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

Bhavini Jeloka
Department of Aerospace Engineering
Georgia Institute of Technology
Atlanta, GA 30318, USA
bjeloka3@gatech.edu

Yue Guan*
Oden Institute for Computational Engineering & Sciences
The University of Texas at Austin
Austin, TX 78712, USA
yue.guan@austin.utexas.edu

Panagiotis Tsiotras
Department of Aerospace Engineering
Georgia Institute of Technology
Atlanta, GA 30318, USA
tsiotras@gatech.edu

Abstract

While multi-agent reinforcement learning (MARL) has shown strong empirical performance, existing methods struggle with a large number of agents, due to the combinatorial growth of joint interactions. Mean-field (MF) approximations address this challenge by replacing pairwise interactions with surrogate interactions against population distributions, yielding tractable formulations to solve for the large-population policies. However, existing MF formulations focus on fully cooperative or purely competitive settings and fail to capture the mixed cooperative–competitive structure of team-based games. We extend mean-field learning to *zero-sum team games*, where agents cooperate within each team and compete at the team level. We show that such games admit ϵ-optimal decentralized policies that depend only on the local states of each agent and the population team distributions. Building on this structure, we propose MF-MAPPO, a scalable algorithm with a shared actor and a minimally informed critic per team. MF-MAPPO is trained directly in finite-population simulators rather than using mean-field oracles, thereby enabling deployment to realistic scenarios with thousands of agents. We further extend MF-MAPPO to partially observable settings via a simple gradient-regularized training scheme. Experiments on large-scale benchmarks in our simulation platform (`MFEnv`), including population games with analytical solutions and high-dimensional battlefield scenarios, demonstrate that MF-MAPPO outperforms existing MARL baselines and yields rich, heterogeneous behaviors.

1 Introduction

Large-scale multi-agent reinforcement learning (MARL) suffers from a scalability bottleneck: the joint state–action space grows exponentially with the number of agents, rendering standard approaches such as MADDPG [Lowe et al., 2017] and MAPPO [Yu et al., 2022] intractable in large populations. A principled remedy is provided by mean-field theory [Huang et al., 2006], which enables tractable analysis and learning by approximating interactions among agents through an interaction between a representative agent and the population distribution (i.e., the mean-field).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

*Work done while at the Georgia Institute of Technology.

Mean-field methods have led to a rich body of work in both non-cooperative settings, through mean-field games (MFGs) [Huang et al., 2006, Lasry and Lions, 2007], and fully cooperative settings, through mean-field control (MFC) [Bensoussan et al., 2013, Gu et al., 2021]. Recent learning-based methods include, MF-Q [Yang et al., 2020], DDPG-based extensions for mean-field team games [Shao et al., 2024a], and GAN-based approaches for continuous domains [Wang et al., 2022]. However, existing methods often rely on access to a **mean-field oracle** for exact population distribution trajectories, which are typically available only for simple scenarios. Additionally, these methods are primarily designed for either fully competitive or fully cooperative regimes. Consequently, scenarios involving **mixed cooperative–competitive interactions**, where agents collaborate within teams while competing across teams, remain underexplored despite their prevalence in real-world systems such as team sports [Gaviria Alzate et al., 2025] and social dilemmas [Leibo et al., 2017].

Recent work has begun to address this gap through team-based mean-field formulations. Early single-team results appear in Arabneydi and Mahajan [2014], while multi-population extensions have been studied in Perolat et al. [2021], typically under restrictive independence assumptions. More recently, Guan et al. [2024a] introduced *zero-sum mean-field team games* (ZS-MFTGs), later generalized to general-sum team settings by Subramanian et al. [2023]. While these formulations provide a principled foundation, existing numerical solution approaches rely primarily on linear–quadratic structure or discretization. Although such methods scale in the number of agents, they struggle *computationally* to handle complex environments with high-dimensional state spaces and nonlinear dynamics, limiting their applicability in modern settings.

Beyond computational challenges, introducing competition fundamentally alters the structure of mean-field interactions. In standard mean-field settings, optimality is characterized through single-agent deviations [Cui and Koeppl, 2022], whereas competitive team settings require reasoning about *team-level deviations*. As a result, optimal policies must depend on feedback from the opponent's distribution [Guan et al., 2024a]. This introduces another key challenge: agents require access to the opponent population distribution during execution, which is difficult under **partial observability of population distributions**. Most existing approaches assume perfect mean-field measurement or communication [Guan et al., 2024a, Subramanian et al., 2023]. While partial-observation and estimation methods [Inoue et al., 2021, Perrin et al., 2021, Benjamin and Abate, 2025a] have been explored, they often struggle in discrete spaces or in competitive environments. Moreover, addressing partial information requires not only accurate estimation of mean-field distributions, but also systematic adaptation of the learning framework to ensure robustness to estimation errors.

In this work, we address these challenges by developing a scalable learning framework for mixed cooperative–competitive mean-field systems.

Contributions. We introduce **Mean-Field Multi-Agent Proximal Policy Optimization** (MF-MAPPO), which achieves scalability for mixed cooperative–competitive settings through three key design principles with theoretical guarantees: (i) a *minimally-informed critic* that depends only on team distributions, ensuring that the input dimension remains independent of the number of agents; (ii) a *shared actor per team*, which reduces the joint strategy space to identical team policies and enables efficient learning (Theorem 3.1); and (iii) the ability to train with N-agent simulators and generalize to larger populations (Theorems 3.2 and 3.3).

To address partial observability, we show that MF-MAPPO can be combined with gradient regularization and deployed using estimated mean-fields without requiring retraining, and we provide performance bounds depending on the estimation accuracy (Proposition 2).

We validate our approach across several environments, demonstrating that MF-MAPPO produces stable learning dynamics and rich behaviors without reward engineering.

Overall, our contributions advance **scalable** reinforcement learning for realistic large-population systems by jointly addressing **mixed cooperative–competitive interactions**, **partial observability**, and **finite-population training**.

Related Work. Table 1 highlights works most relevant to our setting and clarifies our contributions. [Perolat et al., 2021] developed an online mirror descent approach for multi-population MFG, but focuses on open-loop policies and does not capture team-level competition. [Benjamin and Abate, 2025a] extended this framework to networked settings with communication-based MF estimation, but is restricted to a single population. [Shao et al., 2024a] introduced DDPG-MFTG

Table 1: Comparison between our approach and related works.

Method	Multiple Populations	MF-Dependent Policies	Mixed Coop-Comp Setting	Finite-Population Simulator	Partial MF Observation
MFG-MOMD [Perolat et al., 2021]	✓	✗	✗	✗	✗
Networked-MOMD [Benjamin and Abate, 2025a]	✗	✓	✗	✓	✓
DDPG-MFTG [Shao et al., 2024a]	✓	✓	✓	✗	✗
MFTG-Discretization [Guan et al., 2024a]	✓	✓	✓	✗	✗
MF-MAPPO (Ours)	✓	✓	✓	✓	✓

for mixed cooperative–competitive scenarios, but it relies on a mean-field oracle providing exact infinite-population trajectories and assumes fully observable mean-field information, limiting its applicability to complex environments. Guan et al. [2024a] studied zero-sum mean-field team games via discretization, which scales in the number of agents but becomes intractable in large state spaces.

2 Problem Formulation and Mean-Field Approximation

A zero-sum mean-field team game (ZS-MFTG) models a discrete-time stochastic game between two large teams of agents (Red and Blue) [Guan et al., 2024a]. The Blue and Red teams consist of N_1 and N_2 *identical* agents for each team. 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 $j \in [N_2]$ at time t. 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. We denote the space of probability measures over a set E as $\mathcal{P}(E)$. Below, $\mathrm{d_{TV}}(\mu, \mu')$ represents the total variation between $\mu, \mu' \in \mathcal{P}(E)$.

Definition 1. The *empirical distributions* (ED) for the two teams are respectively defined as $\mathcal{M}_t^{N_1}(x) = \frac{1}{N_1} \sum_{i=1}^{N_1} \mathbb{1}_x(X_{i,t}^{N_1})$ and $\mathcal{N}_t^{N_2}(y) = \frac{1}{N_2} \sum_{j=1}^{N_2} \mathbb{1}_y(Y_{j,t}^{N_2})$, where $\mathbb{1}_{\cdot}(\cdot)$ is the indicator.

Specifically, $\mathcal{M}_t^{N_1}(x)$ gives the fraction of Blue agents at state x. 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.

Dynamics and Rewards. We consider weakly-coupled dynamics where the dynamics of each individual agent are coupled with other agents through the EDs [Huang et al., 2006]. For the Blue agent i, its stochastic transition is governed by the kernel $f_t : \mathcal{X} \times \mathcal{U} \times \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y}) \to \mathcal{P}(\mathcal{X})$:

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

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 a Red agent is governed by the transition kernel $g_t(y_{j,t+1}^{N_2} | y_{j,t}^{N_2}, v_{j,t}^{N_2}, \mu_t^{N_1}, \nu_t^{N_2})$. All agents in the Blue team receive an identical *weakly-coupled* team reward, i.e., $r_t \triangleq r_t(\mu_t, \nu_t) : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y}) \to \mathbb{R}$. The Red agents receive $-r_t(\mu_t, \nu_t)$ as their rewards (zero-sum). We assume that the Blue team is *maximizing* while the Red team is *minimizing* and $r_t \in [-R_{\max}, R_{\max}]$ for all t.

Assumption 1 (Lipschitz Model)**.** For all $x \in \mathcal{X}, u \in \mathcal{U}$, $\mu, \mu' \in \mathcal{P}(\mathcal{X})$, $\nu, \nu' \in \mathcal{P}(\mathcal{Y})$ and all t, there exist constants $L_f, L_r > 0$ such that $\sum_{x' \in \mathcal{X}} |f_t(x'|x, u, \mu, \nu) - f_t(x'|x, u, \mu', \nu')| \le L_f\big(\mathrm{d_{TV}}(\mu, \mu') + \mathrm{d_{TV}}(\nu, \nu')\big)$ and $|r_t(\mu, \nu) - r_t(\mu', \nu')| \le L_r\big(\mathrm{d_{TV}}(\mu, \mu') + \mathrm{d_{TV}}(\nu, \nu')\big)$.

We similarly assume that g_t is L_g-Lipschitz. The Lipschitz model is a standard assumption in the mean-field literature [Huang et al., 2006, Gu et al., 2021], without which the mean-field approximation may fail as shown in the counterexamples [Cui et al., 2024, Guan et al., 2024a].

Information Structure. We consider a mean-field sharing information structure [Arabneydi and Mahajan, 2015], where each agent's decision depends on its own state and the two teams' EDs. We start by assuming full observation of EDs and later we relax this assumption in Section 4. Specifically, the Blue and Red teams seek to construct *mixed* Markov policies

$$\phi_{i,t}(u | x_{i,t}^{N_1}, \mu_t^{N_1}, \nu_t^{N_2}) \text{ and } \psi_{j,t}(v | y_{j,t}^{N_2}, \mu_t^{N_1}, \nu_t^{N_2}).$$

We use Φ_t to denote the set of individual Blue 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 $\psi_t^{N_2} \in \Psi_t^{N_2} = \times_{N_2} \Psi_t$.

