Partially Observable Multi-Type Mean-Field Reinforcement Learning

Shuhui Chu* and Chengzhong Xu
Department of Computer and Information Science, University of Macau, Macau, China 999078
{yc07445, czxu}@um.edu.mo

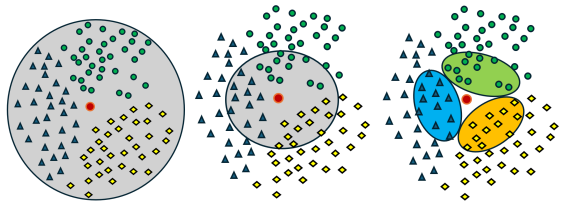
***Abstract*—Multi-agent reinforcement learning (MARL) holds significant promise for real-world applications, but scaling it to environments with many agents remains a fundamental challenge. Mean-field theory has emerged as a powerful approach to this problem, simplifying complex multi-agent interactions into tractable two-agent interactions. While this greatly enhances scalability, existing methods rely on strong and unrealistic assumptions—homogeneous agents and global observability—which limit their practical applicability and theoretical generality. In this work, we relax these restrictive assumptions by introducing agent heterogeneity and partial observability into the mean-field MARL framework. We explicitly model diverse agent behaviors through multiple agent types and operate under realistic local observation constraints. To this end, we propose POMTMFQ (Partially Observable Multi-Type Mean-Field Q-learning) under these more general conditions. We provide a comprehensive theoretical analysis, proving that POMTMFQ converges close to the Nash Q-value. Experimental results on three large-scale games in MAgent platform demonstrate the effectiveness and superiority of POMTMFQ. Notably, POMTMFQ achieves 42% higher winning rates and significantly faster convergence compared to prior methods, confirming its ability to learn more efficient policies in complex, many-agent environments.**

***Index Terms*—mean-field theory, partial observability, multiple types, reinforcement learning**

I. Introduction

Multi-Agent Reinforcement Learning (MARL) investigates how multiple agents interact with a shared environment and with each other [1]. A major obstacle to its widespread application is the curse of dimensionality: most MARL algorithms become computationally intractable as the number of agents grows large. To enable scalability, recent works [2] have adopted mean-field theory [3]. This approach approximates the complex interactions among a massive population by modeling the collective behavior of all other agents as a single mean field. This simplification reduces the multi-agent interactions to the two-agent interactions between a representative agent and this aggregate mean field, significantly lowering computational complexity and enabling scalability to large-scale populations. Mean-field MARL has shown promise in various large-scale applications, including task offloading in large-scale mobile edge computing [4], pickup-and-delivery vehicle routing problems in large-scale multi-agent systems [5], and resource allocation in vehicle-to-everything communication networks [6].

Classical standard mean-field approximation relies on two strong and unrealistic assumptions. First, it presumes agent homogeneity, treating all agents in the environment as behaviorally identical and indistinguishable [2], [7]. Therefore, the whole population can be approximated as a single mean field, as depicted in Fig. 1(a) where all agents within the grey area are aggregated into a single mean field. Second, it assumes global observability, granting each agent access to the information about the entire population in the environment [2]. Thus, as shown in Fig. 1(a), all agents' information in the whole population (the grey area) can be used for mean-field calculation. Recent efforts have begun to relax these assumptions. For instance, the work of [8] extended the framework to partially observable environments, where agents estimate the mean field using only local observations for neighbors within a limited range (shown in Fig. 1(b)). However, this work retains the assumption of agent homogeneity, failing to capture the inherent diversity of behaviors and objectives present in real-world systems—a critical shortcoming that limits practical utility.



(a) MF (b) POMF (c) Ours

Fig. 1: Mean-field approximation in different cases. Agents are represented as points. The red point denotes the representative agent interacting with its neighbors, which are categorized into three distinct types (indicated by different colors). (a) **MF** (Mean Field): All other agents within the grey area are aggregated into a single mean field. (b) **POMF** (Partially Observable Mean Field): Only observable neighboring agents (within a limited range) are used to estimate the mean field. (c) **Ours**: The partially observable neighbors are grouped by type, and a separate mean field is estimated for each agent type, enabling heterogeneity-aware approximation.

In this paper, we address the partially observable mean-field MARL problem with heterogeneous agents, which is described in Fig. 1(c). A direct integration of these elements faces significant **challenges and limitations**: First, agents in real-world environments exhibit distinct characteristics and strategies [9]. Simply aggregating them into a single, homogeneous mean field yields an insufficient representation, which can hinder exploration and policy learning, leading to inefficient policies. Second, agents in real-world large-scale scenarios typically only have access to information from their nearby neighbors. Then, utilizing global observations to calculate the mean field is impractical, as it would require a central controller to gather global information—a solution that incurs significant communication overhead and undermines decentralized decision-making. Third, existing theoretical analyses

in mean-field MARL works are often fragmented, lacking a unified and principled understanding of approximation errors and convergence guarantees in this generalized setting. This motivates our central research question:

*Can we establish a **holistic framework** that bridges these methodological and theoretical gaps in MARL?*

To solve this question, we propose a partially observable multi-type mean-field reinforcement learning framework. Our key idea is to categorize agents within the limited observation range into distinct types, where agents of the same type are homogeneous and aggregated into a type-specific mean field through estimation, as shown in Fig. 1(c). Each type is then modeled by its own type-specific mean field, creating a more expressive, accurate, and heterogeneity-aware representation. This framework seamlessly integrates agent heterogeneity and partial observability in mean-field MARL.

To address this theoretical gap, we develop a unified analytical framework for partially observable multi-type mean-field MARL grounded in stochastic game theory. This analysis provides two key contributions: it delivers a cohesive perspective that unifies and clarifies prior work, and it formally validates our method by establishing a convergence guarantee with respect to the Nash Q-value.

Our main contributions are as follows.

