# Learning Large-Scale Competitive Team Behaviors with Mean-Field Interactions

Bhavini Jeloka
Georgia Institute of Technology
Atlanta, United States of America
bjeloka3@gatech.edu

Yue Guan
Georgia Institute of Technology
Atlanta, United States of America
yguan44@gatech.edu

Panagiotis Tsiotras
Georgia Institute of Technology
Atlanta, United States of America
tsiotras@gatech.edu

## ABSTRACT

State-of-the-art multi-agent reinforcement learning (MARL) algorithms such as MADDPG and MAAC fail to scale in situations where the number of agents becomes large. Mean-field theory has shown encouraging results in modeling macroscopic agent behavior for teams with a large number of agents through a continuum approximation of the agent population and its interaction with the environment. In this work, we extend proximal policy optimization (PPO) to the mean-field domain by introducing the Mean-Field Multi-Agent Proximal Policy Optimization (MF-MAPPO), a novel algorithm that utilizes the effectiveness of the finite-population mean-field approximation in the context of zero-sum *competitive* multi-agent games between two teams. The proposed algorithm can be easily scaled to hundreds and thousands of agents in each team as shown through numerical experiments. In particular, the algorithm is applied to realistic applications such as large-scale offense-defense battlefield scenarios.

## KEYWORDS

Multi-Agent Reinforcement Learning, Game Theory, Large-Scale Systems

## 1 INTRODUCTION

Existing state-of-the-art multi-agent reinforcement learning (MARL) algorithms, such as MADDPG and MAAC [11], encounter significant scalability challenges as the number of agents increases. The associated complexities arise due to well-known *curse of dimensionality*. One promising direction that addresses the scalability challenge in MARL is through mean-field theory, which approximates large-scale agent interactions with the environment at an infinite population limit [6]. Two major areas of mean-field research are the mean-field games (MFGs) [5, 6, 19] which focus on non-cooperative agents, and mean-field control problems (MFC) [6, 15, 18], which study fully cooperative scenarios. However, work in mixed collaborative-competitive settings is relatively sparse.

The recent work in [4] formulated and studied zero-sum *mean-field team games* (ZS-MFTG), which models large-population teams competing against each other while agents within a team cooperate. In a two-team scenario, [4] utilized a common-information decomposition [12] to reduce the original problem to training two fictitious team coordinators, making the approach agnostic to the actual number of agents in the team, thus significantly reducing the computational load. The existence of approximate optimal team policies that are *identical* across agents was also established. However, numerically computing the corresponding team policies is not straightforward. Our work makes a contribution in this direction by leveraging deep reinforcement learning and exploiting theoretical properties of the MFTG obtained in [4], specifically the identical team policies and the common-information decomposition.

We propose the Mean-Field Multi-Agent Proximal Policy Optimization algorithm (MF-MAPPO), a novel multi-agent RL algorithm that extends PPO to accommodate intra-team cooperation and inter-team competition in large-population scenarios. The proposed algorithm employs a shared actor and critic for each team, with the information commonly available to all agents as inputs.

As shown in extensive numerical experiments, backed by theoretical guarantees from [4], our method scales efficiently to teams with hundreds or thousands of agents. Notably, the algorithm operates independently of individual agents' private information and is agnostic to the specific index/identity of each agent. To the best of our knowledge, this is the first algorithm that applies PPO to learning mixed competitive and collaborative mean-field problems.

The main contributions of our work are: 1) the MF-MAPPO algorithm that relies on shared critic and actor to efficiently learn large-scale team games; 2) novel MFTG scenarios (constrained Rock-Paper-Scissor and Battlefield) as future benchmarking for validation of the scalability of different MARL algorithms; 3) demonstration of MF-MAPPO's superior efficiency and performance over existing MARL algorithms through comprehensive numerical experiments.

## 2 RELATED WORK

Recent advances in learning mean-field games have resulted in several model-free reinforcement learning algorithms that span from Q-function based policy gradients to value-function based policy optimization techniques. Reference [24] proposes MF-Q and MF-AC that parameterize the mean-field Q-function by a neural network. However, the mean-field approach proposed in [24] differs substantially from our formulation, as it defines the mean-field over neighboring actions rather than over the entire state-space.

Along the lines of Q-function based MARL, reference [19] proposed an extension of the original Deep Deterministic Policy Gradient (DDPG) algorithm [10], called DDPG-MFTG, to prescribe a team-level policy based on mean-field observations of all teams. We adopt DDPG-MFTG as a baseline and demonstrate that our proposed method consistently outperforms it, both in terms of stability and performance. Notably, DDPG-MFTG has been primarily evaluated in simple grid-world environments with team-decoupled transition dynamics. It has not been evaluated in environments with tightly coupled mean-field interactions or strict collaborative-competitive scenarios like zero-sum games, where a balance between competition and coordination is paramount. Our work, in

*Proc. of the Adaptive and Learning Agents Workshop (ALA 2025), Avalos, Aydeniz, Müller, Mohammedalamen (eds.), May 19 – 20, 2025, Detroit, Michigan, USA, ala-workshop.github.io.* 2025.

contrast, demonstrates robustness and superior performance under precisely these settings.

Reference [25] shows that policy mirror descent (PMD) along with Temporal Difference (TD) learning converges to an approximate Nash Equilibrium of an $N$-player finite horizon dynamic game (FH-DG). This is another instance of Q-function based learning. Although their approach is more general in terms of the reward structure and heterogeneity among agents, the analysis is limited to mean-field games and excludes mixed collaborative-competitive team games. Alternatively, [23] introduced ECA-Net, a GAN-based method to solve a differential game between two adversarial teams of cooperative players in an attack-defense situation. However, their work focuses on a continuous state-action space setting while we consider an MDP-type problem formulation with strictly conflicting zero-sum rewards.

Closest to the proposed algorithm is the paper [2], where the authors introduced a PPO-based algorithm for constructing optimal policies in the context of mean-field control. However, [2] requires the a-priori knowledge of a mapping from the high-level policy to the agent policies which adds an additional computational step to the algorithm; this is in contrast to our method that directly trains MFTGs using a single identical team policy.

Unlike the methods discussed above, our algorithm is centered around a value function based approach and builds upon the successes of PPO and MA-PPO algorithms [17, 26], and extends those to the competitive mean-field team setting. Following the standard PPO architecture, we do not provide the agents' actions to the critic network, which greatly reduces the size of the neural network. In fact, as we will show later on, it is sufficient to consider a critic network with just the team distributions (mean-fields) as the sole inputs to the network. This also makes the value function network independent of the number of agents, thereby leading to better scalability. As MF-MAPPO employs a shared actor and critic for each team, a single buffer per team suffices for storing team data, thus reducing memory usage and streamlining experience collection without sacrificing the performance of the learned policy. Another key feature of our proposed algorithm is the simultaneous training of both competing teams. Unlike iterative best-response methods [9, 20], which involve alternate policy updates, simultaneous training allows both teams to adapt to each other's most recent policies more dynamically.