Optimization Objective. The performance of a team policy pair (ϕ^{N_1}, ψ^{N_2}) is measured by $J^{N,\phi^{N_1},\psi^{N_2}}(\mathbf{x}_0^{N_1}, \mathbf{y}_0^{N_2}) = \mathbb{E}_{\phi^{N_1},\psi^{N_2}}\left[\sum_{t=0}^{T} r_t(\mathcal{M}_t^{N_1}, \mathcal{N}_t^{N_2}) \middle| \mathbf{x}_0^{N_1}, \mathbf{y}_0^{N_2}\right]$. When the maximizing Blue team considers its worst-case performance, we have the following max-min optimization:

$$\underline{J}^{N*}(\mathbf{x}_0^{N_1}, \mathbf{y}_0^{N_2}) = \max_{\phi^{N_1} \in \Phi^{N_1}} \min_{\psi^{N_2} \in \Psi^{N_2}} J^{N,\phi^{N_1},\psi^{N_2}}(\mathbf{x}_0^{N_1}, \mathbf{y}_0^{N_2}), \tag{2}$$

where $\underline{J}^{N*}$ is the lower game value for the N-agent finite-population game. Similarly, the minimizing Red team considers a min-max optimization and the upper game value. Note that the optimization in (2) allows agents within each team to follow different strategies. Such *non-identical team policies* preserve maximum flexibility for the Blue team and, more crucially, avoid underestimating the adversary Red team's strategy flexibility. However, this generality leads to exponential growth in the joint policy space, rendering the problem intractable [Bernstein et al., 2002].

Battlefield Scenario. To motivate and ground the ZS-MFTG formulation, we consider a grid-based battlefield scenario in which Blue agents aim to reach designated targets (green cells), while Red agents attempt to block their access. At each cell, an activation–deactivation mechanism governs interactions: an agent's probability of being deactivated is proportional to the opponent's numerical advantage relative to its own team's presence in that cell. This induces strategic grouping and dispersion behaviors. The Blue team seeks to maximize the fraction of its population that reaches the target cells, while the Red team aims to minimize this quantity, resulting in a zero-sum objective. Applying the Emp operator yields the empirical distributions of both teams.

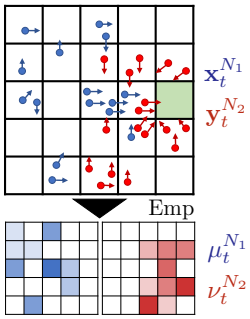

Figure 1: Battlefield Scenario.

Mean-Field Approximation. To reduce the complexity of team policy optimization domains in (2), previous works [Guan et al., 2024a, Subramanian et al., 2023] propose to approximate complex team behaviors using *identical team policies* in the *infinite-population* limit.

Definition 2 (Identical Team Policy)**.** The Blue team policy $\phi_t^{N_1} = (\phi_{1,t}, \ldots, \phi_{N_1,t})$ is *identical*, if $\phi_{i_1,t} = \phi_{i_2,t}$ for all times t and all $i_1, i_2 \in [N_1]$. The set of identical Blue team policies is denoted as Φ. Similarly, Ψ represents the set of identical Red team policies.

With an identical team policy, an infinite-population team's joint states can be represented using the team population distribution, which coincides with the state distribution of a typical agent and is commonly known as the *mean-field* (MF) [Huang et al., 2006]. We use μ_t and ν_t to denote the Blue and Red MFs, respectively. Guan et al. [2024a] showed that MFs induced by identical team policies closely approximate finite-population EDs under non-identical policies. This justifies the restriction of the optimization domain in (2) to identical policies, yielding the infinite-population optimization

$$\underline{J}^{*}_{\infty}(\mu_0, \nu_0) = \max_{\phi \in \Phi} \min_{\psi \in \Psi} J^{\phi,\psi}_{\infty}(\mu_0, \nu_0), \tag{3}$$

where μ_t and ν_t follow *deterministic* dynamics, corresponding to the evolution of the state distributions of the Blue and Red representative agents.

While the lower (max–min) and upper (min–max) values may not coincide in general, i.e., a Nash value may not exist in the infinite-population game, we can show that a Nash equilibrium exists under a convex–concave reward structure and independent dynamics (where transitions depend only on an agent's state and action). See Appendix C for details.

3 Mean-Field Multi-Agent Proximal Policy Optimization

Note that the optimization (3) is defined in the infinite-population limit and depends on the MF trajectories (μ_t, ν_t). To obtain such trajectories, existing learning approaches [Shao et al., 2024b, Perolat et al., 2021] typically rely on mean-field oracles that propagate the state distributions of representative agents. However, these oracles are only available when the scenarios are simple enough to permit exact distributional propagation. In realistic environments with complex dynamics and high-dimensional state spaces, one needs to leverage finite-population simulators to generate ED trajectories using samples of joint states and actions as approximations of the underlying MF

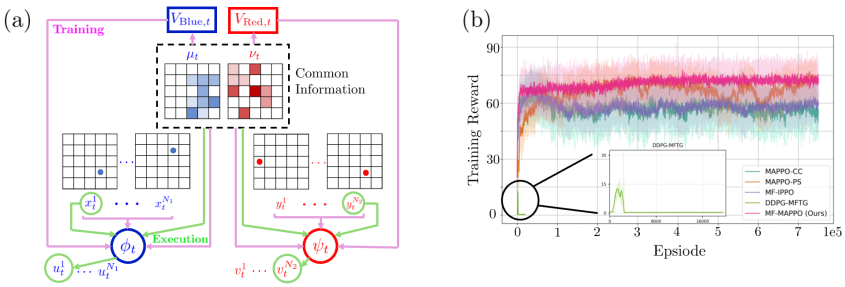

Figure 2: (a) Architecture of MF-MAPPO. (b) Training curve for Battlefield on a 4x4 grid (Blue).

trajectories. Motivated by this need for generality, we design the MF-MAPPO algorithm to operate directly on finite-population simulators, which introduces two key challenges: (i) The learned policies are optimal only for finite-population games, whereas existing theoretical guarantees rely on optimality in the infinite-population limit; thus, new performance guarantees are required. (ii) Policy gradients estimated from finite-population simulators are stochastic and often deviate from their mean-field counterparts, requiring additional convergence analysis.

Guan et al. [2024a] show that the optimal policy ϕ^* obtained from the *infinite-population* optimization in (3) is $\mathcal{O}(1/\sqrt{\underline{N}})$-optimal in the finite-population game, even against non-identical exploiting Red team policies, where $\underline{N} = \min\{N_1, N_2\}$. The following theorem presents a similar guarantee when the identical policy is obtained as an *exact* solution to the finite-population optimization problem.

Theorem 3.1. *Let $\tilde{\phi}^*$ be a solution to the finite-population optimization restricted to identical policies, i.e., $\max_{\phi\in\Phi}\min_{\psi\in\Psi} J^{N,\phi,\psi}(\mathbf{x}_0^{N_1}, \mathbf{y}_0^{N_2})$. Then, the exploitability of $\tilde{\phi}^*$ satisfies*

$$\min_{\psi^{N_2}\in\Psi^{N_2}} J^{N,\tilde{\phi}^*,\psi^{N_2}}(\mathbf{x}^{N_1}, \mathbf{y}^{N_2}) \geq \underline{J}^{N*}(\mathbf{x}^{N_1}, \mathbf{y}^{N_2}) - \mathcal{O}(1/\sqrt{\underline{N}}), \quad \forall \mathbf{x}^{N_1} \in \mathcal{X}^{N_1}, \mathbf{y}^{N_2} \in \mathcal{Y}^{N_2}.$$

Theorem 3.1 provides a principled justification for learning *identical* team policies in finite-population games, even when the opponent may exploit using non-identical strategies (i.e., $\psi^{N_2} \in \Psi^{N_2}$). Moreover, the exploitability vanishes as $N_1, N_2 \to \infty$, recovering the classical infinite-population mean-field result [Huang et al., 2006].

With identical policies justified, we introduce our main contribution: the Mean-Field Multi-Agent Proximal Policy Optimization (MF-MAPPO) algorithm. MF-MAPPO builds on the proximal policy optimization (PPO) framework with two key adaptations: *minimally-informed critics* and *shared team actors*. These components exploit the underlying game structures to enable efficient learning while maintaining scalability. We describe these design details below.

Minimally-Informed Critic. The shared team reward and identical policy approximation imply that the value function depends only on the team distributions (See Proposition 5 in Appendix D). Accordingly, we employ a *minimally-informed critic* per team, which takes (μ, ν) as input, independent of joint agent states and actions. The Blue team critic evaluates the value function $V_{\text{Blue}}(\mu, \nu)$. We use ζ_{Blue} to parameterize the critic network and train it via minimizing the MSE loss as follows

$$L_{\text{critic}}(\zeta_{\text{Blue}}) = \frac{1}{|B|}\sum_{k=1}^{|B|}\left(V_{\zeta_{\text{Blue}}}(\mu_k, \nu_k) - \hat{R}_{\text{Blue},k}\right)^2, \tag{4}$$

where B is the mini-batch size and $\hat{R}_{\text{Blue},k}$ is the discounted reward-to-go for sample k. The Red team critic is parameterized with ζ_{Red} and updated similarly. The following proposition justifies the usage of minimally-informed critics.

This result reduces the learning problem to a single critic per team. Unlike standard centralized training/decentralized execution (CTDE), which requires joint observations of all agents $(\mathbf{x}_t^{N_1}, \mathbf{y}_t^{N_2})$, our critic depends only on the empirical distributions. This design is critical for scalability, as the input dimension is independent of the number of agents.

Shared Team Actor. The exploitability bound of identical team strategies in Theorem 3.1 allows us to employ a shared actor per team. The actor optimizes a PPO-based objective with a decaying

entropy bonus [Schulman et al., 2017, Huang et al., 2022], which promotes exploration and stabilizes learning in mean-field settings [Cui and Koeppl, 2022, Guan et al., 2022]. Identical team policies further allow a single replay buffer per team, reducing memory costs and simplifying experience collection. The PPO-based objective function of the Blue team actor is given by:

$$L_{\text{actor}}(\theta_{\text{Blue}}) = \frac{1}{|B|}\sum_{k=1}^{|B|}\Big[\min\Big(g_k(\theta_{\text{Blue}})A_k, \text{clip}_{[1-\epsilon,1+\epsilon]}(g_k(\theta_{\text{Blue}}))A_k\Big) + \omega S(\phi_{\theta_{\text{Blue}}}(x_k,\mu_k,\nu_k))\Big], \quad (5)$$

where, $g(\theta) = \frac{\phi_\theta(u|x,\mu,\nu)}{\phi_{\theta^{\text{old}}}(u|x,\mu,\nu)}$, A_k is the generalized advantage function estimate [Schulman et al., 2018] and the tunable parameter ω weighs the entropy term $S(\cdot)$ and decays as training progresses.

3.1 Theoretical Guarantees for Finite-Population Training

As discussed earlier, the infinite-population limit remains of significant theoretical interest. It is therefore desirable to ensure that policy gradients computed using finite-population simulators converge to their infinite-population counterparts as the population size grows, which motivates the following theorem.

Theorem 3.2 (*Informal*)**.** *The approximate policy gradient of the infinite-population Blue (Red) team policy computed using identical local policies* $(\phi_{\theta_{\text{Blue}}}, \psi_{\theta_{\text{Red}}})$ *via MF-MAPPO from the* (N_1, N_2)*-agent simulator converges to the infinite-population policy gradient, i.e.,* $\|\nabla_{\theta_{\text{Blue}}} J_\infty(\phi_{\theta_{\text{Blue}}}, \psi) - \nabla_{\theta_{\text{Blue}}} \hat{J}^{N_1,N_2}(\phi_{\theta_{\text{Blue}}}, \psi)\|_2 \to 0$ *as* $(N_1, N_2) \to \infty$.

Theorem 3.2 relates policies learned using finite-population simulators to their infinite-population counterparts. This connection allows us to use the infinite-population solution as a bridge to analyze the performance of the learned policy under different population sizes, leading to Theorem 3.3. In particular, Theorem 3.3 allows MF-MAPPO to be trained on smaller populations and deployed to larger teams without additional tuning, while maintaining performance guarantees.