- **Problem.** In contrast to the standard assumptions of agent homogeneity and global observability in typical mean-field MARL, we explore the challenging mean-field MARL problem under agent heterogeneity and partial observability.
- **Methodology.** We propose the partially observable multi-type mean-field Q-learning framework (POMTMFQ), which models agent heterogeneity through multiple agent types and operates using partial observations.
- **Theory.** We provide a holistic theoretical analysis, deriving generalized bounds on the mean-field approximation error and convergence guarantee regarding the Nash Q-value. This theory offers a unified perspective on prior work by providing insights into agent heterogeneity and partial observability.
- **Performance.** Extensive experiments on large-scale games in the MAgent benchmark [10] demonstrate that POMTMFQ achieves significantly faster convergence and outperforms baselines by up to 42% in winning rates, confirming its superior performance.

The remainder of this paper is organized as follows. Section II reviews related work in mean-field MARL. Section III provides the necessary background and preliminaries. Our proposed POMTMFQ algorithm is detailed in Section IV, which also presents its theoretical analysis. Experimental results are reported in Section V. Finally, Section VI concludes the paper and discusses future work.

II. Related Work

Agent heterogeneity: Numerous studies [11]–[16] have focused on incorporating agent heterogeneity into mean field games (MFGs) [3] which combined game theory with mean field theory to solve large-scale multi-agent problems. One prominent approach uses weights to model the varying relationships among agents. [17] and [14] employed multi-head attention mechanisms to design attention-based weighted mean field structures for richer interaction features. Further developments include adaptive mean field Q-learning, where the mean field is approximated with dynamically adjusted weights [18]. A common limitation of these weighted methods is their reliance on a centralized module to compute the weights, which hinders decentralized execution. [19] introduced graphon mean field theory to capture full heterogeneity and non-uniform interactions. Another significant direction extends homogeneous mean field to heterogeneous settings with multiple agent types [20]. A fundamental practicality issue across many of these works is the assumption of accessible mean fields, which often requires global information or a central controller. This is infeasible in real-world partially observable environments. Our work addresses this gap by enabling the estimation and use of mean fields under partial observations.

Partial observability: A parallel line of research addresses the challenge of partial observability in MFGs [21]–[25]. [8] explicitly modeled agents with limited observation ranges and proposed a partially observable mean-field Q-learning method. [26] incorporated common noise in modeling partial observations in MFGs for realism. [7] proposed decentralized MFGs based on local observations. Techniques to process these local observations include graph attention networks for identifying relevant neighboring agents [27] and consensus algorithms for distributed population estimation [28]. Furthermore, frameworks for cooperative settings have been proposed where agents are trained on and respond to localized information [24]. However, these works predominantly estimate and respond to a single, aggregated mean field, treating the population as a homogeneous entity. This constitutes a significant limitation for environments with inherently diverse agents, as it fails to capture and leverage population substructures. Our work directly addresses this by introducing a framework that estimates and utilizes explicit multiple mean fields from partial observations, thereby integrating the strengths of both research directions.

In contrast to related works discussed, our work uniquely addresses the dual challenges of agent heterogeneity and environmental partial observability. Notably, the studies by [8], [20] are particularly relevant. The former introduces multiple agent types but assumes access to global observations, while the latter operates under partial observability but treats agents as homogeneous. Our work synthesizes these concepts, modeling agent diversity through multiple types and operating within partially observable environments. This integrated approach yields a more general and realistic framework.

III. Background

A. Stochastic Game

A stochastic game with N agents is defined by the tuple $\Gamma \triangleq (S, \{A^j\}_{j=1}^N, \{r^j\}_{j=1}^N, p, \gamma)$. S is the state space, A^j is

the action space of agent j, $r^j : S \times A^1 \times ... \times A^N \to R$ is the reward function for agent j, $p : S \times A^1 \times ... \times A^N \to \Omega(S)$ is the transition function with $\Omega(S)$ denoting the set of probability distributions over S, and $\gamma \in [0, 1)$ is the discount factor. At time t, agents in state s_t select actions according to their policies $\pi^j : S \to \Omega(A^j)$. After executing the joint action $\boldsymbol{a}_t = [a_t^1, \ldots, a_t^N]$, each agent j receives a reward $r_t^j = r^j(s_t, \boldsymbol{a}_t)$ and the system transitions to a new state $s_{t+1} \sim p(\cdot|s_t, \boldsymbol{a}_t)$. The objective of each agent is to maximize its expected cumulative discounted reward. Given an initial state s and a joint policy $\boldsymbol{\pi} = [\pi^1, ..., \pi^N]$, the value function for agent j is $v_{\boldsymbol{\pi}}^j(s) = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{\boldsymbol{\pi},p}[r_t^j | s_0 = s]$. The corresponding action-value function is $Q_{\boldsymbol{\pi}}^j(s, \boldsymbol{a}) = r^j(s, \boldsymbol{a}) + \gamma \mathbb{E}_{s' \sim p}[v_{\boldsymbol{\pi}}^j(s')]$, where s' denotes the next state. Within this framework, a standard formulation for MARL is the non-cooperative stochastic game, where agents act independently without forming explicit coalitions.

In stochastic games, the Nash equilibrium solution is defined as a joint policy $\boldsymbol{\pi}_* = (\pi_*^1, ..., \pi_*^N)$ that satisfies, for any valid state s and policy π^j, $v^j(s; \boldsymbol{\pi}_*) = v^j(s; \pi_*^1, ..., \pi_*^j, ..., \pi_*^N) \geq v^j(s; \pi_*^1, ..., \pi^j, ..., \pi_*^N)$. This condition signifies that each agent's policy is a best-response to the others, and no agent can improve its value function by unilaterally deviating from their equilibrium policy. The corresponding optimal Q function is termed Nash Q-function, denoted as $Q_*^j(s, \boldsymbol{a})$, which is evaluated under the Nash equilibrium policy. [29] proposed Nash Q-learning to compute this Nash equilibrium, introducing the Nash operator $\mathcal{H}^{Nash}\boldsymbol{Q}(s, \boldsymbol{a}) = \mathbb{E}_{s' \sim p}[\boldsymbol{r}(s, \boldsymbol{a}) + \gamma \boldsymbol{v}^{Nash}(s')]$. Here $\boldsymbol{Q}(s, \boldsymbol{a}) = [Q^1(s, \boldsymbol{a}), ..., Q^N(s, \boldsymbol{a})]$, $\boldsymbol{r}(s, \boldsymbol{a}) = [r^1(s, \boldsymbol{a}), ..., r^N(s, \boldsymbol{a})]$, and $\boldsymbol{v}^{Nash}(s') = [v_{\boldsymbol{\pi}_*}^1(s'), ..., v_{\boldsymbol{\pi}_*}^N(s')]$. Under standard assumptions, the Nash operator is a contraction mapping. This guarantees convergence of the iterative learning process to the fixed point at Nash equilibrium, known as the Nash Q-value.

