gamma_N(\tilde{\boldsymbol{\pi}}^*)$. □

5. Experiments

In this section, we provide an empirical verification of our theoretical results, with two examples adapted from existing works on learning MFGs [16, 10] and learning GMFGs [15].

5.1. SIS Graphon Model

We consider a SIS graphon model in [16] under a cooperative setting. In this model, each agent $\alpha \in \mathcal{I}$ shares a state space $\mathcal{S} = \{S, I\}$ and an action space $\mathcal{A} = \{C, NC\}$, where S is susceptible, I is infected, C represents keeping contact with others, and NC means keeping social distance. The transition probability of each agent α is represented as follows

$$\begin{aligned} P^\alpha(s_{t+1} = I | s_t = S, a_t = C, \mu_t^{\alpha,W}) &= \beta_1 \mu_t^{\alpha,W}(I), \\ P^\alpha(s_{t+1} = I | s_t = S, a_t = NC, \mu_t^{\alpha,W}) &= \beta_2 \mu_t^{\alpha,W}(I), \\ P^\alpha(s_{t+1} = S | s_t = I, \mu_t^{\alpha,W}) &= \delta, \end{aligned}$$

where β_1 is the infection rate with keeping contact with others, β_2 is the infection rate under social distance, and δ is the fixed recovery rate. We assume $0 < \beta_2 < \beta_1$, meaning that keeping social distance can reduce the risk of being infected. The individual reward function is defined as

$$r^\alpha(s, \mu_t^{\alpha,W}, a) = -c_1 \mathbf{1}_{\{I\}}(s) - c_2 \mathbf{1}_{\{NC\}}(a) - c_3 \mathbf{1}_{\{I\}}(s) \mathbf{1}_{\{C\}}(a),$$

where c_1 represents the cost of being infected such as the cost of medical treatment, c_2 represents the cost of keeping social distance, and c_3 represents the penalty of going out if the agent is infected.

In our experiment, we set β_1=0.8, β_2=0, $\delta = 0.3$ for the transition dynamics and c_1=2, c_2=0.3, $c_3 = 0.5$ for the reward function. The initial mean field μ_0 is taken as the uniform distribution. We set the episode length to 50.

5.2. Malware Spread Graphon Model

We consider a malware spread model in [10] under a cooperative setting. In this model, let $\mathcal{S} = \{0, 1, \ldots, K-1\}$, $K \in \mathbb{N}$, denote the health level of the agent, where $s_t = 0$ and $s_t = K-1$ represents the best level and the worst level, respectively. All agents can take two actions: $a_t = 0$ means doing nothing, and $a_t = 1$ means repairing. The state transition is given by

$$s_{t+1} = \begin{cases} s_t + \lfloor (K - s_t)\chi_t \rfloor, & \text{if } a_t = 0, \\ 0, & \text{if } a_t = 1, \end{cases}$$

where $\chi_t, t \in \mathbb{N}$ are i.i.d. random variables with a certain probability distribution. Then after taking action a_t, agent α will receive an individual reward

$$r^\alpha(s_t, \mu_t^{\alpha,W}, a_t) = -(c_1 + \langle \mu_t^{\alpha,W} \rangle) s_t / K - c_2 a_t.$$

Here considering the heterogeneity of agents, we use $W(\alpha, \beta)$ to denote the *importance* effect of agent β on agent α. $\langle \mu_t^{\alpha,W} \rangle := \int_{\beta \in \mathcal{I}} \sum_{s \in \mathcal{S}} s W(\alpha, \beta) \mu_t^\beta(s) d\beta$ is the risk of being infected by other agents and c_2 is the cost of taking action a_t.

In our experiment, we set K=3, c_1=0.3, and c_2=0.5. In addition, to stabilize the training of the RL agent, we fix χ_t to a static value, i.e., 0.7. In this model, we set the episode length to 10.

5.3. *Performance of N-agent GMFC on Multi-Agent System*

For both models, we use PPO [47] to train the block GMFC agent in the infinite-agent environment and obtain the policy ensembles and further use Algorithm 2 to deploy them in the finite-agent environment. We test the performance of N-agent GMFC with 10 blocks to different numbers of agents, i.e., from 10 to 100. For each case, we run 1000 times of simulations and show the mean and standard variation (Green shadows in Figure 1 and Figure 2) of the mean episode reward. We can see that in both scenarios and for different types of graphons, the mean episode rewards of the N-agent GMFC become increasingly close to that of block GMFC as the number of agents grows. (See Figure 1 and Figure 2). This verifies our theoretical findings empirically.