Theorem 3.3. *Let* $\mathcal{G}_N$ *denote the finite-population game in which agents use identical team policies* (ϕ_t^*, ψ_t^*) *with an* (N_1, N_2)*-agent simulator. Consider another game* $\mathcal{G}_{\tilde{N}}$ *with the same setup but population sizes* $(\tilde{N}_1, \tilde{N}_2)$ *satisfying* $\tilde{N}_1/\tilde{N}_2 = N_1/N_2$ *and* $\tilde{N}_1 \geq N_1$*. If* (ϕ_t^*, ψ_t^*) *is* ϵ*-optimal in* $\mathcal{G}_N$*, then it remains* ϵ*-optimal in* $\mathcal{G}_{\tilde{N}}$.

4 MF-MAPPO for Partially Observable MFTGs

Unlike standard mean-field formulations based on single-agent deviations, ZS-MFTGs consider team-level deviations (2), requiring strategies that depend on feedback from both teams' distributions. Since access to the opponent distribution is often unrealistic, we extend MF-MAPPO to partially observable settings where opponent distributions must be estimated.

Partially Observable Battlefield. To motivate partial observability, consider a battlefield scenario with a state-dependent visibility graph $\mathcal{G}_t^{\text{viz}}$ and communication graph $\mathcal{G}_t^{\text{comm}}$, which define the states an agent can observe and the agents it can communicate with. In the example to the right, the cyan ego agent observes a 3×3 neighborhood under $\mathcal{G}_t^{\text{viz}}$. Based on $\mathcal{G}_t^{\text{comm}}$, it can communicate with agents in four neighboring cells, leading to information exchange with the teammate to the left.

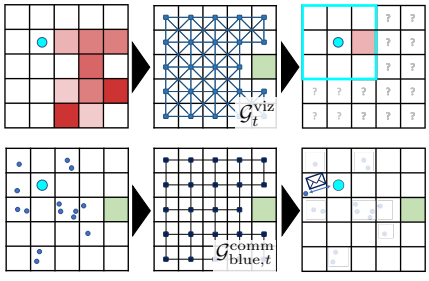

Figure 3: PO-MFTG

Given the inevitable noise in the estimates, it is first necessary to ensure that the learned policies are robust to errors in the estimated mean-field distributions $(\hat{\mu}_t, \hat{\nu}_t)$. To this end, we introduce gradient regularization [Chen et al., 2024] in the actor objective (5) to enforce Lipschitz continuity with respect to the mean-field inputs. This promotes robustness, so that small distribution estimation errors induce only minor changes in the action distributions. Specifically,

$$L_{\text{actor}}^{\text{GR}}(\theta_{\text{Blue}}) = L_{\text{actor}}(\theta_{\text{Blue}}) + \frac{\lambda^{\text{GR}}}{|B|}\sum_{k=1}^{|B|}\|\nabla_\eta \log\phi(u_k|x_k,\eta)\big|_{\eta=(\mu_k,\nu_k)}\|_2^2, \quad (6)$$

where $L_{\text{actor}}(\theta_{\text{Blue}})$ follows from (5), $\eta = (\mu, \nu)$ and λ^{GR} is a tunable hyperparameter to control the regularization. The following proposition relates the gradient regularization to the Lipschitz constant L_ϕ of the learned Blue team policy.

Proposition 1. *If the log-probability of the Blue team policy is uniformly bounded such that for all $x \in \mathcal{X}, u \in \mathcal{U}, \mu \in \mathcal{P}(\mathcal{X}), \nu \in \mathcal{P}(\mathcal{Y})$, $\|\nabla_\eta \log \phi(u \mid x, \mu, \nu)\|_2 \leq L_\phi/2|\mathcal{U}|$, where the gradient is taken with respect to $\eta \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})$, then $\phi(u \mid x, \mu, \nu)$ is L_ϕ-Lipschitz continuous, i.e., $\sum_u |\phi_t(u|x, \mu, \nu) - \phi_t(u|x, \mu', \nu')| \leq L_\phi \left(\mathrm{d_{TV}}(\mu, \mu') + \mathrm{d_{TV}}(\nu, \nu')\right)$ for all $x \in \mathcal{X}$.*

While we still train gradient-regularized MF-MAPPO (GR-MF-MAPPO) with full mean-field information similar to the CTDE framework, we rely on mean-field estimators being ϵ-accurate to ensure performance of the resulting team policies under partial observability.

Definition 3. Suppose Blue agent i receives an observation $o_{i,t} \in \mathcal{O}_t$ at time t. A Red mean-field estimator $\mathcal{H}_\nu : \mathcal{O}_t \to \mathcal{P}(\mathcal{Y})$ is said to be ϵ-accurate if $\mathrm{d_{TV}}\big(\mathcal{H}_\nu(o_{i,t}), \nu_t^{N_2}\big) \leq \epsilon$ for all $i \in [N_1]$, where $\nu_t^{N_2}$ is the true Red distribution. We similarly define $\mathcal{H}_\mu : \mathcal{O}_t \to \mathcal{P}(\mathcal{X})$ for the Red team.

We provide performance guarantees for GR-MF-MAPPO when deployed with ϵ-accurate estimators. Let (ϕ^*, ψ^*) denote the learned policies. Under full mean-field feedback, the induced trajectories are $(\mathcal{M}^{N_1}, \mathcal{N}^{N_2})$, where Blue actions are sampled according to $\phi^*(\cdot|x_{i,t}, \mathcal{M}_t^{N_1}, \mathcal{N}_t^{N_2})$. Under partial observability, Blue agents instead act using estimated opponent distributions, yielding trajectories $(\hat{\mathcal{M}}^{N_1}, \hat{\mathcal{N}}^{N_2})$ with actions sampled from $\phi^*(\cdot|x_{i,t}, \hat{\mathcal{M}}_t^{N_1}, \mathcal{H}_\nu(o_{i,t}))$, with Red actions sampled analogously. We quantify the performance degradation through the cumulative regret $\Delta J(\phi^*, \psi^*) = \mathbb{E}_{\phi^*, \psi^*}\big[\big|\sum_{t=0}^T r_t(\mathcal{M}_t^{N_1}, \mathcal{N}_t^{N_2}) - \sum_{t=0}^T r_t(\hat{\mathcal{M}}_t^{N_1}, \hat{\mathcal{N}}_t^{N_2})\big|\big]$. See Appendix I for details.

Proposition 2. *When a policy pair (ϕ_t^*, ψ_t^*) learned via GR-MF-MAPPO is deployed with an ϵ-accurate estimator and satisfies a uniform Lipschitz bound, the cumulative regret satisfies $\Delta J(\phi_t^*, \psi_t^*) \leq K\epsilon + \mathcal{O}(1/\sqrt{N})$.*

Consequently, GR-MF-MAPPO can incorporate different estimators without retraining, with performance degrading linearly in ϵ. We present two estimation algorithms for the Battlefield scenario, both of which can be shown to be ϵ-accurate. Details of the estimators are provided in Appendix B.

Classical Consensus. In consensus-based mean-field estimation [Benjamin and Abate, 2025b], agents observe opponent counts based on their visibility graph $\mathcal{G}_t^{\text{viz}}$ and exchange the counts through the communication graph $\mathcal{G}_t^{\text{com}}$ over R_{com} rounds. Each agent then constructs a mean-field estimate from the aggregated counts, assigning the remaining probability mass uniformly to unobserved states. The resulting estimator is $\big(1 - \mathcal{O}(R_{\text{com}}/|\mathcal{X}|)\big)$-accurate.

Dynamic-Projected Consensus. Instead of communicating agent counts, we propose the dynamic-projected consensus (D-PC) algorithm that exchanges mean-field estimates directly. The agent at state x first constructs an initial estimate $\hat{\nu}_{x,t}^{(0)} \in \mathcal{R}(x)$, where $\mathcal{R}(x) \subseteq \mathcal{P}(\mathcal{Y})$ denotes the set of distributions consistent with locally observed opponent counts. During communication round τ, agents exchange their previous estimates and compute the weighted consensus $\xi_{x,t}^{(\tau)} = \sum_{x' \in \text{Neighbor}} w(x, x') \hat{\nu}_{x',t}^{(\tau-1)}$, followed by the projection $\hat{\nu}_{x,t}^{(\tau)} = \Omega_{\mathcal{R}(x)}(\xi_{x,t}^{(\tau)})$, which preserves consistency with locally observed data. Compared to classical consensus, D-PC is $\mathcal{O}\big(e^{-cR_{\text{com}}}\big)$-accurate (Theorem B.1), i.e., exponentially decreasing error with R_{com}, and is empirically more robust to noisy communication.

5 Numerical Experiments

We showcase MF-MAPPO's efficacy on the following three large-population scenarios using our novel benchmark simulation platform, `MFEnv`[2]: (1) a population-based rock–paper–scissors game, (2) a battlefield setting where Blue and Red teams engage in attack–defense tasks, and (3) a SIS/epidemiology scenario with probabilistic infection dynamics. Additional details of the experiment setups and results are presented in Appendix M. To the best of our knowledge, DDPG-MFTG [Shao et al., 2024b] is the only method explicitly designed for mixed cooperative–competitive large-population team problems. Accordingly, we compare against DDPG-MFTG and three mean-field "house-made" adaptations of standard MARL methods: MF-IPPO and two MAPPO variants, parameter sharing (MAPPO-PS) and centralized critic (MAPPO-CC) [Yu et al., 2022]. Centralized Q-learning methods

[2] `https://github.com/bhavini-jeloka/mf-env`

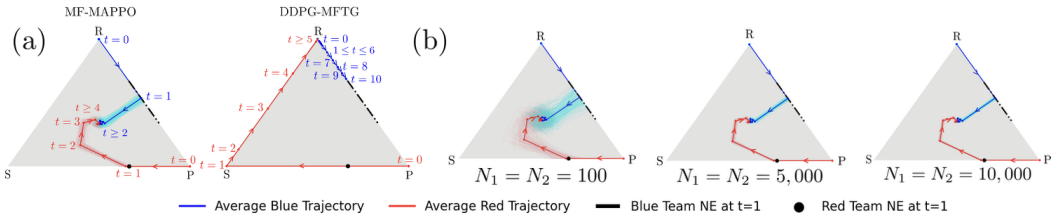

Figure 4: (a) MF-MAPPO and DDPG-MFTG trajectories. (b) MF-MAPPO under varying team sizes.

such as QMix [Rashid et al., 2020], MADDPG [Lowe et al., 2017], and FACMAC [Peng et al., 2021] are omitted, since their centralized Q-functions depend on joint actions and states, making them intractable in large-population settings.

5.1 Validation of Convergence: Constrained Rock-Paper-Scissors (cRPS)

We use a population-based rock–paper–scissors (RPS) game to validate the convergence of MF-MAPPO to equilibrium before deploying it to more complex scenarios. This setting is chosen for its simplicity, which allows the equilibrium to be computed analytically. The individual state space is $\{\texttt{R},\texttt{P},\texttt{S}\}$, and we consider a restriction of the action space to $\{\texttt{CW},\texttt{Stay}\}$, i.e, either move clockwise ($\texttt{R}\to\texttt{P}$, $\texttt{P}\to\texttt{S}$, $\texttt{S}\to\texttt{R}$) or remain on current state. At each time step t, the reward is given by $r_t(\mu_t,\nu_t)=\mu_t^\intercal A\nu_t$ where A is the standard RPS payoff matrix. The equilibrium state distribution is uniform for both teams.