B. Mean-Field MARL

A foundational application of mean field theory to MARL was introduced by [2] which addresses scalability in many-agent systems by approximating complex multi-agent interactions as a simplified two-agent paradigm. This approach reformulates the stochastic game in large populations under the assumption that agents are homogeneous. Specifically, the central agent's interactions with all other agents are abstracted into an interaction with a single mean field, effectively representing the collective behavior of its neighbors.

The mean-field approximation is derived through a two-step process in [2]. First, the Q-function is decomposed into a sum of pairwise local interactions:

$$Q^j(s, \boldsymbol{a}) = \frac{1}{N^j} \sum_{k \in \mathcal{N}(j)} Q^j(s, a^j, a^k), \tag{1}$$

where $\mathcal{N}(j)$ is the set of agent j's neighbors and $N^j = |\mathcal{N}(j)|$.

Subsequently, applying mean-field approximation and Taylor's expansion yields a simplified, tractable form:

$$Q^j(s, \boldsymbol{a}) \approx Q_{MF}^j(s, a^j, \bar{a}^j). \tag{2}$$

Here $\bar{a}^j = \frac{1}{N^j} \sum_{k \in \mathcal{N}(j)} a^k$ is the mean-field action, encapsulating the average behavior of agent j's neighborhood.

IV. Partially Observable Multi-Type Mean-Field MARL

While mean-field MARL successfully addresses scalability, it introduces two stringent requirements that limit its practical applicability. First, it assumes agent homogeneity, treating all agents as identical and indistinguishable. Second, it relies on global observability, requiring each agent to observe either all other agents or a sufficiently large neighborhood. These restrictive assumptions hinder the deployment of mean-field MARL in many real-world environments.

A. Our Method: POMTMFQ

To bridge the gap between theory and practice, this work relaxes the two stringent assumptions of classical mean-field MARL—agent homogeneity and global observability—to enhance its generalizability for real-world applications. We achieve this by introducing heterogeneous agent types and operating under partial observability.

We model heterogeneity by categorizing agents into M distinct types. Agents of the same type share behavioral similarities (preserving local homogeneity), while agents of different types exhibit diverse strategies, capturing the desired heterogeneity. For partial observability, we assume agents lack access to global information. Instead, each agent possesses a limited observation range and can only perceive other agents within this local neighborhood.

Formally, under these relaxed assumptions, we derive partially observable multi-type mean-field Q-function updates:

$$\begin{aligned} Q_{t+1}^j(s^j, a^j, \tilde{a}_1^j, ..., \tilde{a}_M^j) =& (1 - \alpha_t) Q_t^j(s^j, a^j, \tilde{a}_1^j, ..., \tilde{a}_M^j) + \\ & \alpha_t [r_t^j + \gamma v_t^j(s_{t+1}^j)], \end{aligned} \tag{3}$$

where α_t is the learning rate and

$$\begin{aligned} v_t^j(s_{t+1}^j) = \sum_{a^j} \pi_t^j(a^j | s_{t+1}^j, \tilde{a}_1^j, ..., \tilde{a}_M^j) \mathbb{E}_{\boldsymbol{a}^{-j} \sim \boldsymbol{\pi}_t^{-j}}[\\ Q_t^j(s_{t+1}^j, a^j, \tilde{a}_1^j, ..., \tilde{a}_M^j)]. \end{aligned} \tag{4}$$

Here, s^j denotes the local state observed by agent j, replacing the global state s. The key innovation is the introduction of M type-specific mean-field actions, each $\tilde{a}_m^j$ representing the estimated average action of neighboring agents belonging to type m. These mean fields are estimated directly from partial observations, as detailed next.

We focus on discrete action spaces, represented as one-hot vectors. For each agent type m, we maintain a categorical distribution over actions to model its mean field. Since the Dirichlet distribution is the conjugate prior for the categorical distribution, it provides a Bayesian framework for updating our estimates based on local observations:

$$\begin{aligned} \mathcal{D}_m^j(\theta) \propto \theta_1^{\eta_{m,1} - 1 + c_{m,1}} ... \theta_L^{\eta_{m,L} - 1 + c_{m,L}}; \mathcal{D}_m^j(\theta | \eta_m + c_m); \\ \forall m \in \{1, 2, ..., M\}. \end{aligned} \tag{5}$$

Here L is the size of the action space, θ denotes the categorical parameters, η_m denotes the Dirichlet parameters for type m, and c_m is the observed action counts from neighbors of type m. The mean-field action for type m is then estimated by sampling from this distribution:

$$\tilde{a}^j_{i,m} \sim \mathcal{D}^j_m(\theta; \eta_m + c_m); \tilde{a}^j_m = \frac{1}{X}\sum_{i=1}^{X} \tilde{a}^j_{i,m}. \tag{6}$$

Here X is the number of samples. This sampling process inherently introduces stochasticity in the mean-field estimation, facilitating further policy exploration.

Finally, each agent selects the optimal action using the Boltzmann policy derived from its Q-function:

$$\pi^j_t(a^j|s^j, \tilde{a}^j_1, ..., \tilde{a}^j_M) = \frac{exp(-\beta Q^j_t(s^j, a^j, \tilde{a}^j_1, ..., \tilde{a}^j_M))}{\sum\limits_{a\in A^j} exp(-\beta Q^j_t(s^j, a, \tilde{a}^j_1, ..., \tilde{a}^j_M))}, \tag{7}$$

where β is the Boltzmann parameter.

B. Algorithm Implementation

Building on the formulations above, we present the Partially Observable Multi-Type Mean-Field Q-learning algorithm (POMTMFQ), as detailed in Algorithm 1.