# 3 PROBLEM FORMULATION

## 3.1 Zero-Sum Mean-Field Team Game

The zero-sum mean-field team game models a discrete-time stochastic game between two large teams of agents [5]. The Blue and Red teams consist of $N_1$ and $N_2$ *identical* agents for each team, with the total number of agents $N = N_1 + N_2$. Let $X_{i,t}^{N_1} \in \mathcal{X}$ and $U_{i,t}^{N_1} \in \mathcal{U}$ represent the state and action of Blue agent $i \in [N_1]$ at time $t$. Here, $\mathcal{X}$ and $\mathcal{U}$ are the finite state and action spaces of the Blue team. Similarly, $Y_{j,t}^{N_2} \in \mathcal{Y}$ and $V_{j,t}^{N_2} \in \mathcal{V}$ denote the state and action of Red agent $j \in [N_2]$. The joint state-action variables for the Blue and Red teams are denoted as $(\mathbf{X}_t^{N_1}, \mathbf{U}_t^{N_1})$ and $(\mathbf{Y}_t^{N_2}, \mathbf{V}_t^{N_2})$, respectively. Here, we use uppercase letters to denote random variables (e.g., $X$, $\mathcal{M}$) and lowercase letters to denote their realizations (e.g., $x$, $\mu$).

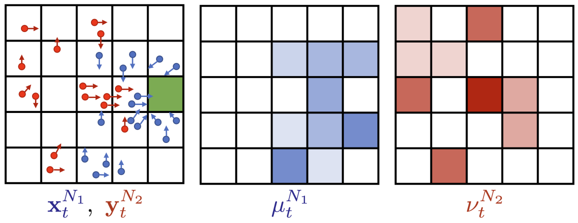

**Figure 1: Battlefield Scenario as an example of ZS-MFTG.**

For a set $E$, we denote the space of probability measures over $E$ as $\mathcal{P}(E)$.

*Definition 3.1.* The *empirical distributions* (ED) for the Blue and Red teams are defined as

$$\mathcal{M}_t^{N_1}(x) = \frac{1}{N_1}\sum_{i=1}^{N_1} \mathbb{1}_x(X_{i,t}^{N_1}), \quad x \in \mathcal{X}, \tag{1a}$$

$$\mathcal{N}_t^{N_2}(y) = \frac{1}{N_2}\sum_{j=1}^{N_2} \mathbb{1}_y(Y_{j,t}^{N_2}), \quad y \in \mathcal{Y}, \tag{1b}$$

where $\mathbb{1}_a(b) = 1$ if $a = b$ and 0 otherwise. Specifically, $\mathcal{M}_t^{N_1}(x)$ gives the fraction of Blue agents at state $x$ and similarly for $\mathcal{N}_t^{N_2}(y)$. We use $\mathcal{M}_t^{N_1} = \mathrm{Emp}_\mu(\mathbf{X}_t^{N_1})$ and $\mathcal{N}_t^{N_2} = \mathrm{Emp}_\nu(\mathbf{Y}_t^{N_2})$ to denote the EDs computed from the given joint states. Note that the Emp operators remove agent index information, so one *cannot* determine the state of a specific Blue agent $i$ from $\mathcal{M}_t^{N_1}$.

We consider weakly-coupled dynamics where the dynamics of each individual agent is coupled with other agents through the EDs [4]. For Blue agent $i$, its stochastic transition is governed by the transition kernel $f_t : \mathcal{X} \times \mathcal{U} \times \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y}) \to \mathcal{P}(\mathcal{X})$ so that

$$\begin{aligned}\mathbb{P}(X_{i,t+1}^{N_1} = x_{i,t+1}^{N_1} | U_{i,t}^{N_1} = u_{i,t}^{N_1}, \mathbf{X}_t^{N_1} = \mathbf{x}_t^{N_1}, \mathbf{Y}_t^{N_2} = \mathbf{y}_t^{N_2}) \\ = f_t(x_{i,t+1}^{N_1} | x_{i,t}^{N_1}, u_{i,t}^{N_1}, \mu_t^{N_1}, \nu_t^{N_2}),\end{aligned} \tag{2}$$

where $\mu_t^{N_1} = \mathrm{Emp}_\mu(\mathbf{x}_t^{N_1})$ and $\nu_t^{N_2} = \mathrm{Emp}_\nu(\mathbf{y}_t^{N_2})$. Similarly, the dynamics of Red agent $j$ is governed by the transition kernel $g_t : \mathcal{Y} \times \mathcal{V} \times \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y}) \to \mathcal{P}(\mathcal{Y})$. All agents in the Blue team receive an identical weakly-coupled team reward, i.e., $r_t \triangleq r_t(\mu_t, \nu_t) : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y}) \to \mathbb{R}$. Following the zero-sum structure, the Red team agents receive $-r_t(\mu_t, \nu_t)$ as their rewards. We assume that the Blue (Red) team is the maximizing (minimizing) team.

Figure 1 depicts a battlefield scenario of an MFTG between two teams (Blue and Red) on an $n \times n$ grid world.

## 3.2 Large-Population Optimization

We consider a mean-field sharing information structure [1], where each agent observes its own state and the two team EDs, where the EDs serve as common information accessible to both teams. Specifically, the Blue and Red agents seek to construct mixed Markov policies with the following structure

$$\phi_{i,t} : \mathcal{U} \times \mathcal{X} \times \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y}) \to [0, 1], \tag{3a}$$

$$\psi_{j,t} : \mathcal{V} \times \mathcal{Y} \times \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y}) \to [0, 1], \tag{3b}$$

where the Blue policy $\phi_{i,t}(u | x_{i,t}^{N_1}, \mu_t^{N_1}, \nu_t^{N_2})$ dictates the probability that Blue agent $i$ selects action $u$ given its state $x_{i,t}^{N_1}$ and the observed team EDs $\mu_t^{N_1}$ and $\nu_t^{N_2}$. Note that each agent's individual state is its private information, while the team EDs are the common information available to all $N$ agents.

Let $\Phi_t$ ($\Psi_t$) denote the set of individual Blue (Red) policies at time $t$. We define the Blue team policy $\phi_t^{N_1} = \{\phi_{i,t}\}_{i=1}^{N_1}$ as the collection of the $N_1$ Blue agent individual policies, and denote the set of Blue team policies as $\Phi_t^{N_1} = \times_{N_1} \Phi_t$. Similarly, the Red team policy is denoted as $\psi_t^{N_2} \in \Psi_t^{N_2} = \times_{N_2} \Psi_t$.

*Definition 3.2 (Identical team policy).* The Blue team policy $\phi_t^{N_1} = (\phi_{1,t}^{N_1}, \dots, \phi_{N_1,t}^{N_1})$ is an *identical*, if $\phi_{i_1,t} = \phi_{i_2,t}$ for all time $t$ and $i_1, i_2 \in [N_1]$. We denote the set of identical Blue team policies as $\Phi$.

The definition and notation extend naturally to the Red team, and the set of identical Red team policies is denoted as $\Psi$.

The performance of the team policy pair $(\phi^{N_1}, \psi^{N_2})$ is given by the expected cumulative reward

$$J^{N,\phi^{N_1},\psi^{N_2}}(\mathbf{x}_0^{N_1}, \mathbf{y}_0^{N_2}) = \mathbb{E}_{\phi^{N_1},\psi^{N_2}}\Big[\sum_{t=0}^{T} r_t(\mathcal{M}_t^{N_1}, \mathcal{N}_t^{N_2})\Big|\mathbf{x}_0^{N_1}, \mathbf{y}_0^{N_2}\Big].$$

When the Blue team considers its worst-case performance, we have the following max-min optimization:

$$\underline{J}^{N*}(\mathbf{x}_0^{N_1}, \mathbf{y}_0^{N_2}) = \max_{\phi^{N_1} \in \Phi^{N_1}} \min_{\psi^{N_2} \in \Psi^{N_2}} J^{N,\phi^{N_1},\psi^{N_2}}(\mathbf{x}_0^{N_1}, \mathbf{y}_0^{N_2}), \quad (4)$$

where $\underline{J}^{N*}$ is the lower game value for the finite-population game. Similarly, the minimizing Red team considers a min-max optimization problem, which leads to the upper game value. Note that we allow both teams to follow *non-identical* team policies in (4).