Table 2 compares MF-MAPPO with baseline algorithms, along with their critic inputs. We test the learned policies with a fixed initial distribution, with all Blue agents initialized at $\texttt{R}$ and Red agents at $\texttt{P}$. One can show analytically that the unique game value is $-1/3$ for this initial condition. Furthermore, we have the optimal conditions for the trajectories: (i) $\mu_t^*=\nu_t^*=[\frac{1}{3},\frac{1}{3},\frac{1}{3}]^\intercal$ for all $t\geq 2$, and (ii) $\mu_1^*=[0,1-\delta,\delta]^\intercal$ where $\delta\in[\frac{1}{3},\frac{2}{3}]$ and $\nu_1^*=[0,\frac{2}{3},\frac{1}{3}]^\intercal$. See Proposition 6 in Appendix J for details. As seen in Table 2, MF-MAPPO achieves rewards closest to the analytical Nash value $(-1/3)$, indicating that the minimally-informed critic preserves value information despite its reduced input dimensionality. This leads to faster and more efficient training as shown by the training time.

Figure 4(a) compares trajectories of MF-MAPPO and DDPG-MFTG and plots the analytical optimal distribution to be achieved at $t=1$. MF-MAPPO follows the prescribed optimal trajectories and converges to the equilibrium uniform distribution; DDPG-MFTG, however, fails to converge, since it conditions only on the ego team's policy, limiting its ability to respond to adversarial behaviors. Figure 4(b) shows trajectories of MF-MAPPO policies trained with 1,000 agents and deployed to larger populations. The average population distribution converges to equilibrium, while larger populations reduce trajectory variance, consistent with Theorem 3.3.

5.2 Emergent Heterogeneous Team Behaviors

With convergence validated, we deploy MF-MAPPO to two more complex grid-based scenarios, whose setups are briefly described below. All experiments use $N_1=N_2=100$ agents.

Battlefield Game. This game, featuring an activation–deactivation mechanism, was introduced in Section 2. The reward is defined as $r_t(\mu_t,\nu_t)=\sum_{x\in\text{Target}}\mu_t(x)$, which incentivizes the (maximiz-

Table 2: Performance comparison for cRPS

Algorithm	Critic Input	Average Reward	NE Attained	Training Time
MF-IPPO	$\{(x_{i,t}^{N_1},\mu_t)\}_{i=1}^{N_1}$	-0.340	✓	2:45:11
MAPPO-PS	$\{(x_{i,t}^{N_1},\mu_t,\nu_t)\}_{i=1}^{N_1}$	-0.313	✓	2:47:54
MAPPO-CC	$(\mathbf{x}_t^{N_1},\mathbf{y}_t^{N_2},\mu_t,\nu_t)$	-0.342	✓	3:20:46
DDPG-MFTG	(ϕ_t,μ_t,ν_t)	3.774	✗	60:49:35
MF-MAPPO (Ours)	(μ_t,ν_t)	**-0.331**	✓	**2:17:15**

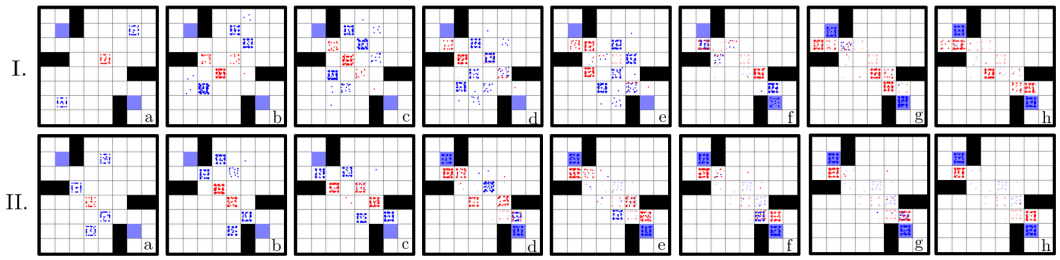

Figure 5: I. Red is concentrated; 30% Blue are at the bottom, the rest are at the top II. Blue is evenly split, Red is concentrated.

ing) Blue team to reach target cells, while the (minimizing) Red team seeks to block or deactivate Blue agents. Additional details are provided in Appendix K.

Epidemiology Game. In this scenario, Red agents emulate a virus that infects the Blue population, with infected Blue agents further spreading the disease. The Blue team aims to maximize the size of its healthy population, leading to the reward $r_t(\mu_t, \nu_t) = \sum_{x \in \text{Healthy}} \mu_t(x)$. Accordingly, Blue agents avoid infected peers and move toward hospital states for recovery, while the Red (virus) team tracks the Blue population and block access to hospitals. See Appendix L for additional details.

We compare MF-MAPPO with baselines via head-to-head evaluation for both scenarios with full mean-field feedback (Table 3). MF-MAPPO consistently achieves superior performance: Fixing the Red policy (row) shows that the MF-MAPPO Blue team always attains the highest performance, while fixing the Blue policy (column) shows that the MF-MAPPO Red team performs the best except against Blue DDPG-MFTG in Epidemiology[3]. Figure 2(b) presents the training curves in the Battlefield scenario, where MF-MAPPO converges faster and achieves higher rewards than all baselines. These results highlight the advantages of combining a minimally informed critic with a shared team actor: the resulting networks are smaller, have input dimensions independent of the population size, and achieve better performance. Notably, DDPG-MFTG exhibits initial performance gains but subsequently fails to maintain and further improve, and we were unable to stabilize the algorithm despite extensive hyperparameter tuning.

We deploy MF-MAPPO to a more challenging 8×8 battlefield in Figure 5 where the Red team faces a dilemma in deciding which target to defend, while the Blue team exploits this ambiguity. Since DDPG-MFTG and other homemade baselines require high computation due to their high network input dimensions (and DDPG-MFTG has a high output dimension), they are excluded from our analysis, and we qualitatively assess MF-MAPPO's performance. Panels I–II present sequences of MF-MAPPO movements under different initial distributions. In subplot I, the Blue team fakes a move towards the top-left target and then switches towards the other target to trick the Red team. In subplot II, each Blue sub-team first reaches towards the closest target, and the Red team splits

Table 3: Performance evaluations averaged over 10 random seeds.

Battlefield	MF-IPPO	MAPPO-PS	MAPPO-CC	DDPG-MFTG	MF-MAPPO (Ours)
MF-IPPO	61.54	73.85	70.60	0.00	**86.26**
MAPPO-PS	55.82	72.78	71.64	0.00	**86.22**
MAPPO-CC	51.60	68.81	65.58	0.00	**82.58**
DDPG-MFTG	65.26	75.14	77.21	0.00	**85.95**
MF-MAPPO (Ours)	**36.10**	**54.43**	**62.00**	0.00	**76.26**
Epidemiology	MF-IPPO	MAPPO-PS	MAPPO-CC	DDPG-MFTG	MF-MAPPO (Ours)
MF-IPPO	62.76	58.26	59.48	35.56	**77.83**
MAPPO-PS	57.95	57.63	55.41	29.52	**79.78**
MAPPO-CC	55.64	55.16	55.13	**27.95**	**76.74**
DDPG-MFTG	99.83	93.99	88.26	80.22	**107.39**
MF-MAPPO (Ours)	**54.71**	**46.38**	**47.42**	30.50	**71.04**

[3]MF-MAPPO adopts a more conservative policy that is robust to stronger opponents, and therefore exploits weak strategies, such as the DDPG-MFTG Blue policy, less aggressively than MAPPO-CC.

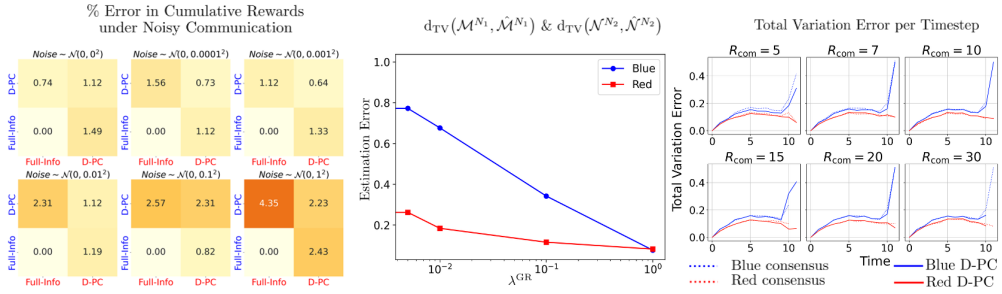

Figure 6: I. Reward error under different communication noises; II. Effect of gradient regularization on estimation quality; III. Estimation error of Classical Consensus and D-PC under varying R_{com}.

and tries to block both corridors. The highlighted Blue subteam realized the corridor would be blocked by the Red team and headed towards the bottom-right target. The Red team then draws part of its force blocking the upper-left target to reinforce the blockage at the bottom-right, and is able to deactivate all agents in the highlighted Blue subteam. These two game sequences illustrate how identical policies can generate complex, heterogeneous team behaviors, including emergent Blue agent target adaptation and Red team target-defense role distribution and reinforcement behaviors, highlighting the flexibility and richness of the mean-field approximation and identical team strategies.

5.3 Performance under Partial Observability

We evaluate GR-MF-MAPPO with ϵ-accurate estimators in a partially observable battlefield setting where each team must estimate the opponent's distribution (Figure 6). When GR-MF-MAPPO policies are paired with D-PC estimators for both teams, the cumulative regret decreases as R_{com} increases from 5 to 12 (Table 4), aligning with Proposition 2. Although $R_{\text{com}} = 3$ yields lower regret than 5 and 10, its substantially higher variability and wider confidence interval indicate less reliable estimation, with additional communication rounds yielding more consistent performance. We further empirically study the impact of communication quality (Figure 6I). With $R_{\text{com}} = 7$, increasing Gaussian noise variance leads to larger regret deviations; however, these remain within 5%, as the D-PC projection step mitigates noise accumulation from communication. Panel II presents an ablation study of gradient regularization. Larger λ^{GR} penalty imposes stronger smoothness (Lipschitz) constraints, yielding reduced trajectory deviation and thus improved robustness against estimation errors. Finally, Panel III shows the estimation error of the opponent distribution over time under varying communication rounds. Both estimators achieve comparable total variation errors, with D-PC performing better under limited communication budgets ($R_{\text{com}} < 10$).

6 Conclusion

In this work, we studied zero-sum mean-field team games and introduced MF-MAPPO, a scalable MARL framework for mixed cooperative–competitive large-population systems. MF-MAPPO combines a minimally informed critic with shared team policies, enabling efficient learning in finite-population simulators while scaling to thousands of agents. We further extended the framework to partially observable settings through gradient regularization and decentralized mean-field estimation algorithms. Experiments on large-scale benchmarks demonstrated that MF-MAPPO outperforms existing MARL and mean-field baselines while generating rich emergent team behaviors. Future work includes improving scalability in high-dimensional state spaces and extending the framework to more general communication and interaction structures.

Table 4: Cumulative regret under varying R_{com}

R_{com}	Average Regret	95% Confidence Interval	CI Width
3	4.563 ± 0.621	[3.021, 6.105]	3.084
5	4.967 ± 0.517	[3.682, 6.252]	2.570
10	4.687 ± 0.408	[3.674, 5.699]	2.026
12	4.530 ± 0.333	[3.703, 5.357]	1.654

Acknowledgments and Disclosure of Funding

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

References

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6382–6393, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

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

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

Jean-Michel Lasry and Pierre-Louis Lions. Mean field games. *Japanese Journal of Mathematics*, 2 (1):229–260, 2007.

Alain Bensoussan, Jens Frehse, Phillip Yam, et al. *Mean field games and mean field type control theory*, volume 101. Springer, 2013.

Haotian Gu, Xin Guo, Xiaoli Wei, and Renyuan Xu. Mean-field controls with q-learning for cooperative marl: Convergence and complexity analysis, 2021. URL `https://arxiv.org/abs/2002.04131`.

Yaodong Yang, Rui Luo, Minne Li, Ming Zhou, Weinan Zhang, and Jun Wang. Mean field multi-agent reinforcement learning, 2020. URL `https://arxiv.org/abs/1802.05438`.