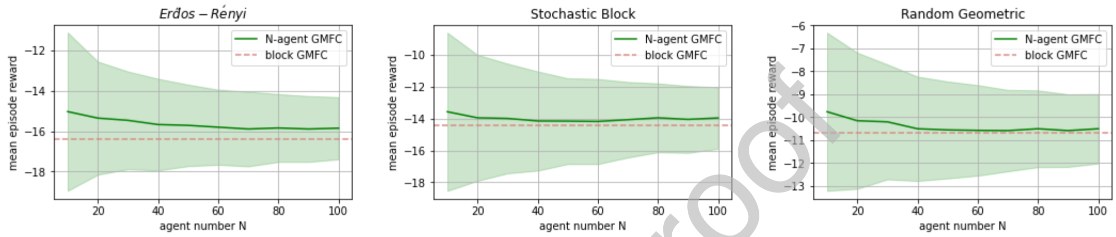


Figure 1: Experiments for different graphons in SIS finite-agent environment

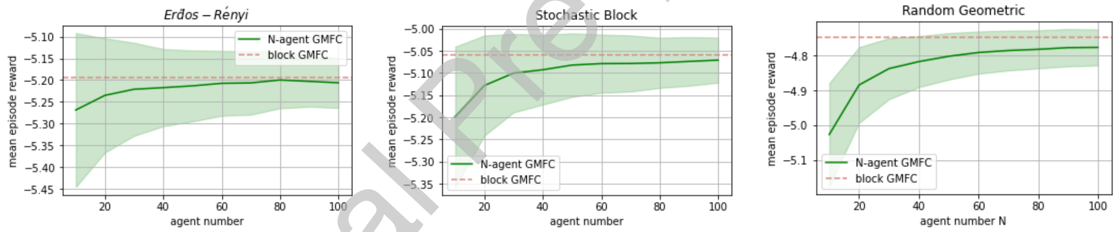


Figure 2: Experiments for different graphons in Malware Spread finite-agent environment

5.4. *Comparison with Other Algorithms*

For different types of graphons, we compare our algorithm N-agent GMFC with three existing MARL algorithms, including two independent learning algorithms, i.e., independent DQN [40], independent PPO [47] and a powerful centralized-training-and-decentralized-execution(CTDE)-based algorithm QMIX [46]. We test the performance of those algorithms with different numbers of blocks, i.e., 2, 5, 10, to the multi-agent systems with 40 agents. The results are reported in Table 1 and Table 2.

In the SIS graphon model, N-agent GMFC shows dominating performance in most cases and outperforms independent algorithms by a large margin. Only QMIX can reach comparable results. And in the malware spread graphon model, N-agent GMFC outperforms other algorithms in more than half of the cases. Only independent DQN has comparable performance in this environment. And we can see that in both environments, the performance gap between N-agent GMFC and other MARL algorithms is shrinking as the number of blocks goes larger. This is mainly because the action space of block GMFC increases more quickly than MARL algorithms as the block number increases. And it is hard to train RL agents when the action space is too large.

Beyond the visible results shown in Tables 1 and 2, when the number of agents N grows larger, classic MARL methods become infeasible because of the curse of dimensionality and the restriction of memory storage, while N-agent GMFC is trained only once and independent of the number of agents N, hence is easier to scale up in a large-scale regime and enjoys a more stable performance. We can see that N-agent GMFC shows more stable results when N increases as shown in Figure 1 and Figure 2.