The Q-function is implemented using neural networks, parameterized by weights ϕ^j for each agent j. During training, each agent minimizes the loss function:

$$\mathcal{L}(\phi^j) = (y^j - Q_{\phi^j}(s^j, a^j, \tilde{a}^j_1, ..., \tilde{a}^j_M))^2, \tag{8}$$

where the target value $y^j = r^j + \gamma v_{\bar{\phi}^j}(s')$ with weights $\bar{\phi}^j$. The corresponding gradient is:

$$\nabla_{\phi^j}\mathcal{L}(\phi^j) = 2(Q_{\phi^j}(s^j, a^j, \tilde{a}^j_1, ..., \tilde{a}^j_M) - y^j)\times \nabla_{\phi^j} Q_{\phi^j}(s^j, a^j, \tilde{a}^j_1, ..., \tilde{a}^j_M). \tag{9}$$

C. Theoretical Results

We present the core theoretical contributions: two theorems that unify the analysis of partial observability and multiple agent types. The first quantifies the approximation error introduced by our POMTMFQ Q-function relative to the exact Q-function. The second establishes its convergence properties. We begin by introducing a series of supporting lemmas, forming the foundation for our main results.

Lemma 1. *Let $\tilde{a}_{i,m,t}$ be the i-th component of $\tilde{a}_{m,t}$, the POMTMFQ estimated mean action for type m at time t, and let $\bar{a}_{i,m,t}$ be the corresponding component of $\bar{a}_{m,t}$, the oracle mean action. Thus, the following inequality holds as $t \to \infty$.*

$$|\tilde{a}_{i,m,t} - \bar{a}_{i,m,t}| \leq \sqrt{\frac{1}{2n}\log\frac{2}{\delta}}, \tag{10}$$

with probability $\geq \delta$. n is the number of observed samples.

This bound on individual action components follows directly from Hoeffding's inequality.

Lemma 2. *Both the estimated mean action $\tilde{a}_{m,t}$ and the oracle mean action $\bar{a}_{m,t}$ yield as $t \to \infty$:*

$$\|\tilde{a}_{m,t} - \bar{a}_{m,t}\|_2 \leq \sqrt{\frac{L}{2n}\log\frac{2}{\delta}}, \tag{11}$$

Algorithm 1 Partially Observable Multi-Type Mean-Field Q-learning (POMTMFQ)

Initialize each mean-field action $\tilde{a}^j_m$.
Initialize Q_{ϕ^j} and $Q_{\bar{\phi}^j}$ for each agent j.
Initialize Dirichlet distributions $\mathcal{D}^j_m(\theta)$.
Initialize number of episodes E and number of steps T.
while Episode $< E$ **do**
  while Step $< T$ **do**
    Calculate action a^j with Boltzmann policy in Eq. 7 for each agent j.
    Compute Dirichlet parameters $\mathcal{D}^j_m(\theta)$ by Eq. 5 for each agent type m.
    Update mean-field action $\tilde{a}^j_m$ by Eq. 6 for each agent type m.
    Take joint action $\boldsymbol{a}$, receive reward $\boldsymbol{r}$, and transition to next state $\boldsymbol{s}'$ for all agents.
    Store $< \boldsymbol{s}, \boldsymbol{a}, \boldsymbol{r}, \boldsymbol{s}', \tilde{\boldsymbol{a}}_1, ..., \tilde{\boldsymbol{a}}_M >$ in replay buffer.
  end while
  Sample K experiences from replay buffer for each agent j.
  Sample action a^j from $Q_{\bar{\phi}^j}$.
  Compute $y^j = r^j + \gamma v_{\bar{\phi}^j}(s')$ by Eq. 4.
  Update Q network for each agent j by minimizing the loss: $\mathcal{L}(\phi^j) = \frac{1}{K}\sum(y^j - Q_{\phi^j}(s^j, a^j, \tilde{a}^j_1, ..., \tilde{a}^j_M))^2$.
  Update target network with learning rate τ for each agent j: $\bar{\phi}^j \leftarrow \tau\phi^j + (1-\tau)\bar{\phi}^j$.
end while

with probability $\geq \delta^{L-1}$. L is the size of the action space.

This bound is derived by applying Lemma 1 across all L dimensions and accounting for the joint probability.

Lemma 3. *Let $\tilde{\boldsymbol{a}}_t = [\tilde{a}_{1,t}; \tilde{a}_{2,t}; ...; \tilde{a}_{M,t}]$ and $\bar{\boldsymbol{a}}_t = [\bar{a}_{1,t}; \bar{a}_{2,t}; ...; \bar{a}_{M,t}]$, which yield the following inequality.*

$$\|\tilde{\boldsymbol{a}}_t - \bar{\boldsymbol{a}}_t\|_2 \leq \sqrt{\frac{ML}{2n}\log\frac{2}{\delta}}, \tag{12}$$

with probability $\geq \delta^{M(L-1)}$ as $t \to \infty$.

This extends Lemma 2 by aggregating the error bounds across all M agent types.

Lemma 4. *Assume the Q-function is K-Lipschitz continuous with respect to the multi-type mean actions. Thus, the following inequality holds.*

$$|Q^j(s^j_t, a^j_t, \tilde{\boldsymbol{a}}^j_t) - Q^j(s^j_t, a^j_t, \bar{\boldsymbol{a}}^j_t)| \leq K\sqrt{\frac{ML}{2n}\log\frac{2}{\delta}}, \tag{13}$$

with probability $\geq \delta^{M(L-1)}$ as $t \to \infty$.

This bound follows from Lemma 3 and the Lipschitz continuous condition.

Lemma 5. *Let ϵ_0 bound the average deviation between each action and the oracle mean action of its type. If the Q-function is K_0-smooth, then the following inequality holds.*

$$|Q^j(s, \boldsymbol{a}) - Q^j(s, a^j, \bar{a}^j_1, ..., \bar{a}^j_M)| \leq \frac{1}{2}K_0\epsilon_0, \tag{14}$$

where $\boldsymbol{a}$ is the joint action of all agents and $Q^j(s, \boldsymbol{a})$ is the exact Q-function.

This quantifies the inherent approximation error when replacing the joint action with multi-type mean fields, using the smoothness property.