Kai Shao, Jiacheng Shen, Chijie An, and Mathieu Laurière. Reinforcement learning for finite space mean-field type games, 2024a. URL `https://arxiv.org/abs/2409.18152`.

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

Samuel Jose Octavio Gaviria Alzate, Wilder Valencia Sánchez, and Elkin Arias Arias. A socio-constructivist framework for tactical development in team sports: fostering critical thinking through collaborative learning. *Frontiers in Psychology*, 16:1610750, 2025.

Joel Z Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. Multi-agent reinforcement learning in sequential social dilemmas. *arXiv preprint arXiv:1702.03037*, 2017.

Jalal Arabneydi and Aditya Mahajan. Team optimal control of coupled subsystems with mean-field sharing. In *53rd IEEE Conference on Decision and Control*, pages 1669–1674. IEEE, 2014.

Julien Perolat, Sarah Perrin, Romuald Elie, Mathieu Laurière, Georgios Piliouras, Matthieu Geist, Karl Tuyls, and Olivier Pietquin. Scaling up mean field games with online mirror descent, 2021. URL `https://arxiv.org/abs/2103.00623`.

Yue Guan, Mohammad Afshari, and Panagiotis Tsiotras. Zero-sum games between mean-field teams: Reachability-based analysis under mean-field sharing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(9):9731–9739, Mar. 2024a. doi: 10.1609/aaai.v38i9.28831. URL `https://ojs.aaai.org/index.php/AAAI/article/view/28831`.

Jayakumar Subramanian, Akshat Kumar, and Aditya Mahajan. Mean-field games among teams. *arXiv preprint arXiv:2310.12282*, 2023.

Kai Cui and Heinz Koeppl. Approximately solving mean field games via entropy-regularized deep reinforcement learning, 2022. URL `https://arxiv.org/abs/2102.01585`.

Daisuke Inoue, Yuji Ito, Takahito Kashiwabara, Norikazu Saito, and Hiroaki Yoshida. Model predictive mean field games for controlling multi-agent systems. In *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 982–987. IEEE, 2021.

Sarah Perrin, Mathieu Laurière, Julien Pérolat, Matthieu Geist, Romuald Élie, and Olivier Pietquin. Mean field games flock! the reinforcement learning way, 2021. URL https://arxiv.org/abs/2105.07933.

Patrick Benjamin and Alessandro Abate. Networked communication for mean-field games with function approximation and empirical mean-field estimation, 2025a. URL https://arxiv.org/abs/2408.11607.

Kai Cui, Sascha Hauck, Christian Fabian, and Heinz Koeppl. Learning decentralized partially observable mean field control for artificial collective behavior, 2024. URL https://arxiv.org/abs/2307.06175.

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

Daniel S Bernstein, Robert Givan, Neil Immerman, and Shlomo Zilberstein. The complexity of decentralized control of markov decision processes. *Mathematics of operations research*, 27(4): 819–840, 2002.

Kai Shao, Jiacheng Shen, Chijie An, and Mathieu Laurière. Reinforcement learning for finite space mean-field type games. *arXiv preprint arXiv:2409.18152*, 2024b.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Shengyi Huang, Rousslan Fernand Julien Dossa, Antonin Raffin, Anssi Kanervisto, and Weixun Wang. The 37 implementation details of proximal policy optimization, 2022. URL https://iclr-blog-track.github.io/2022/03/25/ppo-implementation-details/. https://iclr-blog-track.github.io/2022/03/25/ppo-implementation-details/.

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

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation, 2018. URL https://arxiv.org/abs/1506.02438.

Zixuan Chen, Xialin He, Yen-Jen Wang, Qiayuan Liao, Yanjie Ze, Zhongyu Li, S Shankar Sastry, Jiajun Wu, Koushil Sreenath, Saurabh Gupta, et al. Learning smooth humanoid locomotion through lipschitz-constrained policies. *arXiv preprint arXiv:2410.11825*, 2024.

Patrick Benjamin and Alessandro Abate. Networked communication for decentralised cooperative agents in mean-field control. *arXiv preprint arXiv:2503.09400*, 2025b.

Tabish Rashid, Mikayel Samvelyan, Christian Schroeder De Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Monotonic value function factorisation for deep multi-agent reinforcement learning. *Journal of Machine Learning Research*, 21(178):1–51, 2020.

Bei Peng, Tabish Rashid, Christian A. Schroeder de Witt, Pierre-Alexandre Kamienny, Philip H. S. Torr, Wendelin Böhmer, and Shimon Whiteson. Facmac: Factored multi-agent centralised policy gradients, 2021. URL https://arxiv.org/abs/2003.06709.

Aditya Mahajan and Ashutosh Nayyar. Sufficient statistics for linear control strategies in decentralized systems with partial history sharing. *IEEE Transactions on Automatic Control*, 60(8):2046–2056, 2015. doi: 10.1109/TAC.2015.2398884.

Sina Sanjari, Naci Saldi, and Serdar Yüksel. Nash equilibria for exchangeable team against team games, their mean field limit, and role of common randomness. *arXiv preprint arXiv:2210.07339*, 2022.

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

Batuhan Yardim and Niao He. Exploiting approximate symmetry for efficient multi-agent reinforcement learning, 2024. URL https://arxiv.org/abs/2408.15173.

Sriram Ganapathi Subramanian, Matthew E Taylor, Mark Crowley, and Pascal Poupart. Partially observable mean field reinforcement learning. *arXiv preprint arXiv:2012.15791*, 2020.

Sriram Ganapathi Subramanian, Pascal Poupart, Matthew E. Taylor, and Nidhi Hegde. Multi type mean field reinforcement learning, 2022a. URL https://arxiv.org/abs/2002.02513.

Sriram Ganapathi Subramanian, Matthew E. Taylor, Mark Crowley, and Pascal Poupart. Decentralized mean field games, 2022b. URL https://arxiv.org/abs/2112.09099.

Marc Lanctot, Vinicius Zambaldi, Audrūnas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Pérolat, David Silver, and Thore Graepel. A unified game-theoretic approach to multiagent reinforcement learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4193–4206, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Max Olan Smith, Thomas Anthony, and Michael P. Wellman. Iterative empirical game solving via single policy best response, 2021. URL https://arxiv.org/abs/2106.01901.

René Carmona, Mathieu Laurière, and Zongjun Tan. Model-free mean-field reinforcement learning: Mean-field mdp and mean-field q-learning, 2021. URL https://arxiv.org/abs/1910.12802.

Xin Guo, Anran Hu, Matteo Santamaria, Mahan Tajrobehkar, and Junzi Zhang. Mfglib: A library for mean-field games. *arXiv preprint arXiv:2304.08630*, 2023.

J Terry, Benjamin Black, Nathaniel Grammel, Mario Jayakumar, Ananth Hari, Ryan Sullivan, Luis S Santos, Clemens Dieffendahl, Caroline Horsch, Rodrigo Perez-Vicente, et al. Pettingzoo: Gym for multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 34: 15032–15043, 2021.

Lianmin Zheng, Jiacheng Yang, Han Cai, Ming Zhou, Weinan Zhang, Jun Wang, and Yong Yu. Magent: A many-agent reinforcement learning platform for artificial collective intelligence. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Pam Richards, Dave Collins, and Duncan RD Mascarenhas. Developing rapid high-pressure team decision-making skills. the integration of slow deliberate reflective learning within the competitive performance environment: A case study of elite netball. *Reflective Practice*, 13(3):407–424, 2012.

Lin Xiao, Stephen Boyd, and Sanjay Lall. Distributed average consensus with time-varying metropolis weights. *Automatica*, 1:1–4, 2006.

Eduardo Pignatelli, Johan Ferret, Matthieu Geist, Thomas Mesnard, Hado van Hasselt, Olivier Pietquin, and Laura Toni. A survey of temporal credit assignment in deep reinforcement learning, 2024. URL https://arxiv.org/abs/2312.01072.

Angelia Nedić and Ji Liu. On convergence rate of weighted-averaging dynamics for consensus problems. *IEEE Transactions on Automatic Control*, 62(2):766–781, 2016.

Randy Freeman and Petar V Kokotovic. *Robust Nonlinear Control Design: State-space and Lyapunov Techniques*. Springer Science & Business Media, 2008.

Daishi Kuroiwa. Convexity for set-valued maps. *Applied Mathematics Letters*, 9(2):97–101, 1996.

Guillermo Owen. *Game Theory*. Emerald Group Publishing, 2013.

Yue Guan, Mohammad Afshari, and Panagiotis Tsiotras. Zero-sum games between large-population teams: Reachability-based analysis under mean-field sharing, 2024b. URL https://arxiv.org/abs/2303.12243.

Kai Cui, Christian Fabian, Anam Tahir, and Heinz Koeppl. Major-minor mean field multi-agent reinforcement learning. *arXiv preprint arXiv:2303.10665*, 2023.

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

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

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. URL https://arxiv.org/abs/1412.6980.

TES Raghavan. Zero-sum two-person games. *Handbook of game theory with economic applications*, 2:735–768, 1994.

A Related Work

Mean-Field Teams. The single-team problem was explored in Arabneydi and Mahajan [2014], where agents share a common team reward (MFC). In contrast, Mahajan and Nayyar [2015] established optimality for finite-population games *only* in the LQG setting, while Sanjari et al. [2022] analyzed a two-team setting with continuous states and actions—unlike our finite state-space formulation that directly admits the familiar MDP-type structure. Similarly, *multi-population* MFGs (MP-MFGs) have been studied in the past but often restrict agent dynamics and policies to be independent of both other agents and other population distributions, see Perolat et al. [2021] and references therein. More recently, Guan et al. [2024a] introduced *zero-sum mean-field team games* (ZS-MFTGs), modeling large-population teams that compete while cooperating internally. A common-information decomposition [Nayyar et al., 2013] enables team-size-independent learning with identical optimal team policies using MF feedback, unlike aforementioned MP-MFGs (open-loop MF policies). However, computing such policies numerically, especially in large state-action spaces using dynamic programming, is costly. By contrast, MF-MAPPO (Algorithm A) leverages shared actor–critic networks per team and uses only commonly accessible information (exact/estimated), ensuring tractability and scalability.

Mean-Field Theory and Learning. Recent advances in mean-field learning span Q-function–based methods, such as MF-Q and MF-AC [Yang et al., 2020] and DDPG-MFTG [Shao et al., 2024a], to value-function–based methods like Dec-POMFPPO [Cui et al., 2024] (MFC only). While DDPG-MFTG incorporates team games, it is restricted to simple grid worlds, unlike our focus on tightly coupled collaborative–competitive domains (ZS-MFTGs). We adopt it as a baseline and show that MF-MAPPO consistently outperforms it in stability and performance. Other related works include PMD-TD for MFGs [Yardim and He, 2024] and GAN-based ECA-Net for continuous-space attack–defense games [Wang et al., 2022], both differing in scope and structure. Moreover, Yang et al. [2020], Subramanian et al. [2020, 2022a,b] define mean fields over (neighboring) agents' actions rather than their state distribution, yielding a rich line of research that is orthogonal to our formulation. MF-MAPPO extends PPO [Schulman et al., 2017] to competitive MFTGs, using team distributions as critic inputs for scalability, a shared actor–critic per team with a single buffer for efficiency, and simultaneous team training to avoid the inefficiencies of iterative best-response methods [Lanctot et al., 2017, Smith et al., 2021]. Unlike prior MF methods that rely on infinite-population oracles [Shao et al., 2024b, Perolat et al., 2021, Carmona et al., 2021], MF-MAPPO is trained directly in finite-population simulators, making it suitable for realistic deployment. Finally, to standardize evaluation in large-scale MFTGs, we create novel MFTG benchmark environments (Constrained Rock–Paper–Scissors, Battlefield) going beyond existing ones that either focus only on MFGs [Guo et al., 2023] or omit MF coupling altogether [Terry et al., 2021, Zheng et al., 2018].