### 3.3 Infinite-Population Solution

To reduce the complexity of team policy optimization domains in (4), the authors of [4] proposed to examine team behaviors under *identical team policies* at the *infinite-population* limit. It was shown that the team joint states can be represented using the team population distribution, which coincides with the state distribution of a *typical agent*. Such distributions are referred to as the mean-fields (MFs), and we denoted them as $\mu_t$ and $\nu_t$ for the Blue and Red teams, respectively. As proved in [4], MFs induced by identical team policies in an infinite-population game closely approximate the EDs induced by *non-identical* team policies in the corresponding finite-population game, which justifies the simplification of the optimization domain in (4) to identical team policies.

For the infinite-population game, the performance of the *identical* team policies $(\phi, \psi) \in \Phi \times \Psi$ is measured by

$$J^{\phi,\psi}(\mu_0, \nu_0) = \sum_{t=0}^{T} r_t(\mu_t, \nu_t), \quad (5)$$

where $\mu_t$ and $\nu_t$ follow a *deterministic* dynamics [4] similar to the state distribution propagation of a controlled Markov chain. The worst-case performance of the Blue team in this infinite-population game is then given by the lower game value

$$\underline{J}^*(\mu_0, \nu_0) = \max_{\phi \in \Phi} \min_{\psi \in \Psi} J^{\phi,\psi}(\mu_0, \nu_0), \quad (6)$$

where the optimization domain is restricted to identical team policies. Reference [4] exploited the simplified optimization domain in (6) and proposed to transform the optimization to an equivalent zero-sum game between two fictitious coordinators. The optimal identical team policies can then be solved via dynamic programming. The following performance guarantees was established in [4].

Theorem 3.3. *The optimal identical Blue team policy $\phi^*$ obtained from the equivalent zero-sum coordinator game is an $\epsilon$-optimal Blue team policy. Formally, for all joint states $\mathbf{x}^{N_1}$ and $\mathbf{y}^{N_2}$,*

$$\min_{\psi^{N_2} \in \Psi^{N_2}} J^{N,\phi^*,\psi^{N_2}}(\mathbf{x}^{N_1}, \mathbf{y}^{N_2}) \geq \underline{J}^{N*}(\mathbf{x}^{N_1}, \mathbf{y}^{N_2}) - O\Big(\frac{1}{\sqrt{\underline{N}}}\Big) \quad (7)$$

*where $\underline{N} = \min\{N_1, N_2\}$.*

This result ensures that identical team policies resulting from the solution of the equivalent zero-sum coordinator game are still $\epsilon$-optimal for the original max-min optimization problem in (4). From Theorem 3.3 we can further show that the performance of the optimal identical policy learned from the finite population ZS-MFTG remains within an $\epsilon$-bound of the identical policies derived from the optimal coordinator game.

Theorem 3.4. *The value of the optimal identical Blue team policy $\phi^*_{\text{finite}}$ obtained from the finite population game is within $\epsilon$ of the value of the optimal identical Blue team policy $\phi^*$ obtained from the equivalent zero-sum coordinator game. Formally, for all joint states $\mathbf{x}^{N_1}$ and $\mathbf{y}^{N_2}$,*

$$\min_{\psi^{N_2}} J^{N,\phi^*_{\text{finite}},\psi^{N_2}}(\mathbf{x}^{N_1}, \mathbf{y}^{N_2}) - \min_{\psi^{N_2}} J^{N,\phi^*,\psi^{N_2}}(\mathbf{x}^{N_1}, \mathbf{y}^{N_2}) \leq \epsilon, \quad (8)$$

*where $\epsilon = O(1/\sqrt{\underline{N}})$ and $\underline{N} = \min\{N_1, N_2\}$.*

The first figure depicts the individual agents' local positions, with the target marked by the green colored cell. The subsequent figures illustrate the state distributions $\mu_t^{N_1}$ and $\nu_t^{N_2}$ of both teams, which constitute the common information available to the agents of both teams. The agents interact based on weakly-coupled dynamics, which depend only on $\mu_t^{N_1}$ and $\nu_t^{N_2}$ as described in (2). In this typical scenario of an MFTG, each agent takes its action after observing its own local position and the common (i.e., mean-field) information, in order to achieve its own team's objective.

## 4 MEAN-FIELD MULTI-AGENT PROXIMAL POLICY OPTIMIZATION

Motivated by Theorem 3.3, we present an algorithm to learn the optimal identical team policy. We build our algorithm based on the proximal policy optimization (PPO) framework due to its simplicity and effectiveness. While PPO has shown promising performance in cooperative tasks including mean-field control problems [2, 26], its application in mixed competitive-collaborative scenarios is less studied, especially in the MFTG settings. In the sequel, we introduce our key contribution: the MFTG learning algorithm, which we refer to as Mean-Field Multi-Agent Proximal Policy Optimization (MF-MAPPO).

We initialize two pairs of actor-critic networks, one for each team, deployed to learn the identical policy used by each team (see Figure 2). Specifically, we introduce a *minimally-informed critic* network by exploiting the shared mean-field information. The key point to note here is that we only require the common information for the critic network in order to learn the value function (Proposition 4.1). Furthermore, the private information available to each agent only *individually* enters the actor during training. This results in neural networks that scale well with the number of agents.

## 4.1 Minimally-Informed Critic

The MF-MAPPO critic network of the Blue team evaluates the value function $V_{\text{Blue}}(\mu, \nu)$, which depends only on the common information, and is *independent* of the joint agent states and actions. We use the parameter vector $\zeta_{\text{Blue}}$ to parameterize the critic network while minimizing the MSE loss

$$L_{\text{critic}}(\zeta_{\text{Blue}}) = \frac{1}{|B|}\sum_{k=1}^{|B|}\Big(V_{\text{Blue}}(\mu_k, \nu_k|\zeta_{\text{Blue}}) - \hat{R}_{\text{Blue},k}\Big)^2, \qquad (9)$$

where the optimization is performed over a mini-batch of size $B$ and $\hat{R}_{\text{Blue},k}$ is the discounted reward-to-go for sample $k$. The reward-to-go for sample $k$ obtained at a time step $t$ is computed using Monte-Carlo roll-outs starting at $t$ until the episode ends, and is given by $\hat{R}_{\text{Blue},k} = \sum_{t'=t}^{T} \gamma^{t'-t} r_t(\mu_t, \nu_t)$ [22]. Similar learning rules apply to the Red team critic $V_{\text{Red}}(\mu, \nu|\zeta_{\text{Red}})$ with the negative reward $\hat{R}_{\text{Red},k} = -\hat{R}_{\text{Blue},k}$ due to the zero-sum structure.