Lemma 6. *If the Q-function based on multi-type mean fields is K_1-Lipschitz continuous with respect to the agent j's action, then for any two distinct actions a^j and b^j:*

$$|Q^j(s, a^j, \bar{a}_1^j, ..., \bar{a}_M^j) - Q^j(s, b^j, \bar{a}_1^j, ..., \bar{a}_M^j)| \leq K_1\sqrt{2}. \quad (15)$$

This bound leverages the one-hot encoding of actions and the Lipschitz continuous condition.

Lemma 7. *Let v^{EMT} and v^{OMT} be the value functions using the estimated and oracle multi-type mean actions, respectively. Then, they satisfy*

$$|v^{EMT}(s_{t+1}) - v^{OMT}(s_{t+1})| \leq K\sqrt{\frac{ML}{2n}\log\frac{2}{\delta}} + K_1\sqrt{2}, \quad (16)$$

with probability $\geq \delta^{M(L-1)}$ as $t \to \infty$.

This bound is derived by combining Lemmas 4 and 6.

Theorem 1 (Approximation Error Bound)**.** *If the Q-function is K_0-smooth and K-Lipschitz continuous w.r.t. the mean actions, the approximation error between the proposed Q-function and the exact Q-function is bounded as follows.*

$$|Q^j(s, \boldsymbol{a}) - Q^j(s, a^j, \tilde{a}_1^j, ..., \tilde{a}_M^j)| \leq \frac{1}{2}K_0\epsilon_0 + K\sqrt{\frac{ML}{2n}\log\frac{2}{\delta}}, \quad (17)$$

with probability $\geq \delta^{M(L-1)}$.

Proof. The proof follows from the triangle inequality and the application of Lemmas 5 and 4:

$$\begin{aligned}
&|Q^j(s, \boldsymbol{a}) - Q^j(s, a^j, \tilde{a}_1^j, ..., \tilde{a}_M^j)| \\
\leq &|Q^j(s, \boldsymbol{a}) - Q^j(s, a^j, \bar{a}_1^j, ..., \bar{a}_M^j)| + \\
&|Q^j(s, a^j, \bar{a}_1^j, ..., \bar{a}_M^j) - Q^j(s, a^j, \tilde{a}_1^j, ..., \tilde{a}_M^j)| \\
\leq &\frac{1}{2}K_0\epsilon_0 + K\sqrt{\frac{ML}{2n}\log\frac{2}{\delta}},
\end{aligned}$$

with probability at least $\delta^{M(L-1)}$. □

Remark 1. In a fully observable environment with large n, Theorem 1 reduces to the approximation bound of MTMFQ [20]. When $M = 1$, it further simplifies to the bound of standard MFQ [2].

Remark 2. When $M = 1$, Theorem 1 recovers the approximation bound of POMFQ [8] for homogeneous, partially observable settings.

What does the approximation error bound reveal? *Theorem 1 provides a unified perspective on the mean-field approximation error, jointly accounting for the effects of multiple agent types and partial observations.* The bound in Eq. 17 contains two terms: (1) mean-field approximation error due to modeling multiple agent types, and (2) estimation error introduced by inferring mean fields from partial observations. This bound yields several key insights. First, the error decreases as the number of observed samples n increases, indicating a larger n improves approximation accuracy. However, a very large n can reduce the exploration noise, potentially degrading learning performance. Thus, a trade-off exists when selecting n to balance accurate estimation with sufficient exploration. Second, the bound scales with the number of agent types M. A smaller M yields a tighter bound but may inadequately capture agent heterogeneity. Conversely, a larger M better models agent diversity at the cost of increased computational complexity and a looser bound. Thus, a second critical trade-off exists between model accuracy and computational efficiency.

We now establish the convergence guarantee for the proposed POMTMFQ to the Nash Q-value. We begin with three standard assumptions commonly used in previous research [2].

Assumption 1. Each action-value pair is visited infinitely often, and the reward remains bounded.

Assumption 2. Agents' policies are Greedy in the Limit with Infinite Exploration (GLIE).

Assumption 3. For each stage game of the stochastic game, the Nash equilibrium can be recognized as either a global optimum or a saddle point.

Theorem 2 (Convergence Guarantee)**.** *Under Assumptions 1-3, the proposed Q-function in POMTMFQ converges to a bounded distance of the Nash Q-value Q_* as $t \to \infty$:*

$$|Q_*(s_t, \boldsymbol{a}_t) - Q(s_t, a_t, \tilde{a}_{1,t}, ..., \tilde{a}_{M,t})| \leq 2Z, \quad (18)$$

with probability $\geq \delta^{M(L-1)}$. $Z \triangleq K\sqrt{\frac{ML}{2n}\log\frac{2}{\delta}} + K_1\sqrt{2}$.

Proof. We frame the proposed Q-function update as a stochastic process of the following form:

$$x_i(t+1) = x_i(t) + \alpha_i(t)(F_i(x^i(t)) - x_i(t) + w_i(t)). \quad (19)$$

By identifying $x(t)$ with the Q-function and verifying the four assumptions and the bounded requirement of Theorem 1 in [8], we show that they are all satisfied and then this process is bounded. Applying Theorem 1 in [8] yields the result that $Q_t^j(s_t^j, a_t^j, \tilde{a}_{1,t}^j, ..., \tilde{a}_{M,t}^j)$ is bounded in the interval $[Q_*(s_t, \boldsymbol{a}_t) - 2Z, Q_*(s_t, \boldsymbol{a}_t) + 2Z]$ as $t \to \infty$. Due to space limit, detailed verification and derivation are omitted here. □

V. Experimental Results

Experimental settings. We evaluate POMTMFQ on three large-scale cooperative-competitive games from the open-source MAgent framework [10]: Multibattle, Battle-Gathering, and Predator-Prey. In each game, six distinct groups of agents (labeled Group A through Group F) interact. Agents within the same group cooperate, while groups compete against each other. The winning group is determined by the highest number of surviving agents at the end of a game. We have 50 agents per group for both Multibattle and Battle-Gathering, and 35 predators per group (3 groups) and 70 prey per group (3 groups) for Predator-Prey. Our evaluation consists of two stages: In the training stage, all groups are trained against each other for 2000 episodes using the same algorithm. Then, in the faceoff stage, groups trained with different algorithms

fight each other over 1200 games. All experiments are repeated 50 times, and we report the average results in both stages.