Mean-Field Estimation and Opponent Modeling. To enable reactive behavior to opponents' actions, most existing MF approaches assume centralized or exact knowledge of the opponent's MF, which is rarely practical. While Cui et al. [2024] considers partial observability, it is limited to MFC. Estimation methods such as kernel density estimation [Inoue et al., 2021] or normalizing flows [Perrin et al., 2021] struggle with discrete spaces and limited agent-visibility. Communication-based methods [Benjamin and Abate, 2025a,b] address these constraints but assume uniform estimates for unobserved states and rely on perfect multi-round communication while being restricted to fully cooperative or competitive regimes; noisy variants may even produce invalid distributions. We instead propose Dynamic-Projected Consensus (D-PC), a constrained consensus algorithm that ensures validity, exponential convergence, and bounded deviations when paired with a *gradient-regularized* MF-MAPPO policy. Gradient regularization naturally stabilizes MF-MAPPO in partially observable MFTGs. Experiments show that D-PC matches baseline performance and even outperforms them under limited communication, especially critical in adversarial settings requiring rapid adaptation [Richards et al., 2012], enhancing robustness and fault tolerance. To our knowledge, this is the first MF estimator for opponent modeling in competitive team settings.

B Details of D-PC

Definition 4 (Visibility Graph)**.** The bipartite state-based visibility graph at time t, denoted by $\mathcal{G}_t^{\text{viz}} = (\mathcal{X}, \mathcal{Y}, \mathcal{E}_t^{\text{viz}})$, is an undirected graph where (i) The vertex sets consist of the states $x \in \mathcal{X}$ and $y \in \mathcal{Y}$; (ii) An undirected edge $(x, y) \in \mathcal{E}_t^{\text{viz}}$ exists if and only if all Blue agents at state x are

Algorithm A 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
    $u_{i,t} \sim \phi_{\theta^{\text{old}}_{\text{Blue}}}(x_{i,t}, \mu_t^{N_1}, \nu_t^{N_2})$, $v_{j,t} \sim \psi_{\theta^{\text{old}}_{\text{Red}}}(y_{j,t}, \mu_t^{N_1}, \nu_t^{N_2})$
    Step environment according to kernels (f_t, g_t)
    Collect samples $(\mathbf{x}_{t+1}^{N_1}, \mathbf{y}_{t+1}^{N_2}, \mu_{t+1}^{N_1}, \nu_{t+1}^{N_2}, \mathbf{u}_t^{N_1}, \mathbf{v}_t^{N_2}, r_t)$
  end for
  for K epochs **do**
    Update $\{\theta_{\text{Blue}}, \zeta_{\text{Blue}}\}$ and $\{\theta_{\text{Red}}, \zeta_{\text{Red}}\}$ using (4–5)
  end for
end for
Return: $(\phi_{\theta_{\textbf{Blue}}}, \psi_{\theta_{\textbf{Red}}})$

mutually visible with all Red agents at state y at time t; that is, every Blue agent at state x is visible to every Red agent at state y at time t.

We further define the $|\mathcal{X}| \times |\mathcal{X}|$ matrix M_t as $M_t[p, q] = 1$ if $(x_p, x_q) \notin \mathcal{E}_t^{\text{viz}}$, else 0. $M_t[p, q] = 1$ indicates that state x_q is **not** visible from state x_p at time t, and vice versa.

Assumption 2. M_t is invertible.

Before describing the communication architecture, we define $\mathcal{X}_t^o \subseteq \mathcal{X}$ such that $\hat{\mu}_t(x) > 0$ if and only if $x \in \mathcal{X}_t^o$.

Definition 5 (Communication Graph)**.** The state-based communication graph at time t, denoted by $\mathcal{G}_{\text{Blue},t}^{\text{com}} = (\mathcal{X}_t^o, \mathcal{E}_t^{\text{com}})$, is an undirected graph where (i) The vertex set $\mathcal{X}_t^o$ consists of the states $x \in \mathcal{X}_t^o$; (ii) An undirected edge $(x_p, x_q) \in \mathcal{E}_t^{\text{com}}$ exists if and only if all agents at state x_p can communicate with all agents at state x_q at time t.

We denote the set of neighbors of vertex v of $\mathcal{G}_t^{\text{viz}}$ and $\mathcal{G}_{\text{Blue},t}^{\text{com}}$ respectively, as $\mathcal{N}_t^{\text{viz}}(v)$ and $\mathcal{N}_{\text{Blue},t}^{\text{com}}(v)$. $\mathcal{Y}_t^o \subseteq \mathcal{Y}$, $\mathcal{G}_{\text{Red},t}^{\text{com}}$ and $\mathcal{N}_{\text{Red},t}^{\text{com}}(v)$ are defined similarly. The following assumption on the topology of the communication graph will simplify the mean-field estimator design problem.

Assumption 3. We assume that $\mathcal{G}_{\text{Blue},t}^{\text{com}}$ and $\mathcal{G}_{\text{Red},t}^{\text{com}}$ are connected for all t.

This assumption ensures that information from all agents contributes to the mean-field estimate. Without connectivity, disconnected agents cannot communicate their states, and convergence of the estimated mean-field distribution to the true distribution is no longer guaranteed [Benjamin and Abate, 2025b]. Studying disconnected or intermittently connected communication graphs is an interesting direction for future work.

At time t, we assume that each agent $i \in [N]$ observes the population distribution of its opponent's visible states $y \in \mathcal{N}_t^{\text{viz}}(x_{i,t})$. Thus, from the perspective of agent i, the ED $\hat{\nu}_{i,t}^{N_2}$ is defined on the constrained simplex $\hat{\nu}_{i,t}^{N_2} \in \mathcal{R}(x_{i,t}) \triangleq \{\eta \in \mathcal{P}(\mathcal{Y}) \mid \eta(z) = \hat{\nu}_t^{N_2}(z), \ \forall \ z \in \mathcal{N}_t^{\text{viz}}(x_{i,t})\}$. given by $\mathcal{R}(x_{i,t})$ combines the Red team's mean-field components known with certainty by Blue agent i—i.e., the observable states $\mathcal{N}_t^{\text{viz}}(x_{i,t})$—with those that must be estimated.

Proposition 3. *For all $x \in \mathcal{X}$, $\mathcal{R}(x)$ is a closed and convex set.*

It can be easily seen that the true mean-field distribution is an element of this constrained simplex for each agent, i.e., $\hat{\nu}_t^{N_2} \in \mathcal{R}(x_{i,t})$ for all $i \in [N_1]$ and time t. The following assumption imposes uniformity of the estimate of the ED across agents in a given state.

Assumption 4. We assume that $\hat{\nu}_{i,t}^{N_2} = \hat{\nu}_{j,t}^{N_2} \triangleq \hat{\nu}_{x,t}^{N_2}$ for all agents satisfying $x_{i,t}^{N_1} = x_{j,t}^{N_1} = x$.

Definition 6. Let $x \in \mathcal{X}$. The projection operator $\Omega_{\mathcal{R}(x)} : \mathbb{R}^{|\mathcal{Y}|} \to \mathcal{R}(x)$ where $\mathcal{R}(x)$ is a closed and convex set is defined as $\Omega_{\mathcal{R}(x)}[\eta] = \omega^* \triangleq \arg\min_{\omega \in \mathcal{R}(x)} \|\eta - \omega\|_2, \ \ \eta \in \mathbb{R}^{|\mathcal{Y}|}$, where $\|\cdot\|_2$ is the standard Euclidean norm.

At each communication round, D-PC performs the following two steps:

1. Weighted average consensus:

$$\xi_{x,t}^{\tau} = \sum_{z \in \mathcal{N}_{\text{Blue},t}^{\text{com}}(x) \cup x} w_{(x,z)}^{\text{Blue}} \hat{\nu}_{z,t}^{\tau-1}, \quad \sum_{z} w_{(x,z)}^{\text{Blue}} = 1, \quad x \in \mathcal{X}^o. \tag{B.1}$$

2. Projection onto the constraint set:

$$\hat{\nu}_{x,t}^{\tau} = \Omega_{\mathcal{R}(x)} \left[\xi_{x,t}^{\tau} \right], \; x \in \mathcal{X}^o. \tag{B.2}$$

Assumption 5. The matrix $W^{\text{Blue}} \in \mathbb{R}^{|\mathcal{X}^o| \times |\mathcal{X}^o|}$ formed by the non-negative weights is symmetric and doubly-stochastic. Furthermore, it respects the sparsity structure of the communication graph $\mathcal{G}^{\text{com}}$, that is, $w_{(x,z)}^{\text{Blue}} > 0$ if and only if $z \in \mathcal{N}^{\text{com}}(x) \cup x$. W^{Red} is constructed similarly.

To satisfy Assumption 5, one may use the well-known Metropolis Matrix [Xiao et al., 2006]. We define the smallest non-zero entry of W as

$$\theta \triangleq \min_{(x,y)} \{ w_{(x,y)} > 0 \}. \tag{B.3}$$

For clarity of the D-PC pseudocode presented in Algorithm B, the estimation is presented from the Blue team's perspective which models the distribution of its opponent (Red team). One can simultaneously run this estimation algorithm from the Red team's perspective.

Algorithm B Dynamic Projected Consensus for Mean-Field Estimation (**D-PC**)

Initialize: $\hat{\mu}_{t=0}^{N_1} = \mu_{t=0}^{N_1}$ and $\hat{\nu}_{t=0}^{N_2} = \nu_{t=0}^{N2}$, identical MF-MAPPO trained Lipschitz policies (ϕ^*, ψ^*) and graphs $\mathcal{G}_{t=0}^{\text{viz}}$ and $(\mathcal{G}_{\text{Blue},t}^{\text{com}}, \mathcal{G}_{\text{Red},t}^{\text{com}})$
for $t = 0, \ldots, T$ **do**
  Update the graphs $\mathcal{G}_t^{\text{viz}}$ and $(\mathcal{G}_{\text{Blue},t}^{\text{com}}, \mathcal{G}_{\text{Red},t}^{\text{com}})$.
  Receive reward $r_t(\hat{\mu}_t^{N_1}, \hat{\nu}_t^{N_2})$
  Set $\tau = 0$
  Initialize $\hat{\nu}_{x,t}^{N_2,\tau=0}$ ($\hat{\mu}_{y,t}^{N_1,\tau=0}$) for all $x \in \mathcal{X}_t^o$ ($y \in \mathcal{Y}_t^o$) using $\mathcal{G}_{t=0}^{\text{viz}}$
  Compute matrix $W^{\text{Blue}}(W^{\text{Red}})$
  for $\tau < R_{\text{com}}$ **do**
    Communicate mean-field estimate to neighbors based on $\mathcal{G}_{\text{Blue},t}^{\text{com}}(\mathcal{G}_{\text{Red},t}^{\text{com}})$
    $\tau = \tau + 1$
    for all $x \in \mathcal{X}_t^o$ ($y \in \mathcal{Y}_t^o$) **do**
      Compute weighted average consensus (B.1)
      Project result onto constraint set (B.2)
    end for
  end for
  Return: $\hat{\nu}_{x,t}^{N_2,R_{\text{com}}}(\hat{\mu}_{y,t}^{N_1,R_{\text{com}}})$
end for

Remark 1. One could alternatively cast this problem as a Partially Observable Markov Decision Process (PO-MDP) [Bernstein et al., 2002], where the environment provides arbitrary observations. However, our focus is on policies that explicitly depend on the mean-field distributions of the two teams, for two key reasons: (a) Distinct opponent strategies induce different mean-field trajectories, and closed-loop feedback on these distributions enables an appropriate response to the strategies deployed by the opponent [Guan et al., 2024a]; and (b) Access to team-level distributions facilitates credit assignment [Pignatelli et al., 2024] in MARL by allowing agents to reason about which collective distributions are optimal. Thus, having access to MF information—particularly that of the opponent—is desirable. Yet, it is unrealistic to assume that the environment directly provides these distributions, which motivates the design of estimation algorithms proposed in this work. Finally, unlike general PO-MDP formulations that often require maintaining long observation histories, our setting involves population sizes for which such history-based tracking is computationally infeasible in terms of memory.