Table 1: Mean Episode Reward for SIS with 40 agents

Graphon Type	M	Algorithm			
		N-agent GMFC	I-DQN	I-PPO	QMIX
Erdős Rényi	2	**-15.37**	-17.58	-20.63	-20.51
	5	**-15.74**	-16.17	-20.42	16.94
	10	-15.67	-17.55	-21.38	**-14.45**
Stochastic Block	2	**-13.58**	-16.05	-18.38	-17.69
	5	**-13.67**	-15.91	-20.13	-13.79
	10	**-13.57**	-15.52	-14.87	-13.86
Random Geometric	2	**-12.45**	-17.93	-14.82	-14.52
	5	**-9.82**	-12.81	-12.99	-10.84
	10	**-10.52**	-11.68	-12.66	-12.60

Table 2: Mean Episode Reward for Malware Spread with 40 agents

Graphon Type	M	Algorithm			
		N-agent GMFC	I-DQN	I-PPO	QMIX
Erdős Rényi	2	-5.21	**-5.11**	-5.31	-6.05
	5	**-5.21**	-5.30	-5.26	-6.13
	10	-5.21	**-5.14**	-5.27	-5.21
Stochastic Block	2	**-5.16**	-5.21	-5.37	-5.88
	5	**-5.10**	-5.19	-5.31	-5.70
	10	-5.09	**-5.05**	-5.28	-5.27
Random Geometric	2	**-5.02**	-5.21	-5.27	-5.35
	5	**-4.85**	-5.03	-5.04	-5.05
	10	**-4.82**	-4.83	-5.14	-4.83

5.5. Implementation Details

We use three graphons in our experiments: (1) Erdős Rényi: $W(\alpha,\beta) = 0.8$; (2) Stochastic block model: $W(\alpha,\beta) = 0.9$, if $0 \leqslant \alpha,\beta \leqslant 0.5$ or $0.5 \leqslant \alpha,\beta \leqslant 1$, $W(\alpha,\beta) = 0.4$, otherwise; (3) Random geometric graphon: $W(\alpha, \beta) = f(\min(|\beta - \alpha|, 1 - |\beta - \alpha|))$, where $f(x) = \mathrm{e}^{-\frac{x}{0.5-x}}$.

For the RL algorithms, we use the implementation of RLlib [36] (version 1.11.0, Apache-2.0 license). For PPO used to learn an optimal policy ensemble in block GFMC, we use a 64-dimensional linear layer to encode the observation and 2-layer MLPs with 256 hidden units per layer for both value network and actor network. For independent DQN and independent PPO, we use the default weight-sharing model with 64-dimensional embedding layers. We train the GMFC PPO agent for 1000 iterations, and other three MARL agents for 200 iterations. The specific hyper-parameters are listed in Table 3.

Table 3: RL Algorithm Settings

Algorithms	GMFC PPO	I-DQN	I-PPO	QMIX
Learning rate	0.0005	0.0005	0.0001	0.00005
Learning rate decay	True	True	True	False
Discount factor	0.95	0.95	0.95	0.95
Batch size	128	128	128	128
KL coefficient	0.2	-	0.2	-
KL target	0.01	-	0.01	-
Buffer size	-	2000	-	2000
Target network update frequency	-	2000	-	1000

6. Conclusion

In this work, we have proposed a discrete-time GMFC framework for MARL with nonuniform interactions and heterogeneous reward functions and transition functions across the agents on dense graphs. Theoretically, we have shown that under suitable assumptions, GMFC approximates MARL well with approximation error of order $\mathcal{O}(\frac{1}{\sqrt{N}})$. To reduce the dimension of GMFC, we have introduced block GMFC by discretizing the graphon index and shown that it also approximates MARL well. Empirical studies on several examples have verified the plausibility of the GMFC framework. For future research, we wish to explore more on how to extract the optimal policy of cooperative MARL without the simulator for population state distribution ensemble and to extend our framework to heterogeneous MARL on sparse graphs.

References

[1] Adler, J. L., Blue, V. J., 2002. A cooperative multi-agent transportation management and route guidance system. Transportation Research Part C: Emerging Technologies 10 (5-6), 433–454.

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

[3] Bayraktar, E., Chakraborty, S., Wu, R., 2020. Graphon mean field systems. arXiv preprint arXiv:2003.13180.

[4] Bet, G., Coppini, F., Nardi, F. R., 2020. Weakly interacting oscillators on dense random graphs. arXiv preprint arXiv:2006.07670.

[5] Caines, P. E., Huang, M., 2019. Graphon mean field games and the GMFG equations: ε-Nash equilibria. In: 2019 IEEE 58th conference on decision and control (CDC). IEEE, pp. 286–292.