Hyperparameter settings. To ensure a fair comparison, we maintain nearly identical hyperparameter configurations across all evaluated algorithms. The learning rate α is 10^{-4}, and the exploration rate decays linearly from 1 to 0 during training. The discount factor γ is 0.95, the mini-batch size is 64, and the replay buffer size is 2^{10}. We have a maximum of 500 steps for each game. The number of samples from Dirichlet distribution is 100 for both POMTMFQ and POMFQ algorithms.

Reward functions for Multibattle. In the Multibattle game, core environmental penalties include: -0.01 for moving, -1 for dying, and -0.1 for attacking an empty grid. Each agent begins with 10 health and deals 2 damage per attack, and dies when its health reaches zero, with a health recovery rate of 0.1 per step from attacks. The primary rewards are given for attacking and killing opponents from other groups. The reward values are asymmetric and depend on the specific groups involved, as detailed in Table I. For instance, an agent from Group A receives a reward of 0.2 for attacking an agent from Group B, and a reward of 80 for killing one. These values create a competitive hierarchy influencing strategic targeting.

TABLE I: Rewards for attack (kill) actions between different group agents (the rewards in brackets are for kill actions)

	A	B	C	D	E	F
A	–	0.2(80)	0.3(90)	0.4(100)	0.5(110)	0.6(120)
B	0.6(120)	–	0.2(80)	0.3(90)	0.4(100)	0.5(110)
C	0.5(110)	0.6(120)	–	0.2(80)	0.3(90)	0.4(100)
D	0.4(100)	0.5(110)	0.6(120)	–	0.2(80)	0.3(90)
E	0.3(90)	0.4(100)	0.5(110)	0.6(120)	–	0.2(80)
F	0.2(80)	0.3(90)	0.4(100)	0.5(110)	0.6(120)	–

Reward functions for Battle-Gathering. The Battle-Gathering game extends the Multibattle mechanics by introducing limited food resources into the environment. The rewards (for attack, kill, and penalties) remain identical to those in Multibattle. The key addition is the food mechanic: an agent receives a reward of 80 for successfully capturing food, which requires repeatedly attacking the grid containing food (each such attack yields a reward of 0.5) until the food's power is depleted. This adds a strategic layer of competition over limited resources alongside direct combat.

Reward functions for Predator-Prey. This game features asymmetric roles. There are three predator groups (A, B, C) and three prey groups (D, E, F). Predators have 10 health, an attack range of 2, and a speed of 2. They incur penalties of -0.1 for dying and -0.2 for attacking an empty grid. Prey have only 2 health, no attack capability, but a faster speed of 2.5. All agents have a health recovery rate of 0.1 per step. A fixed reward of 5 is given for killing any agent from a different group. Additionally, agents receive attack rewards as specified in Table II. The attacked agent simultaneously receives a penalty equal to the attacker's reward.

Compared methods. We evaluate our approach against five established baselines: (1) **IL** [30], a fundamental MARL approach that treats other agents as part of the environment, neglecting explicit interactions; (2) **MFQ** [2], a standard mean-field MARL method that assumes homogeneous agents and global observability; (3) **POMFQ** [8], which considers partially observable settings but retains the assumption of agent homogeneity in mean-field MARL; (4) **MTMFQ** [20], which models agent heterogeneity with multiple types but requires global observability in mean-field MARL; (5) **IA2C++DM** [26], a non-mean-field method under partial observability.

TABLE II: Rewards for attack actions between different group agents

	A	B	C	D	E	F
A	–	0.5	0.8	1	3	5
B	0.8	–	0.5	3	5	1
C	0.5	0.8	–	5	1	3

POMTMFQ achieves significantly faster convergence. We train all agent groups on the three games using the methods outlined above. The results are shown in Figs. 2(a), 2(b), and 2(c), which present the cumulative group reward for each method. Since all groups use the same algorithm during training, we report the results for Group A as a representative example. The results show that POMTMFQ converges substantially faster than all baselines and achieves higher cumulative rewards after convergence. This further demonstrates that our method learns more effective policies. This performance advantage comes from two key design elements. First, POMTMFQ outperforms MTMFQ because it operates under partial observability, where the stochastic noise introduced during mean-field estimation fosters more robust exploration. Second, it surpasses both POMFQ and IA2C++DM by explicitly modeling agent heterogeneity through multiple types, enabling a more accurate representation of the population structure. We also observe that convergence requires more episodes in Battle-Gathering and Predator-Prey than in Multibattle. This is expected, as these games present greater complexity: Battle-Gathering introduces strategic competition over limited food resources, while Predator-Prey features fundamentally asymmetric agent roles (predators vs. prey).

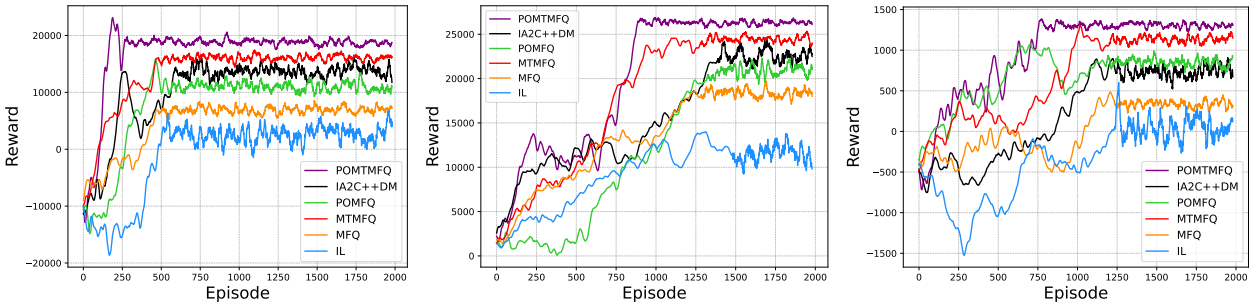
TABLE III: Winning rate (%) of different methods on three games. MB: Multibattle. BG: Battle-Gathering. PP: Predator-Prey.