The following theorem establishes the convergence guarantees of D-PC.

Theorem B.1. *Under Assumption 2, D-PC is $\mathcal{O}\big(e^{-cR_{\text{com}}}\big)$-accurate.*

Sketch of Proof. The exponential convergence of D-PC at each time-step t follows directly from Theorems 6-8 in Nedić and Liu [2016] as the number of communication rounds $R_{\text{com}} \to \infty$. Therefore, for finite communication rounds, there exist constants $c_x, c_y > 0$ and $\rho_x, \rho_y \in (0,1)$ such that for any number of communication rounds R_{com} and any underlying distributions $(\hat{\mu}_t^{N_1}, \hat{\nu}_t^{N_2})$ D-PC satisfies

$$\begin{aligned} \mathrm{d_{TV}}\big(\hat{\nu}_{x,t}^{N_2}, \hat{\nu}_t^{N_2}\big) < c_x \rho_x^{R_{\text{com}}} \triangleq \epsilon_x, \quad \text{for all } x \in \mathcal{X}^o \text{ and } t \\ \mathrm{d_{TV}}\big(\hat{\mu}_{y,t}^{N_1}, \hat{\mu}_t^{N_1}\big) < c_y \rho_y^{R_{\text{com}}} \triangleq \epsilon_y, \quad \text{for all } y \in \mathcal{Y}^o \text{ and } t. \end{aligned} \tag{B.4}$$

We define $\epsilon = \max(\epsilon_x, \epsilon_y)$. □

Proof. We begin by defining set regularity.

Definition 7 (Set Regularity [Nedić and Liu, 2016])**.** Let $Z \subseteq \mathbb{R}^n$ be a nonempty set. A (finite) collection of closed convex sets $\{Y_j\}_{j=1}^J \subseteq \mathbb{R}^n$ is regular (in Euclidean norm) with respect to the set Z, if there exists a constant $r \geq 1$ such that

$$\inf_{y \in Y} \|z - y\|_2 \leq r \max_j \inf_{y \in Y_j} \|z - y\|_2, z \in Z,$$

where $Y \triangleq \bigcap_{j=1,\ldots,J} Y_j$ is non-empty. We refer to the scalar r as the regularity constant. When the preceding relation holds with $Z = \mathbb{R}^n$, we say that the sets $\{Y_j\}_{j=1}^J$ are uniformly regular.

Proposition 4. *The constraint sets $\mathcal{R}(x)$ for all $x \in \mathcal{X}^o$ are uniformly regular with regularity constant $\kappa \geq 1$, i.e.,*

$$\inf_{\nu \in \mathcal{F}} \|\omega - \nu\|_2 \leq \kappa \max_{x \in \mathcal{X}} \inf_{\nu_x \in \mathcal{R}(x)} \|\omega - \nu_x\|_2, \omega \in \mathbb{R}^{|\mathcal{X}|}.$$

Proof. Refer to Proposition 1 from Nedić and Liu [2016]. □

Following Theorem 8 from Nedić and Liu [2016], we have

$$\begin{aligned} \rho_x &= \left(1 - \frac{\theta^2}{4|\mathcal{X}^o|\mathrm{diam}(\mathcal{G}_{\text{Blue}}^{\text{com}})(\kappa+1)^2}\right), \quad c_x = \frac{1}{2}\sqrt{|\mathcal{Y}|} \sum_{x \in \mathcal{X}^o} \|\hat{\nu}_x^{\tau=0} - \xi_x^{\tau=1}\|_2^2 \\ \rho_y &= \left(1 - \frac{\theta^2}{4|\mathcal{Y}^o|\mathrm{diam}(\mathcal{G}_{\text{Red}}^{\text{com}})(\kappa+1)^2}\right), \quad c_y = \frac{1}{2}\sqrt{|\mathcal{X}|} \sum_{x \in \mathcal{Y}^o} \|\hat{\mu}_y^{\tau=0} - \xi_y^{\tau=1}\|_2^2, \end{aligned}$$

where κ is the regularity constant as in Proposition 4 and θ follows from (B.3). □

C Existence of Infinite-Population Game Value

Theorem C.1. *Suppose the reward function r_t is concave-convex in its mean-field arguments and the dynamics are independent, i.e., for all $(\mu_t, \nu_t) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})$,*

$$f_t(x_{t+1}|x_t, u_t, \mu_t, \nu_t) = \bar{f}_t(x_{t+1}|x_t, u_t), \quad g_t(y_{t+1}|y_t, v_t, \mu_t, \nu_t) = \bar{g}_t(y_{t+1}|y_t, v_t). \tag{C.5}$$

Then, the Nash value exists.

To examine the game value, we first formulate an equivalent max-min optimization leveraging a reachability analysis. At the infinite-population limit, the mean-field dynamics is deterministic, and thus selecting the local policies π_t and σ_t at time t is equivalent to selecting the desirable MFs at the next time step. Consequently, we examine the set of MFs that can be reached from the current MFs.

Definition 8. The Blue reachable set, starting from the MFs μ_t and ν_t, is defined as the set comprising all the possible next Blue team MFs μ_{t+1} that can be achieved by employing a local policy $\pi_t \in \Pi_t$. Formally,

$$\mathcal{R}_{\mu,t}(\mu_t, \nu_t) = \{\mu_{t+1} \mid \exists \pi_t \in \Pi_t \text{ s.t. } \mu_{t+1} = \mu_t F_t(\mu_t, \nu_t, \pi_t)\}. \tag{C.6}$$

Similarly, the Red reachable set is defined as

$$\mathcal{R}_{\nu,t}(\mu_t, \nu_t) = \{\nu_{t+1} \mid \exists \sigma_t \in \Sigma_t \text{ s.t. } \nu_{t+1} = \nu_t G_t(\mu_t, \nu_t, \sigma_t)\}. \tag{C.7}$$

We will regard the reachable sets as set-valued functions (correspondences) [Freeman and Kokotovic, 2008]. In this case, we write $\mathcal{R}_{\mu,t} : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y}) \rightsquigarrow \mathcal{P}(\mathcal{X})$, and similarly $\mathcal{R}_{\nu,t} : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y}) \rightsquigarrow \mathcal{P}(\mathcal{Y})$.

We can now change the optimization domains in (3) from the policy spaces to the corresponding reachable sets.

$$\begin{aligned}
&\underline{J}^*_{\infty,T}(\mu_T, \nu_T) = r_T(\mu_T, \nu_T) \\
&\underline{J}^*_{\infty,t}(\mu_t, \nu_t) = r_t(\mu_t, \nu_t) + \max_{\mu_{t+1} \in \mathcal{R}_{\mu,t}(\mu_t, \nu_t)} \min_{\nu_{t+1} \in \mathcal{R}_{\nu,t}(\mu_t, \nu_t)} \underline{J}^*_{\infty,t+1}(\mu_{t+1}, \nu_{t+1}), \quad t = 0, \dots, T-1.
\end{aligned} \tag{C.8}$$

To examine the optimal value in the infinite-population limit, we first examine the convexity of the reachability correspondences.

Definition 9 ([Kuroiwa, 1996])**.** Let $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ be convex sets. The correspondence $\Gamma : \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{Z}$ is convex with respect to $\mathcal{X}$ if, for all $x_1, x_2 \in \mathcal{X}$, $y \in \mathcal{Y}$, $z_1 \in \Gamma(x_1, y)$, $z_2 \in \Gamma(x_2, y)$, and $\lambda \in (0, 1)$, there exists $z \in \Gamma(\lambda x_1 + (1-\lambda)x_2, y)$ such that

$$z = \lambda z_1 + (1-\lambda) z_2.$$

Remark 2. The above definition of convexity is equivalent to the following set inclusion

$$\Gamma(\lambda x_1 + (1-\lambda)x_2, y) \supseteq \lambda\Gamma(x_1, y) + (1-\lambda)\Gamma(x_2, y), \tag{C.9}$$

for all $\lambda \in (0, 1)$, $x_1, x_2 \in \mathcal{X}$ and $y \in \mathcal{Y}$, where the Minkowski sum is defined as

$$\lambda\Gamma(x_1, y) + (1-\lambda)\Gamma(x_2, y) = \{z | z = \lambda z_1 + (1-\lambda) z_2 \text{ where } z_1 \in \Gamma(x_1, y), z_2 \in \Gamma(x_2, y)\}.$$

Definition 10. The correspondence $\Gamma : \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{Z}$ is constant with respect to $\mathcal{Y}$ if

$$\Gamma(x, y_1) = \Gamma(x, y_2) \quad \forall x \in \mathcal{X},\ \forall y_1, y_2 \in \mathcal{Y}. \tag{C.10}$$

We have the following concave-convex result for a max-min marginal function defined in (C.11).

$$f(x, y) = \max_{p \in \Gamma(x,y)} \min_{q \in \Theta(x,y)} g(p, q) \tag{C.11}$$

Lemma 1. *Consider two compact-valued correspondences $\Gamma : \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{X}$ and $\Theta : \mathcal{X} \times \mathcal{Y} \rightsquigarrow \mathcal{Y}$. Let Γ be convex with respect to $\mathcal{X}$ and constant with respect to $\mathcal{Y}$, and let Θ be constant with respect to $\mathcal{X}$ and convex with respect to $\mathcal{Y}$. Let $g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be concave-convex. Then, the max-min marginal function $f(x, y) = \max_{p \in \Gamma(x,y)} \min_{q \in \Theta(x,y)} g(p, q)$ is also concave-convex.*

Proof. Fix $y \in \mathcal{Y}$, and consider $x_1, x_2 \in \mathcal{X}$ and $\lambda \in (0, 1)$. Denote $x = \lambda x_1 + (1-\lambda)x_2$. It follows that

$$\begin{aligned}
f(\lambda x_1 + (1-\lambda)x_2, y) &= \max_{p \in \Gamma(\lambda x_1 + (1-\lambda)x_2, y)} \min_{q \in \Theta(\lambda x_1 + (1-\lambda)x_2, y)} g(p, q) \\
&\overset{\text{(i)}}{\geq} \max_{p \in \lambda\Gamma(x_1, y) + (1-\lambda)\Gamma(x_2, y)} \min_{q \in \Theta(x,y)} g(p, q) \\
&= \max_{\substack{p_1 \in \Gamma(x_1, y) \\ p_2 \in \Gamma(x_2, y)}} \min_{q \in \Theta(x,y)} g(\lambda p_1 + (1-\lambda)p_2, q) \\
&\overset{\text{(ii)}}{\geq} \max_{\substack{p_1 \in \Gamma(x_1, y) \\ p_2 \in \Gamma(x_2, y)}} \min_{q \in \Theta(x,y)} (\lambda g(p_1, q) + (1-\lambda) g(p_2, q)) \\
&\overset{\text{(iii)}}{\geq} \max_{\substack{p_1 \in \Gamma(x_1, y) \\ p_2 \in \Gamma(x_2, y)}} \Big(\underbrace{\min_{q_1 \in \Theta(x,y)} \lambda g(p_1, q_1)}_{A(p_1)} + \underbrace{\min_{q_2 \in \Theta(x,y)} (1-\lambda) g(p_2, q_2)}_{B(p_2)} \Big) \\
&= \lambda \max_{p_1 \in \Gamma(x_1, y)} \min_{q_1 \in \Theta(x,y)} g(p_1, q_1) + (1-\lambda) \max_{p_2 \in \Gamma(x_2, y)} \min_{q_2 \in \Theta(x,y)} g(p_2, q_2) \\
&\overset{\text{(iv)}}{=} \lambda \max_{p_1 \in \Gamma(x_1, y)} \min_{q_1 \in \Theta(x_1, y)} g(p_1, q_1) + (1-\lambda) \max_{p_2 \in \Gamma(x_2, y)} \min_{q_2 \in \Theta(x_2, y)} g(p_2, q_2) \\
&= \lambda f(x_1, y) + (1-\lambda) f(x_2, y),
\end{aligned}$$

where inequality (i) holds from restricting the maximization domain using the convexity of Γ in (C.9); inequality (ii) holds from the assumption that g is concave with respect to its p-argument; inequality (iii) is the result of distributing the minimization; equality (iv) is due to Θ being constant with respect to $\mathcal{X}$. The above result implies the concavity of f with respect to its x-argument.