[6] Cardaliaguet, P., Lehalle, C.-A., 2018. Mean field game of controls and an application to trade crowding. Mathematics and Financial Economics 12 (3), 335–363.

[7] Carmona, R., Cooney, D. B., Graves, C. V., Lauriere, M., 2022. Stochastic graphon games: I. the static case. Mathematics of Operations Research 47 (1), 750–778.

[8] Carmona, R., Delarue, F., 2013. Probabilistic analysis of mean-field games. SIAM Journal on Control and Optimization 51 (4), 2705–2734.

[9] Carmona, R., Laurière, M., Tan, Z., 2019. Model-free mean-field reinforcement learning: mean-field MDP and mean-field Q-learning. arXiv preprint arXiv:1910.12802.

[10] Chen, Y., Liu, J., Khoussainov, B., 2021. Agent-level maximum entropy inverse reinforcement learning for mean field games. arXiv preprint arXiv:2104.14654.

[11] Choi, J., Oh, S., Horowitz, R., 2009. Distributed learning and cooperative control for multi-agent systems. Automatica 45 (12), 2802–2814.

[12] Cortes, J., Martinez, S., Karatas, T., Bullo, F., 2004. Coverage control for mobile sensing networks. IEEE Transactions on Robotics and Automation 20 (2), 243–255.

[13] Cui, J., Liu, Y., Nallanathan, A., 2019. Multi-agent reinforcement learning-based resource allocation for UAV networks. IEEE Transactions on Wireless Communications 19 (2), 729–743.

[14] Cui, K., Koeppl, H., 2021. Approximately solving mean field games via entropy-regularized deep reinforcement learning. In: International Conference on Artificial Intelligence and Statistics. PMLR, pp. 1909–1917.

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

[16] Cui, K., Tahir, A., Sinzger, M., Koeppl, H., 2021. Discrete-time mean field control with environment states. In: 2021 60th IEEE Conference on Decision and Control (CDC). IEEE, pp. 5239–5246.

[17] Delarue, F., 2017. Mean field games: A toy model on an Erdös-Renyi graph. ESAIM: Proceedings and Surveys 60, 1–26.

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

[19] Esunge, J. N., Wu, J., 2014. Convergence of weighted empirical measures. Stochastic Analysis and Applications 32 (5), 802–819.

[20] Flyvbjerg, H., Sneppen, K., Bak, P., 1993. Mean field theory for a simple model of evolution. Physical review letters 71 (24), 4087.

[21] Gao, S., Caines, P. E., 2019. Graphon control of large-scale networks of linear systems. IEEE Transactions on Automatic Control 65 (10), 4090–4105.

[22] Gao, S., Caines, P. E., 2019. Spectral representations of graphons in very large network systems control. In: 2019 IEEE 58th conference on decision and Control (CDC). IEEE, pp. 5068–5075.

[23] Gu, H., Guo, X., Wei, X., Xu, R., 2021. Mean-field controls with Q-learning for cooperative MARL: convergence and complexity analysis. SIAM Journal on Mathematics of Data Science 3 (4), 1168–1196.

[24] Gu, H., Guo, X., Wei, X., Xu, R., 2023. Dynamic programming principles for mean-field controls with learning. Operations Research.

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

[26] Hadikhanloo, S., Silva, F. J., 2019. Finite mean field games: fictitious play and convergence to a first order continuous mean field game. Journal de Mathématiques Pures et Appliquées 132, 369–397.

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

[28] Hughes, E., Leibo, J. Z., Phillips, M., Tuyls, K., Dueñez-Guzman, E., García Castañeda, A., Dunning, I., Zhu, T., McKee, K., Koster, R., et al., 2018. Inequity aversion improves cooperation in intertemporal social dilemmas. Advances in neural information processing systems 31.

[29] Lacker, D., Soret, A., 2022. A label-state formulation of stochastic graphon games and approximate equilibria on large networks. Mathematics of Operations Research.

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

[31] Lee, J. W., Park, J., Jangmin, O., Lee, J., Hong, E., 2007. A multiagent approach to Q-learning for daily stock trading. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans 37 (6), 864–877.