Method	**MB**	**BG**	**PP**
IL	0	1	1
MFQ	3	6	6
POMFQ	6	10	8
MTMFQ	22	15	25
IA2C++DM	6	11	9
POMTMFQ (Ours)	**63**	**57**	**51**

TABLE IV: Average total reward of different methods on three games. MB: Multibattle. BG: Battle-Gathering. PP: Predator-Prey.

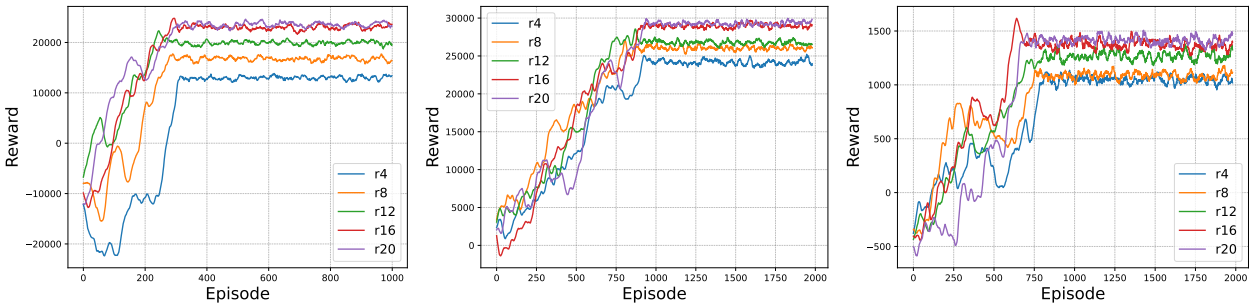
Method	**MB**	**BG**	**PP**
IL	-5520	7811	81
MFQ	8210	13051	243
POMFQ	11032	16291	402
MTMFQ	15821	21318	930
IA2C++DM	12027	18021	501
POMTMFQ (Ours)	**21127**	**30021**	**1503**

POMTMFQ achieves higher winning rates and total rewards. To evaluate the trained policies, we conduct a faceoff



(a) Training curve in MB (b) Training curve in BG (c) Training curve in PP

Fig. 2: Convergence of different methods in the training stage on three games. MB: Multibattle. BG: Battle-Gathering. PP: Predator-Prey.



(a) Training curve in MB (b) Training curve in BG (c) Training curve in PP

Fig. 3: Training results of POMTMFQ with different observation range on three games. MB: Multibattle. BG: Battle-Gathering. PP: Predator-Prey.

stage where the six groups, each trained with a different algorithm, compete across all three games. To ensure a fair comparison and account for potential group advantages, we rotate the algorithm assigned to each group every 200 episodes so that each algorithm is evaluated in each group position. The competitive results, summarized in Tables III and IV, demonstrate the clear superiority of our method. As shown in Table III, the group using POMTMFQ achieves a winning rate approximately 42% higher than any baseline. Furthermore, Table IV indicates that our method also secures the highest average total reward per group than all baselines. These consistent results—dominant winning rates and superior total rewards—provide strong empirical evidence that POMTMFQ learns more effective and robust policies than existing methods. This underscores its enhanced capability in heterogeneous, and partially observable environments.

Impact of the observation range on POMTMFQ. We investigate how the agent's observation range affects the performance of POMTMFQ by conducting additional training experiments across all three games. The observation range, which defines an agent's observable radius, is set to 4, 8, 12, 16, and 20 units (denoted as r4, r8, r12, r16, and r20). The results are presented in Figs. 3(a), 3(b), and 3(c). As shown in the results, the reward generally increases with a larger observation range. This trend is expected: a wider observation radius allows an agent to perceive more neighboring agents, thereby providing richer interaction data from which to learn more efficient policies. Notably, the performance improvement begins to plateau at larger ranges. The results for r16 and r20 are nearly identical after convergence. This suggests that POMTMFQ effectively extracts and utilizes the most salient interaction information from a moderately sized local neighborhood. Once a critical observation threshold is reached, additional observable agents contribute marginal utility. This demonstrates the method's efficiency in learning effective strategies under significant partial observability constraints, without requiring full global vision.

Discussion. The experimental results on three games illustrate that our method consistently outperforms all baselines in terms of convergence speed, winning rate, and total reward. Analyzing the performance hierarchy among the baselines provides further insight. MTMFQ consistently ranks as the strongest baseline, which is expected given that our experimental setting involves six distinct agent groups. MTMFQ's explicit modeling of agent diversity through multiple types gives it a significant advantage over methods that assume homogeneity. The results also show that POMFQ outperforms MFQ. This can be attributed to the exploration noise introduced in POMFQ's partially observable framework, which helps agents avoid local optima. Furthermore, IA2C++DM exceeds POMFQ's performance, likely due to its explicit handling of observation noise, which introduces greater exploratory

randomness. As anticipated, IL performs the worst, as it fails to model inter-agent interactions altogether. The superior performance of our POMTMFQ comes from its integrated design, which simultaneously addresses the two key limitations of prior work: it models agent heterogeneity through multiple types and operates under partial observability.

VI. Conclusion

Mean field theory provides a foundational framework for scalable MARL in large-scale multi-agent systems. In this paper, we address two critical challenges of conventional mean-field MARL—agent homogeneity and global observability—by introducing POMTMFQ, a novel method for Partially Observable Multi-Type Mean-Field Q-learning. We provide a comprehensive theoretical analysis that unifies and clarifies prior works in this space. Experimental results across several challenging games demonstrate the effectiveness and superiority of our approach. For future work, theoretically, we plan to explore mean-field approximations for both actions and states, and consider modeling in continuous spaces. Practically, we intend to validate the feasibility and effectiveness of our method in real-world application scenarios, such as resource allocation for UAV-assisted vehicle-to-everything communication networks.

Acknowledgment

This work is jointly funded by MOST and The Science and Technology Development Fund (FDCT), Macau SAR (File no. 0074/2025/AMJ).