The following proposition follows immediately from the expression of the team reward (5) and the use of identical team policies, and justifies the deployment of a minimally-informed critic network with only the mean-fields as inputs.

Proposition 4.1. *Let $\mu_t^{N_1}$, and $\nu_t^{N_2}$ denote the EDs of a finite-population game obtained from identical Blue and Red team policies $\phi_t \in \Phi_t$ and $\psi_t \in \Psi_t$, respectively. The team reward structure admits a critic that depends only on $\mu_t^{N_1}$ and $\nu_t^{N_2}$. Specifically, for each Blue team agent $i \in \{1, 2, \dots, N_1\}$, the individual critic value function $V_{i,t}^{N_1,\phi_t}(x_{i,t}, \mu_t^{N_1}, \nu_t^{N_2})$ satisfies*

$$V_{i,t}^{N_1,\phi_t}(x_{i,t}, \mu_t^{N_1}, \nu_t^{N_2}) = V_{\text{Blue},t}^{N_1,\phi_t}(\mu_t^{N_1}, \nu_t^{N_2}), \qquad (10)$$

*where $V_{\text{Blue},t}^{N_1,\phi_t}(\mu_t, \nu_t)$ is the team-level critic.*

The above proposition extends to the Red team critic network. Importantly, it reduces the learning problem to one critic network per team. Specifically, the shared team reward structure along with the assumption of homogeneous agents in each team enables us to evaluate the performance of a team's agent using the minimally-informed critic—even if the individual agent has additional local observations such as their actions and private states.

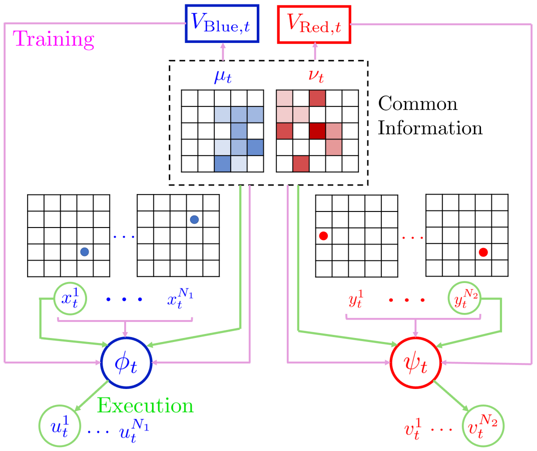

**Figure 2: Overview of the architecture of MF-MAPPO.**

## 4.2 Shared-Team Actor

We consider a single-actor network for each team to learn the identical team policies. According to [4], identical policies derived from an equivalent coordinator game can approximate team behaviors induced by non-identical ones. A single coordinator policy corresponds to the probability distribution over actions for each state, conditioned on a given mean-field. As a result, its dimensionality scales with the joint state-action space, leading to substantial computational overhead and degraded empirical performance of the policy network. Notably, DDPG-MFTG adopts this formulation, but, as demonstrated later in the numerical examples section, its performance is suboptimal.

Furthermore, while the coordinator game serves as a valuable theoretical construct, we found it more practical and computationally efficient to learn finite-population local identical policies in place of coordinator policies. These policies continue to respect the underlying mean-field information-sharing structure, while significantly reducing complexity and improving tractability in practice. Theorem 3.4 ensures performance guarantees.

The actor network maximizes a PPO-based objective with an entropy term to encourage exploration [7, 17], which decays during training as teams learn reward-maximizing policies. It has also been shown in the mean-field game literature that entropy regularization stabilizes the learning process [3, 5]. Since the agents are permutation invariant and an identical policy is being learned for each team, a single buffer per team suffices for storing observations and actions, thereby reducing memory overhead and simplifying the experience collection pipeline.

The PPO-based objective function of the Blue actor is given by:

$$L(\theta_{\text{Blue}}) = \frac{1}{|B|}\sum_{k=1}^{|B|}\Big[\min\Big(g_k(\theta_{\text{Blue}})A_k, \text{clip}_{[1-\epsilon,1+\epsilon]}(g_k(\theta_{\text{Blue}}))A_k\Big) + \omega S(\phi_{\theta_{\text{Blue}}}(x_k, \mu_k, \nu_k))\Big], \qquad (11)$$

where,

$$g(\theta) = \frac{\phi_\theta(u|x, \mu, \nu)}{\phi_{\theta^{\text{old}}}(u|x, \mu, \nu)},$$

and $A_k$ is the generalized advantage function estimate function [16]. The tunable parameter $\omega$ weighs the contribution of the entropy term, which is given by

$$S\big(\phi_\theta(x, \mu, \nu)\big) = -\mathbb{E}_{u\sim\mathcal{U}}\Bigg[\sum_u \phi_\theta(u|x, \mu, \nu)\log\phi_\theta(u|x, \mu, \nu)\Bigg].$$

We decay $\omega$ as training progresses. A similar learning rule is used for the Red team actor network. Algorithm 1 presents the pseudo-code of the MF-MAPPO algorithm.

**Algorithm 1** Mean-Field Multi-Agent Proximal Policy Optimization (**MF-MAPPO**)

```
Initialize: NN parameters $\{\theta_{\text{Blue}}, \zeta_{\text{Blue}}\}$ and $\{\theta_{\text{Red}}, \zeta_{\text{Red}}\}$; step size sequences $\{\alpha_m\}$ and $\{\beta_m\}$; entropy decay sequence $\{\omega_m\}$
for $i = 1, 2, \ldots$ do
    $(\phi_{\theta^{\text{old}}_{\text{Blue}}}, \psi_{\theta^{\text{old}}_{\text{Red}}}) \leftarrow (\phi_{\theta_{\text{Blue}}}, \psi_{\theta_{\text{Red}}})$
    for $t = 0, 1, \ldots, T_{\text{rollout}}$ do
        Sample joint actions
        $u_{i,t} \sim \phi_{\theta^{\text{old}}_{\text{Blue}}}(x_{i,t}, \mu^{N_1}_t, \nu^{N_2}_t)$, $v_{j,t} \sim \psi_{\theta^{\text{old}}_{\text{Red}}}(y_{j,t}, \mu^{N_1}_t, \nu^{N_2}_t)$
        Step environment according to kernels $(f_t, g_t)$
        Collect samples $(\mathbf{x}^{N_1}_{t+1}, \mathbf{y}^{N_2}_{t+1}, \mu^{N_1}_{t+1}, \nu^{N_2}_{t+1}, \mathbf{u}^{N_1}_t, \mathbf{v}^{N_2}_t, r_t)$
    end for
    for $K$ epochs do
        Update $\{\theta_{\text{Blue}}, \zeta_{\text{Blue}}\}$ and $\{\theta_{\text{Red}}, \zeta_{\text{Red}}\}$ using (9-11)
    end for
end for
Return: $(\phi_{\theta_{\textbf{Blue}}}, \psi_{\theta_{\textbf{Red}}})$
```

### 4.3 Scalability of MF-MAPPO

We further demonstrate the scalability of MF-MAPPO as a direct consequence of Theorem 3.3, by showing that, under certain conditions, the learned team policies generalize to varying population sizes $(\tilde{N}_1, \tilde{N}_2)$ while maintaining performance guarantees.