[32] Lee, J. W., Zhang, B.-T., 2002. Stock trading system using reinforcement learning with cooperative agents. In: Proceedings of the Nineteenth International Conference on Machine Learning. pp. 451–458.

[33] Leibo, J. Z., Zambaldi, V., Lanctot, M., Marecki, J., Graepel, T., 2017. Multi-agent reinforcement learning in sequential social dilemmas. arXiv preprint arXiv:1702.03037.

[34] Lerer, A., Peysakhovich, A., 2017. Maintaining cooperation in complex social dilemmas using deep reinforcement learning. arXiv preprint arXiv:1707.01068.

[35] Li, Y., Wang, L., Yang, J., Wang, E., Wang, Z., Zhao, T., Zha, H., 2021. Permutation invariant policy optimization for mean-field multi-agent reinforcement learning: A principled approach. arXiv preprint arXiv:2105.08268.

[36] Liang, E., Liaw, R., Nishihara, R., Moritz, P., Fox, R., Goldberg, K., Gonzalez, J. E., Jordan, M. I., Stoica, I., 2018. RLlib: Abstractions for distributed reinforcement learning. In: International Conference on Machine Learning (ICML).

[37] Lin, Y., Qu, G., Huang, L., Wierman, A., 2021. Multi-agent reinforcement learning in stochastic networked systems. Advances in Neural Information Processing Systems 34, 7825–7837.

[38] Liu, B., Cai, Q., Yang, Z., Wang, Z., 2019. Neural proximal/trust region policy optimization attains globally optimal policy. arXiv preprint arXiv:1906.10306.

[39] Lovász, L., 2012. Large networks and graph limits. Vol. 60. American Mathematical Soc.

[40] Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., Riedmiller, M., 2013. Playing Atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602.

[41] Mondal, W., Aggarwal, V., Ukkusuri, S., 2022. On the near-optimality of local policies in large cooperative multi-agent reinforcement learning. Transactions on Machine Learning Research.

[42] Mondal, W. U., Agarwal, M., Aggarwal, V., Ukkusuri, S. V., 2022. On the approximation of cooperative heterogeneous multi-agent reinforcement learning (MARL) using mean field control (MFC). Journal of Machine Learning Research 23 (129), 1–46.

[43] Mondal, W. U., Aggarwal, V., Ukkusuri, S. V., 2022. Can mean field control (MFC) approximate cooperative multi agent reinforcement learning (MARL) with non-uniform interaction? In: Uncertainty in Artificial Intelligence. PMLR, pp. 1371–1380.

[44] Pasztor, B., Bogunovic, I., Krause, A., 2021. Efficient model-based multi-agent mean-field reinforcement learning. arXiv preprint arXiv:2107.04050.

[45] Qu, G., Wierman, A., Li, N., 2020. Scalable reinforcement learning of localized policies for multi-agent networked systems. In: Learning for Dynamics and Control. PMLR, pp. 256–266.

[46] Rashid, T., Samvelyan, M., Schroeder, C., Farquhar, G., Foerster, J., Whiteson, S., 2018. QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning. In: International Conference on Machine Learning. PMLR, pp. 4295–4304.

[47] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O., 2017. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.

[48] Subramanian, J., Mahajan, A., 2019. Reinforcement learning in stationary mean-field games. In: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems. pp. 251–259.

[49] Vasal, D., Mishra, R., Vishwanath, S., 2021. Sequential decomposition of graphon mean field games. In: 2021 American Control Conference (ACC). IEEE, pp. 730–736.

[50] W Axhausen, K., Horni, A., Nagel, K., 2016. The multi-agent transport simulation MATSim. Ubiquity Press.

[51] Wainwright, M. J., 2019. High-dimensional statistics: A non-asymptotic viewpoint. Vol. 48. Cambridge University Press.

[52] Yang, Y., Luo, R., Li, M., Zhou, M., Zhang, W., Wang, J., 2018. Mean field multi-agent reinforcement learning. In: International Conference on Machine Learning. PMLR, pp. 5571–5580.

[53] Yin, H., Mehta, P. G., Meyn, S. P., Shanbhag, U. V., 2010. Learning in mean-field oscillator games. In: 49th IEEE Conference on Decision and Control (CDC). IEEE, pp. 3125–3132.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.