References

[1] T. T. Nguyen, N. D. Nguyen, and S. Nahavandi, "Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications," *IEEE transactions on cybernetics*, vol. 50, no. 9, pp. 3826–3839, 2020.

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

[3] J.-M. Lasry and P.-L. Lions, "Mean field games," *Japanese journal of mathematics*, vol. 2, no. 1, pp. 229–260, 2007.

[4] D. Wang, W. Wang, H. Gao, Z. Zhang, and Z. Han, "Delay-optimal computation offloading in large-scale multi-access edge computing using mean field game," *IEEE Transactions on Wireless Communications*, vol. 23, no. 3, pp. 1684–1698, 2023.

[5] K. Cui, S. Azem, C. Fabian, K. Kuroptev, R. Khalili, O. Abboud, F. Steinke, and H. Koeppl, "Mean-field rl for large-scale unit-capacity pickup-and-delivery problems," *Transactions on Machine Learning Research*, 2025.

[6] Y. Xu, L. Zheng, X. Wu, Y. Tang, W. Liu, and D. Sun, "Joint resource allocation for uav-assisted v2x communication with mean field multi-agent reinforcement learning," *IEEE Transactions on Vehicular Technology*, vol. 74, no. 1, pp. 1209–1223, 2024.

[7] S. G. Subramanian, M. E. Taylor, M. Crowley, and P. Poupart, "Decentralized mean field games," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 9, 2022, pp. 9439–9447.

[8] S. Ganapathi Subramanian, M. E. Taylor, M. Crowley, and P. Poupart, "Partially observable mean field reinforcement learning," in *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, 2021, pp. 537–545.

[9] A. Ghosh and V. Aggarwal, "Model free reinforcement learning algorithm for stationary mean field equilibrium for multiple types of agents," *arXiv preprint arXiv:2012.15377*, 2020.

[10] L. Zheng, J. Yang, H. Cai, M. Zhou, W. Zhang, J. Wang, and Y. Yu, "Magent: A many-agent reinforcement learning platform for artificial collective intelligence," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[11] K. Cui and H. Koeppl, "Learning graphon mean field games and approximate nash equilibria," in *International Conference on Learning Representations*, 2021.

[12] A. Sridhar and S. Kar, "Mean-field approximations for stochastic population processes with heterogeneous interactions," *SIAM Journal on Control and Optimization*, vol. 61, no. 6, pp. 3442–3466, 2023.

[13] C. Yu, "Hierarchical mean-field deep reinforcement learning for large-scale multiagent systems," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 10, 2023, pp. 11 744–11 752.

[14] B. Wang, S. Li, X. Gao, and T. Xie, "Weighted mean field reinforcement learning for large-scale uav swarm confrontation," *Applied Intelligence*, vol. 53, no. 5, pp. 5274–5289, 2023.

[15] S. Dey and H. Xu, "Distributed adaptive flocking control for large-scale multiagent systems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 2, pp. 3126–3135, 2024.

[16] Y. Liang, B.-C. Wang, and H. Zhang, "Asymptotically optimal distributed control for linear quadratic mean field social systems with heterogeneous agents," *IEEE Transactions on Automatic Control*, 2024.

[17] X. Yuan, H. Wang, and W. Yu, "A weighted mean field reinforcement learning algorithm for large-scale multi-agent collaboration," *Guidance, Navigation and Control*, vol. 3, no. 02, p. 2350007, 2023.

[18] X. Wang, L. Ke, G. Zhang, and D. Zhu, "Adaptive mean field multi-agent reinforcement learning," *Information Sciences*, vol. 669, p. 120560, 2024.

[19] Y. Hu, X. Wei, J. Yan, and H. Zhang, "Graphon mean-field control for cooperative multi-agent reinforcement learning," *Journal of the Franklin Institute*, vol. 360, no. 18, pp. 14 783–14 805, 2023.

[20] S. Ganapathi Subramanian, P. Poupart, M. E. Taylor, and N. Hegde, "Multi type mean field reinforcement learning," in *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, 2020, pp. 411–419.

[21] N. Sen and P. E. Caines, "Mean field games with partial observation," *SIAM Journal on Control and Optimization*, vol. 57, no. 3, pp. 2064–2091, 2019.

[22] B. Yardim, S. Cayci, M. Geist, and N. He, "Policy mirror ascent for efficient and independent learning in mean field games," in *International Conference on Machine Learning*. PMLR, 2023, pp. 39 722–39 754.

[23] K. Cui, S. H. Hauck, C. Fabian, and H. Koeppl, "Partially observable multi-agent reinforcement learning using mean field control," in *ICML 2024 Workshop: Aligning Reinforcement Learning Experimentalists and Theorists*, 2024.

[24] H. Gu, X. Guo, X. Wei, and R. Xu, "Mean-field multiagent reinforcement learning: A decentralized network approach," *Mathematics of Operations Research*, 2024.

[25] D. Xie, Z. Wang, C. Chen, and D. Dong, "Depthwise convolution for multi-agent communication with enhanced mean-field approximation," *IEEE transactions on neural networks and learning systems*, vol. 35, no. 6, pp. 8557–8569, 2024.

[26] K. He, P. Doshi, and B. Banerjee, "Modeling and reinforcement learning in partially observable many-agent systems," *Autonomous Agents and Multi-Agent Systems*, vol. 38, no. 1, p. 12, 2024.

[27] M. Yang, G. Liu, Z. Zhou, and J. Wang, "Partially observable mean field multi-agent reinforcement learning based on graph attention network for uav swarms," *Drones*, vol. 7, no. 7, p. 476, 2023.

[28] F. Rajabali and R. Malhamé, "Can mean field game equilibria amongst exchangeable agents survive under partial observability of their competitors' states?" in *2023 62nd IEEE Conference on Decision and Control (CDC)*. IEEE, 2023, pp. 8188–8193.

[29] J. Hu and M. P. Wellman, "Nash q-learning for general-sum stochastic games," *Journal of machine learning research*, vol. 4, no. Nov, pp. 1039–1069, 2003.

[30] M. Tan, "Multi-agent reinforcement learning: Independent vs. cooperative agents," in *Proceedings of the tenth international conference on machine learning*, 1993, pp. 330–337.