Remark 1. *Let $\mathcal{G}_1$ denote the finite-population game where the agents utilize the identical team policies $\phi^*_t$ and $\psi^*_t$ derived from the equivalent, infinite-population, zero-sum coordinator game, and let the finite-population game $\mathcal{G}_2$ with the same state-action space, dynamics, and rewards, but with population sizes $\tilde{N}_1$ and $\tilde{N}_2$ such that $\tilde{N}_1/\tilde{N}_2 = N_1/N_2$ and $\min(\tilde{N}_1, \tilde{N}_2) \geq \min(N_1, N_2)$. Then, the policies $\phi^*_t$ and $\psi^*_t$ remain $\epsilon$-optimal for the game $\mathcal{G}_2$.*

Remark 1 describes policies from the equivalent coordinator game. Empirically, we show that identical team policies from the finite-population ZS-MFTG (Theorem 3.4) yield similar results. This allows MF-MAPPO to be trained on a smaller population and deployed to larger teams without additional tuning, significantly reducing computational costs while maintaining performance consistency and generalizability across different population sizes.

## 5 NUMERICAL EXPERIMENTS

In this section, we present several large-population scenarios to demonstrate the efficacy of MF-MAPPO. The first two scenarios are mean-field extensions of the rock-paper-scissors game [14] with different action spaces. For these examples, we can analytically compute the mean-field trajectory induced by the equilibrium/optimal policies, and thus use these scenarios to validate the optimality of MF-MAPPO. We then present a more complex battlefield scenario where the Blue and Red teams play an attack-defense game as shown in Figure 1. This scenario has higher (finite) dimensional state and action spaces and the teams are required to learn more complex collective behaviors.

### 5.1 Rock-Paper-Scissors (RPS)

We first extend the two-player Rock-Paper-Scissors (RPS) game [14] to a game played between two populations. The state space of each individual agent is $\mathcal{S} = \{\texttt{R}, \texttt{P}, \texttt{S}\}$, representing rock, paper, and scissors, respectively. Let $\mu, \nu \in \mathcal{P}(\mathcal{S})$ denote the EDs for the Blue and Red teams. Following the mean-field sharing information structure, an agent observes its local state and the EDs of both teams. The action space, $\mathcal{A} = \{\texttt{CW}, \texttt{CCW}, \texttt{Stay}\}$, allows agents to either move clockwise, counter-clockwise, or remain idle, respectively.

We assume deterministic transitions, where each action leads to a unique next state *deterministically*. For example, if an agent at R takes action CW, it will deterministically end in state P, as shown in Figure 3. At each time step $t$, the Blue team receives a team reward $r_t(\mu_t, \nu_t) = \mu_t^\intercal A \nu_t$ where $A$ is the standard RPS payoff matrix given by $A = [0, -1, 1; 1, 0, -1; -1, 1, 0]$. We let the Blue team maximize the expected cumulative reward while the Red team minimizes it. The Nash equilibrium for this population-based RPS game is the uniform population distribution $[1/3, 1/3, 1/3]$ over the 3 states [13, 14].

We compare MF-MAPPO with DDPG-MFTG [19] based on the training time, average test rewards and attainment of the computed Nash distributions for $N_1 = N_2 = 1,000$ agents. A key distinction between the two algorithms lies in their design philosophy: DDPG-MFTG relies on a mean-field oracle that provides the next mean-field given the current team policy—an object that exists only in the infinite population limit and cannot be directly simulated. In contrast, MF-MAPPO is trained directly within a simulated finite-population environment. DDPG-MFTG introduces “central players” that observe mean-field distributions and output deterministic local policies (akin to the role of the coordinator) via a Q-function, following the standard DDPG architecture. Crucially, in DDPG-MFTG, the input to the Q-function for a given team consists of the mean-field information of all teams, along with only the local policy of the team itself. This contrasts with multi-agent extensions of DDPG (e.g., MADDPG), which also incorporates local policy information of other teams.

Furthermore, the DDPG-MFTG policy is updated at every time step post-exploration without any explicit clipping or regularization mechanisms to constrain policy updates. This lack of stabilization—combined with the high computational complexity and limited inter-team policy awareness—contributes to the algorithm's training instability and poor generalization, especially in complex environments as we discuss below.

We exclude MADDPG [11] from our comparison, as it scales poorly to hundreds or thousands of agents due to its reliance on all

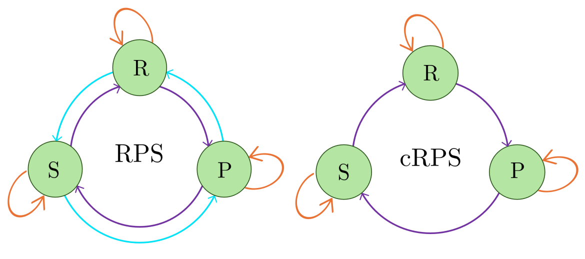

**Figure 3: States and actions for RPS and cRPS.**

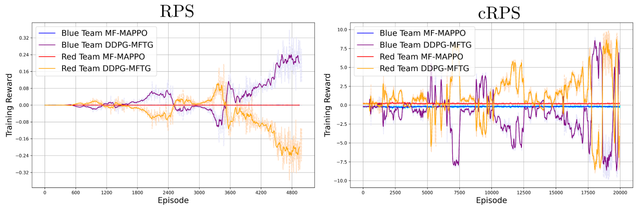

**Figure 4: Training reward curves for RPS and cRPS.**

**Table 1: Performance Comparison for RPS**

| Approach | Training Time | Average Reward | NE Attained? |
|---|---|---|---|
| MF-MAPPO | 5min 17s | 0.0 | ✓ |
| DDPG-MFTG | 1min 34s | 0.334 | ✗ |

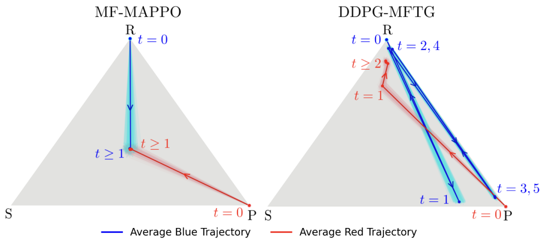

**Figure 5: ED trajectories induced by learned team policies on the state distribution simplex. Mean trajectories are averaged based on the 150 runs from fixed initialization $\mu_{t=0} = [1, 0, 0]^\top$ and $\nu_{t=0} = [0, 1, 0]^\top$; $N_1 = N_2 = 1,000$.**

agents' local and global observations and actions as inputs to its critic networks.

From the learning curves in Figure 4 one can see that the DDPG-MFTG algorithm failed to converge to the analytical game value of zero, while MF-MAPPO almost immediately attained the Nash game value. However, as shown in Table 1, MF-MAPPO does take slightly longer to train since, unlike DDPG-MFTG, since MF-MAPPO avoids mini-batch training, following [26].

We tested the learned policy with a fixed initial distribution $\mu_{t=0} = [1, 0, 0]^\top$ and $\nu_{t=0} = [0, 1, 0]^\top$, and the resulting trajectories are visualized in Figure 5. All simulations were run for 150 instances. The trajectories of the Blue and Red team ED are depicted in cyan and pink, respectively, alongside the mean trajectory. The randomness in these trajectories arises from the finite-population approximation under a stochastic optimal policy, resulting in stochastic EDs. As shown in Figures 4 and 5, DDPG-MFTG diverges from the Nash equilibrium whereas MF-MAPPO converges immediately.

## 5.2 Constrained Rock-Paper-Scissors (cRPS)

We now consider a non-trivial modification to the RPS problem by restricting the action space to $\mathcal{A} = \{\texttt{CW}, \texttt{Stay}\}$. As shown in Figure 3, with this restricted action space, the teams cannot immediately achieve their desired uniform distribution and need to strategically plan for the intermediate distributions before reaching the desired target uniform distribution. We again consider the fixed initial distributions $\mu_{t=0} = [1, 0, 0]^\top$ and $\nu_{t=0} = [0, 1, 0]^\top$. One may obtain analytically the conditions for the optimal trajectories of the cRPS game for these initial conditions.

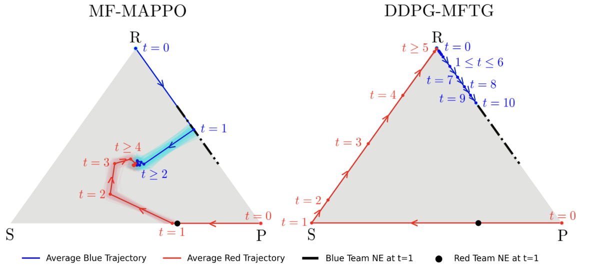

**Figure 6: 150 initializations (cyan/pink) of $\mu_{t=0} = [1, 0, 0]^\top$ and $\nu_{t=0} = [0, 1, 0]^\top$ on a 3D simplex for cRPS for $N_1 = N_2 = 1,000$.**

Proposition 5.1. *With initial conditions $\mu_{t=0} = [1, 0, 0]^\top$ and $\nu_{t=0} = [0, 1, 0]^\top$, all mean-field optimal trajectories satisfy $\mu^*_t = \nu^*_t = [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]^\top$ for all $t \geq 2$, and $\mu^*_1 = [0, 1-\eta, \eta]^\top$ where $\eta \in [\frac{1}{3}, \frac{2}{3}]$ and $\nu^*_1 = [0, \frac{2}{3}, \frac{1}{3}]^\top$. Furthermore, the unique game value is given by $-\frac{1}{3}$.*

Figure 6 shows the trajectories over ten time steps induced by the team policies learned by MF-MAPPO and DDPG-MFTG. Using MF-MAPPO, the teams successfully reach the uniform distribution from the initial condition in Proposition 5.1, as shown in Figure 6, while the DDPG-MFTG's trajectories diverge. For MF-MAPPO, the teams follow the optimal trajectory at $t = 1$, and the target optimal distribution is reached within three to four time steps. The transient time can be attributed to the finite population approximation and the entropy term in the optimization problem. Despite this, the Blue team's game value after training (-0.331) closely matches the theoretical value $-1/3$. Figure 7 shows trajectories for various population sizes using identical policies trained on $N_1 = N_2 = 1,000$ with MF-MAPPO. As the population size increases, the trajectories exhibit reduced noise and variance while maintaining strong performance, consistent with the scalability result in Remark 1.

Table 2 highlights the training times between MF-MAPPO and DDPG-MFTG for the same number of episodes. DDPG-MFTG updates its networks at every time step, causing its computational overhead to scale with the episode length, whereas MF-MAPPO updates every $T_{\text{rollout}}$ steps independently of the episode length. Consequently, DDPG-MFTG trains significantly slower on cRPS despite using mini-batch updates.

These experiments show that MF-MAPPO learns stable policies that are consistent with the theoretical predictions, while DDPG-MFTG, despite its successes in [19], struggles to stabilize in these simple ZS-MFTGs.

**Table 2: Performance Comparison for cRPS**

| Approach | Training Time | Average Reward | NE Attained? |
|---|---|---|---|
| MF-MAPPO | 2h 17min 15s | -0.331 | ✓ |
| DDPG-MFTG | 60h 49min 41s | 3.774 | ✗ |

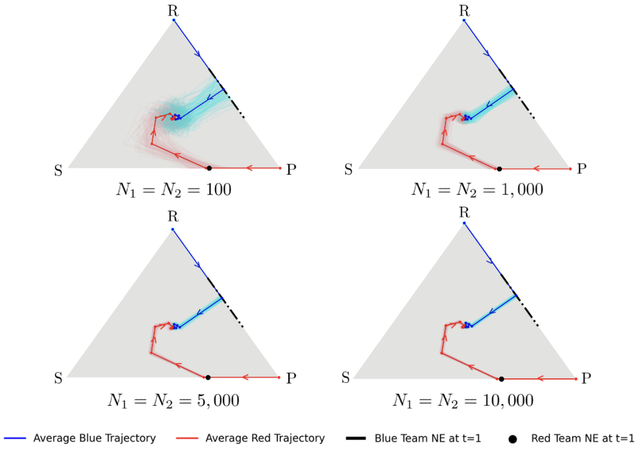

**Figure 7: Deploying MF-MAPPO trained on $N_1 = N_2 = 1,000$ to varying team sizes (150 trajectories) with the same initialization.**

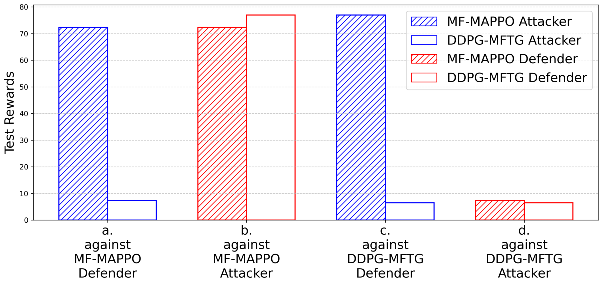

**Figure 8: Average test rewards of MF-MAPPO vs. DDPG-MFTG on a 4x4 grid world for 100 random initializations.**

## 5.3 Battlefield Game

To fully test the capability of MF-MAPPO on a more complex scenario, we propose a battlefield game where an individual agent's dynamics is highly coupled with both teams' distributions.

We consider a large-scale two-team (Blue and Red) ZS-MFTG on an $n \times n$ grid world, modeling a target capture-type offense/defense scenario. The Blue agents aim to reach target locations without being deactivated, while the Red agents must learn to guard these targets. Each team can deactivate its opponents in a cell by maintaining numerical advantage over the opposing team at that cell. In all battlefield illustrations, the targets are shown in lilac color and the obstacles are shown in black; the bottom left cell is $[0, 0]$.

*5.3.1 Problem Setup and Objective.* The state of the $i^{th}$ Blue agent is defined as the pair $x_i = (p_i^x, s_i^x)$ where $p_i^x \in \mathcal{S}_{\text{position}}$ denotes the position of the agent in the grid world and $s_i^x \in \mathcal{S}_{\text{status}} = \{0, 1\}$ defines the status of the agent: 0 being inactive and 1 being active. Similarly, we define the state of the Red agent as $y_i = (p_i^y, s_i^y)$. The state spaces for the Blue and Red teams are denoted by $\mathcal{X} = \mathcal{Y} = \mathcal{S}_{\text{position}} \times \mathcal{S}_{\text{status}}$, respectively. The mean-fields of the Blue ($\mu$) and Red ($\nu$) teams are distributions over the joint position and status space, i.e., $\mu, \nu \in \mathcal{P}(\mathcal{S}_{\text{position}} \times \mathcal{S}_{\text{status}})$. The action spaces are given by $\mathcal{U} = \mathcal{V} = \{\texttt{Up}, \texttt{Down}, \texttt{Left}, \texttt{Right}, \texttt{Stay}\}$ for both teams, representing discrete movements in the grid world. The learned identical team policy assigns actions based on an agent's local position and status, as well as the observed mean-fields of both teams.

An agent can be deactivated by the opponent with a nonzero probability if the opponent's ED at the agent's location exceeds that of the agent's own team, which we refer to as the *numerical advantage*. The total transition probability from state $(p, s)$ to state $(p', s')$ by taking an action $a$ is given by

$$\mathbb{P}\big((p', s') \mid (p, s), a, \mu, \nu\big) = \mathbb{P}\big(p' \mid (p', s'), a\big)\, \mathbb{P}\big(s' \mid (p, s), \mu, \nu\big),$$

where the first term on the right-hand side corresponds to the deterministic position transition when the agent is active. The second term corresponding to the status transition is given by

$$\mathbb{P}\big(0 \mid (p, 1), \mu, \nu\big) = \text{clip}_{[0,1]}\big(\alpha_x(\nu(p) - \mu(p))\big),$$
$$\mathbb{P}\big(1 \mid (p, 1), \mu, \nu\big) = 1 - \mathbb{P}\big(0 \mid (p, 1), \mu, \nu\big),$$

where $\nu(p) - \mu(p)$ is the Red team's numerical advantage over the Blue team at $p$, and $\alpha_x$ is a tuning parameter to control the Red team's deactivation power. We can formulate a similar expression for the deactivation probability of the Red team based on the Blue team's numerical advantage $\mu(p) - \nu(p)$.

These dynamics incentivize both teams to aggregate, i.e., to minimize the numerical advantage of the opponent team and ultimately reduce the risk of being deactivated. The Blue team's reward depends on the fraction of agents active and at the target, and Red's reward follows from the zero-sum structure. To avoid degeneracy, the Red agents are not allowed to enter the target.

*5.3.2 Results and Discussion.* We conducted several experiments on various grid world maps with different target and obstacle layouts. In our experiments, we trained with $N_1 = N_2 = 100$ agents and experimented with different initial distributions. Additional results are presented in the supplementary material.

*Map 1:* A $4 \times 4$ grid features a target partially obstructed by a diagonal obstacle (Figures 9-10). The Red team must position itself along left and the right corridors (cells $[0, 2]$ and $[1, 3]$) to hinder Blue team's advance. We compare MF-MAPPO and DDPG-MFTG by pitting them against each other in both offensive and defensive roles. As shown in Figure 8, MF-MAPPO consistently outperforms DDPG-MFTG across various initial positions, achieving up to 10× higher rewards when attacking (Figures 8(a) and 8(c)). While both methods perform similarly in defense, visualizations reveal that DDPG-MFTG agents often fail to learn effective strategies, instead remaining stationary and aiming for a zero-reward outcome—an exploitable weakness.

Figure 9 compares the two algorithms against the baseline defending team. MF-MAPPO Blue agents exhibit coordinated maneuvering, forming coalitions to reach the target (a), whereas DDPG-MFTG Blue agents (b) show limited coordination, with only nearby agents reaching the target and distant agents failing to engage. Figure 9(c) represents MF-MAPPO vs. MF-MAPPO to illustrate the goal strategies and expected behavior.

We next evaluated MF-MAPPO's performance as the Red defending team against the DDPG-MFTG Blue team (Figure 10). Due to the agents' initial positions, the target entryway near $[0, 2]$ remains unguarded, requiring the Red team to mobilize its agents to defend

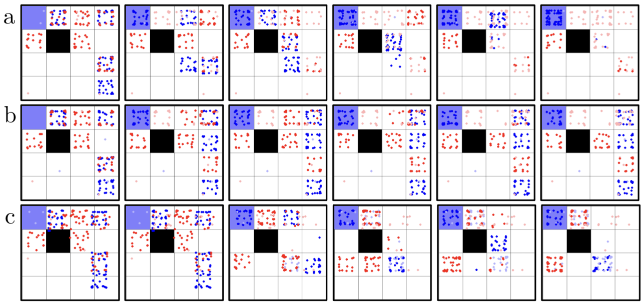

**Figure 9: a. MF-MAPPO Blue vs. DDPG-MFTG Red; b. DDPG-MFTG Blue vs. DDPG-MFTG Red; c. MF-MAPPO Blue vs. MF-MAPPO Red.**

the area. MF-MAPPO Red agents successfully cover the entryway and deactivate several Blue attackers (a), whereas the DDPG-MFTG Red agents (b) fail to secure the second target. Moreover, (b) highlights that DDPG-MFTG Blue agents do not aggressively pursue the target, further illustrating their tendency to passively seek zero reward outcomes rather than take goal-directed actions, unlike (c).

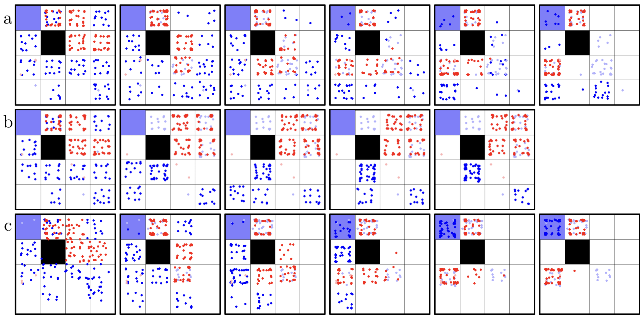

**Figure 10: a. MF-MAPPO Red vs. DDPG-MFTG Blue; b. DDPG-MFTG Red vs. DDPG-MFTG Blue; c. MF-MAPPO Red vs. MF-MAPPO Blue.**

*Map 2:* We design a more complex $8 \times 8$ grid with two targets (Figures 11 and 12). The Red team now faces a dilemma in determining which target to defend, while the Blue team must exploit this ambiguity to its advantage. Due to DDPG-MFTG's high computational cost—which scales with the joint state-action space—and its high network update frequency, it is excluded from our analysis. In the absence of other baselines for such large scale complex games, we only qualitatively assess MF-MAPPO's performance.

In Figure 11, the Red team is concentrated at one cell, while the Blue team is split: 30% at cell [1, 1] and the rest at cell [6, 6]. The Red team proceeds to deactivate the smaller group of Blue agents due to its numerical advantage. Upon seeing both the Blue groups advance towards the upper target, the Red team changes its trajectory in order to block the target as seen in (c). The Blue team then intelligently shifts toward the lower target, as assembling its forces there is more feasible than at the upper target, where the Red team maintains a strong presence (Figures 11(d)-11(g)).

In Figure 12, the Blue team is evenly distributed across four cells, while the Red team is concentrated at [3, 3]. Blue agents move toward the nearest target, demonstrating heterogeneous behavior despite operating under an identical policy—highlighting the strength of the mean-field approximation. Similarly, the Red team also strategically divides to defend both targets its identical policy. Due to policy stochasticity and finite population effects, the split is uneven, prompting the upper Blue subgroup to redirect toward the lower target. This allows both lower Blue subgroups to reach their objective. The Red team quickly reallocates its agents in response and ultimately uses its numerical advantage to deactivate several Blue agents (Figure 12(h)).

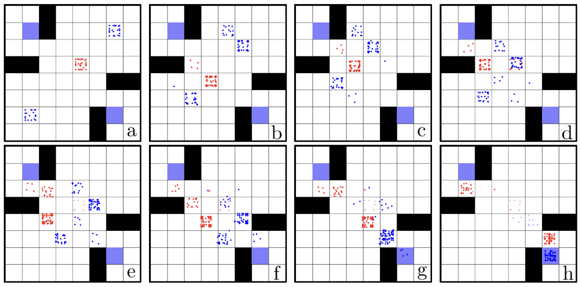

**Figure 11: Red is concentrated; 30% Blue are at $[1, 1]$ and the rest are at $[6, 6]$.**

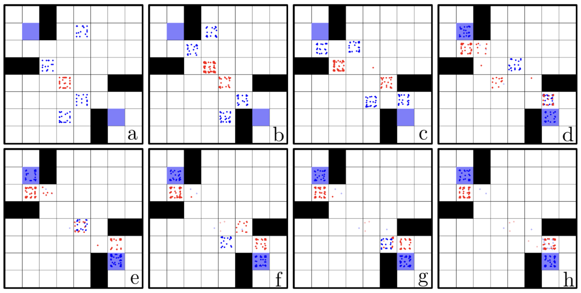

**Figure 12: Blue is evenly split, Red is concentrated.**

## 6 CONCLUSION

We introduced MF-MAPPO, a novel MARL algorithm for large-population competitive team games, leveraging finite mean-field approximation. Our design—featuring a minimally-informed critic and a shared team actor—achieves scalability without sacrificing performance. We evaluated MF-MAPPO against baselines like DDPG-MFTG on standard and constrained RPS, as well as a new MFTG battlefield scenario. Despite shared team policies, heterogeneous sub-population behaviors emerged, showing that mean-field approximations do not significantly limit performance. The battlefield testbed provides a rigorous benchmark for future research, supporting evaluations on accumulated rewards, sample efficiency, and computational complexity. A current limitation is that input dimensionality grows with the state space, which we aim to address through dimensionality reduction techniques (e.g., kernel embeddings).

## ACKNOWLEDGMENTS

This work was supported in part by ONR Award N00014-23-1-2308 and by ARL DCIST CRA W911NF-17-2-0181.

# REFERENCES

[1] Jalal Arabneydi and Aditya Mahajan. 2015. Team-optimal solution of finite number of mean-field coupled LQG subsystems. In *54th IEEE Conference on Decision and Control*. IEEE, Osaka, Japan, Dec. 15–18, 2015, 5308–5313.

[2] Kai Cui, Sascha Hauck, Christian Fabian, and Heinz Koeppl. 2024. Learning Decentralized Partially Observable Mean Field Control for Artificial Collective Behavior. arXiv:2307.06175 [cs.LG] https://arxiv.org/abs/2307.06175

[3] Kai Cui and Heinz Koeppl. 2022. Approximately Solving Mean Field Games via Entropy-Regularized Deep Reinforcement Learning. arXiv:2102.01585 [cs.MA] https://arxiv.org/abs/2102.01585

[4] Yue Guan, Mohammad Afshari, and Panagiotis Tsiotras. 2024. Zero-Sum Games between Mean-Field Teams: Reachability-Based Analysis under Mean-Field Sharing. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 9 (Mar. 2024), 9731–9739. https://doi.org/10.1609/aaai.v38i9.28831

[5] Yue Guan, Mi Zhou, Ali Pakniyat, and Panagiotis Tsiotras. 2022. Shaping large population agent behaviors through entropy-regularized mean-field games. In *2022 American Control Conference (ACC)*. IEEE, IEEE, Atlanta, USA, June 08–10, 2022, 4429–4435.

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

[7] Shengyi Huang, Rousslan Fernand Julien Dossa, Antonin Raffin, Anssi Kanervisto, and Weixun Wang. 2022. The 37 Implementation Details of Proximal Policy Optimization. https://iclr-blog-track.github.io/2022/03/25/ppo-implementation-details/ https://iclr-blog-track.github.io/2022/03/25/ppo-implementation-details/.

[8] Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs.LG] https://arxiv.org/abs/1412.6980

[9] Marc Lanctot, Vinicius Zambaldi, Audrūnas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Pérolat, David Silver, and Thore Graepel. 2017. A unified game-theoretic approach to multiagent reinforcement learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 4193–4206.

[10] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2019. Continuous control with deep reinforcement learning. arXiv:1509.02971 [cs.LG] https://arxiv.org/abs/1509.02971

[11] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 6382–6393.

[12] Ashutosh Nayyar, Aditya Mahajan, and Demosthenis Teneketzis. 2013. Decentralized stochastic control with partial history sharing: A common information approach. *IEEE Trans. Automat. Control* 58, 7 (2013), 1644–1658.

[13] Martin J Osborne. 2004. An Introduction to Game Theory. *Oxford University Press* 2 (2004), 672–713.

[14] TES Raghavan. 1994. Zero-sum two-person games. *Handbook of game theory with economic applications* 2 (1994), 735–768.

[15] Naci Saldi, Tamer Başar, and Maxim Raginsky. 2023. Partially observed discrete-time risk-sensitive mean field games. *Dynamic Games and Applications* 13, 3 (2023), 929–960.

[16] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. 2018. High-Dimensional Continuous Control Using Generalized Advantage Estimation. arXiv:1506.02438 [cs.LG] https://arxiv.org/abs/1506.02438

[17] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. arXiv:1707.06347 [cs.LG] https://arxiv.org/abs/1707.06347

[18] Nevroz Sen and Peter E Caines. 2019. Mean field games with partial observation. *SIAM Journal on Control and Optimization* 57, 3 (2019), 2064–2091.

[19] Kai Shao, Jiacheng Shen, Chijie An, and Mathieu Laurière. 2024. Reinforcement Learning for Finite Space Mean-Field Type Games. arXiv:2409.18152 [cs.GT] https://arxiv.org/abs/2409.18152

[20] Max Olan Smith, Thomas Anthony, and Michael P. Wellman. 2021. Iterative Empirical Game Solving via Single Policy Best Response. arXiv:2106.01901 [cs.MA] https://arxiv.org/abs/2106.01901

[21] Sylvain Sorin. 2002. *A First Course on Zero-Sum Repeated Games*. Vol. 37. Springer Science & Business Media, Berlin.

[22] Aviv Tamar, Dotan Di Castro, and Shie Mannor. 2016. Learning the variance of the reward-to-go. *Journal of Machine Learning Research* 17, 13 (2016), 1–36.

[23] Guofang Wang, Ziming Li, Wang Yao, and Sikai Xia. 2022. A multi-population mean-field game approach for large-scale agents cooperative attack-defense evolution in high-dimensional environments. *Mathematics* 10, 21 (2022), 4075.

[24] Yaodong Yang, Rui Luo, Minne Li, Ming Zhou, Weinan Zhang, and Jun Wang. 2020. Mean Field Multi-Agent Reinforcement Learning. arXiv:1802.05438 [cs.MA] https://arxiv.org/abs/1802.05438

[25] Batuhan Yardim and Niao He. 2024. Exploiting Approximate Symmetry for Efficient Multi-Agent Reinforcement Learning. arXiv:2408.15173 [cs.GT] https://arxiv.org/abs/2408.15173

[26] Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. 2022. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in Neural Information Processing Systems* 35 (2022), 24611–24624.