Fix $x \in \mathcal{X}$, and let $y_1, y_2 \in \mathcal{Y}$ and $\lambda \in (0,1)$. Denote $y = \lambda y_1 + (1-\lambda)y_2$. Then,

$$\begin{aligned}
f(x, \lambda y_1 + (1-\lambda)y_2) &= \max_{p \in \Gamma(x, \lambda y_1 + (1-\lambda)y_2)} \ \min_{q \in \Theta(x, \lambda y_1 + (1-\lambda)y_2)} g(p,q) \\
&\leq \max_{p \in \Gamma(x,y)} \ \min_{\substack{q_1 \in \Theta(x,y_1) \\ q_2 \in \Theta(x,y_2)}} g(p, \lambda q_1 + (1-\lambda)q_2) \\
&\leq \max_{p \in \Gamma(x,y)} \ \min_{\substack{q_1 \in \Theta(x,y_1) \\ q_2 \in \Theta(x,y_2)}} (\lambda g(p,q_1) + (1-\lambda) g(p,q_2)) \\
&\leq \lambda \max_{p_1 \in \Gamma(x,y)} \ \min_{q_1 \in \Theta(x,y_1)} g(p_1,q_1) + (1-\lambda) \max_{p_2 \in \Gamma(x,y)} \ \min_{q_2 \in \Theta(x,y_2)} g(p_2,q_2) \\
&= \lambda \max_{p_1 \in \Gamma(x,y_1)} \ \min_{q_1 \in \Theta(x,y_1)} g(p_1,q_1) + (1-\lambda) \max_{p_2 \in \Gamma(x,y_2)} \ \min_{q_2 \in \Theta(x,y_2)} g(p_2,q_2) \\
&= \lambda f(x,y_1) + (1-\lambda) f(x,y_2),
\end{aligned}$$

which implies that f is convex with respect to its y-argument. □

The following lemma shows the convexity of the reachability correspondences $\mathcal{R}_{\mu,t}$ and $\mathcal{R}_{\nu,t}$ under independent dynamics.

Lemma 2. *Under the independent dynamics in* (C.5)*, the reachability correspondences satisfy*

$$\mathcal{R}_{\mu,t}(\mu_t, \nu_t) = \mathcal{R}_{\mu,t}(\mu_t, \nu_t'), \quad \forall \mu_t \in \mathcal{P}(\mathcal{X}), \ \ \forall \nu_t, \nu_t' \in \mathcal{P}(\mathcal{Y}), \tag{C.12a}$$

$$\mathcal{R}_{\nu,t}(\mu_t, \nu_t) = \mathcal{R}_{\nu,t}(\mu_t', \nu_t), \quad \forall \mu_t, \mu_t' \in \mathcal{P}(\mathcal{X}), \ \ \forall \nu_t \in \mathcal{P}(\mathcal{Y}). \tag{C.12b}$$

Furthermore, for all $\lambda \in [0,1]$,

$$\mathcal{R}_{\mu,t}\big(\lambda\mu_t + (1-\lambda)\mu_t', \nu_t\big) = \lambda \mathcal{R}_{\mu,t}\big(\mu_t, \nu_t\big) + (1-\lambda)\mathcal{R}_{\mu,t}\big(\mu_t', \nu_t\big), \quad \forall \mu_t, \mu_t' \in \mathcal{P}(\mathcal{X}), \ \ \forall \nu_t \in \mathcal{P}(\mathcal{Y}), \tag{C.13a}$$

$$\mathcal{R}_{\nu,t}\big(\mu_t, \lambda\nu_t + (1-\lambda)\nu_t'\big) = \lambda \mathcal{R}_{\nu,t}\big(\mu_t, \nu_t\big) + (1-\lambda)\mathcal{R}_{\nu,t}\big(\mu_t, \nu_t'\big), \quad \forall \mu_t \in \mathcal{P}(\mathcal{X}), \ \ \forall \nu_t, \nu_t' \in \mathcal{P}(\mathcal{Y}). \tag{C.13b}$$

Proof. Due to symmetry, we only prove the results for the Blue reachability correspondence. The results for the Red team can be obtained through a similar argument.

Consider an arbitrary pair of distributions $\mu_t \in \mathcal{P}(\mathcal{X})$ and $\nu_t \in \mathcal{P}(\mathcal{Y})$. If $\mu_{t+1} \in \mathcal{R}_{\mu,t}(\mu_t, \nu_t)$ under the independent dynamics in (C.5), there exists a local policy $\pi_t \in \Pi_t$, such that

$$\mu_{t+1}(x') = \sum_{x \in \mathcal{X}} \Big[\sum_{u \in \mathcal{U}} \bar{f}_t(x'|x,u)\pi_t(u|x) \Big] \mu_t(x),$$

which is independent of ν_t. Consequently,

$$\mathcal{R}_{\mu,t}(\mu_t, \nu_t) = \mathcal{R}_{\mu,t}(\mu_t, \nu_t'), \quad \forall\ \mu_t \in \mathcal{P}(\mathcal{X}) \text{ and } \forall\ \nu_t, \nu_t' \in \mathcal{P}(\mathcal{Y}).$$

Next, we prove the property in (C.13a). For the $\subseteq$ direction, consider two Blue mean fields $\mu_t, \mu_t' \in \mathcal{P}(\mathcal{X})$ and a Red mean field $\nu_t \in \mathcal{P}(\mathcal{Y})$. Let $\bar{\mu}_{t+1} \in \mathcal{R}_{\mu,t}(\lambda\mu_t + (1-\lambda)\mu_t', \nu_t)$. From the definition of the reachable set, there exists a local policy $\pi_t \in \Pi_t$, such that

$$\begin{aligned}
\bar{\mu}_{t+1}(x') &= \sum_{x \in \mathcal{X}} \Big[\sum_{u \in \mathcal{U}} \bar{f}_t(x'|x,u)\pi_t(u|x) \Big] (\lambda\mu_t(x) + (1-\lambda)\mu_t'(x)) \\
&= \lambda \underbrace{\sum_{x \in \mathcal{X}} \Big[\sum_{u \in \mathcal{U}} \bar{f}_t(x'|x,u)\pi_t(u|x) \Big] \mu_t(x)}_{\mu_{t+1}(x')} + (1-\lambda) \underbrace{\sum_{x \in \mathcal{X}} \Big[\sum_{u \in \mathcal{U}} \bar{f}_t(x'|x,u)\pi_t(u|x) \Big] \mu_t'(x)}_{\mu_{t+1}'(x')},
\end{aligned}$$

where $\mu_{t+1} \in \mathcal{R}_{\mu,t}(\mu_t, \nu_t)$ and $\mu_{t+1}' \in \mathcal{R}_{\mu,t}(\mu_t', \nu_t)$. Hence, we have

$$\mathcal{R}_{\mu,t}(\lambda\mu_t + (1-\lambda)\mu_t', \nu_t) \subseteq \lambda \mathcal{R}_{\mu,t}\big(\mu_t, \nu_t\big) + (1-\lambda)\mathcal{R}_{\mu,t}\big(\mu_t', \nu_t\big).$$

Let now $\bar{\mu}_{t+1} = \lambda\mu_{t+1} + (1-\lambda)\mu'_{t+1}$, where $\mu_{t+1} \in \mathcal{R}_{\mu,t}(\mu_t, \nu_t)$ and $\mu'_{t+1} \in \mathcal{R}_{\mu,t}(\mu'_t, \nu_t)$. Further, define $\bar{\mu}_t = \lambda\mu_t + (1-\lambda)\mu'_t$. From the definition of the reachable set, there exist local policies π_t and π'_t such that

$$
\begin{aligned}
\bar{\mu}_{t+1} &= \lambda \sum_{x\in\mathcal{X}} \Big[\sum_{u\in\mathcal{U}} \bar{f}_t(x'|x,u)\pi_t(u|x) \Big] \mu_t(x) + (1-\lambda) \sum_{x\in\mathcal{X}} \sum_{u\in\mathcal{U}} \bar{f}_t(x'|x,u)\pi'_t(u|x) \Big] \mu'_t(x) \\
&= \sum_{x\in\mathcal{X}} \sum_{u\in\mathcal{U}} \bar{f}_t(x'|x,u) \Big[\lambda\pi_t(u|x)\mu_t(x) + (1-\lambda)\pi'_t(u|x)\mu'_t(x) \Big] \\
&= \sum_{x\in\mathcal{X}} \sum_{u\in\mathcal{U}} \mathbb{1}_{\bar{\mu}_t(x)>0}\ \bar{f}_t(x'|x,u) \Big[\lambda\pi_t(u|x)\mu_t(x) + (1-\lambda)\pi'_t(u|x)\mu'_t(x) \Big] \\
&+ \sum_{x\in\mathcal{X}} \sum_{u\in\mathcal{U}} \mathbb{1}_{\bar{\mu}_t(x)=0}\ \bar{f}_t(x'|x,u) \underbrace{\Big[\lambda\pi_t(u|x)\mu_t(x) + (1-\lambda)\pi'_t(u|x)\mu'_t(x) \Big]}_{=0\,[4]} \\
&= \sum_{x\in\mathcal{X}} \sum_{u\in\mathcal{U}} \mathbb{1}_{\bar{\mu}_t(x)>0}\ \bar{f}_t(x'|x,u) \frac{\lambda\pi_t(u|x)\mu_t(x) + (1-\lambda)\pi'_t(u|x)\mu'_t(x)}{\bar{\mu}_t(x)} \bar{\mu}_t(x) \\
&+ \sum_{x\in\mathcal{X}} \sum_{u\in\mathcal{U}} \mathbb{1}_{\bar{\mu}_t(x)=0}\ \bar{f}_t(x'|x,u) \underbrace{\Big[\lambda\pi_t(u|x) + (1-\lambda)\pi'_t(u|x) \Big] \bar{\mu}_t(x)}_{=0} \\
&= \sum_{x\in\mathcal{X}} \left[\sum_{u\in\mathcal{U}} \bar{f}_t(x'|x,u)\bar{\pi}_t(u|x) \right] \bar{\mu}_t(x),
\end{aligned}
$$

where the "averaged" local policy $\bar{\pi}_t$ is given by

$$
\bar{\pi}_t(u|x) = \mathbb{1}_{\bar{\mu}_t(x)>0} \frac{\lambda\pi_t(u|x)\mu_t(x) + (1-\lambda)\pi'_t(u|x)\mu'_t(x)}{\lambda\mu_t(x) + (1-\lambda)\mu'_t(x)} + \mathbb{1}_{\bar{\mu}_t(x)=0} \Big(\lambda\pi_t(u|x) + (1-\lambda)\pi'_t(u|x) \Big).
$$

Consequently, we have

$$
\bar{\mu}_{t+1} \in \mathcal{R}_{\mu,t}(\lambda\mu_t + (1-\lambda)\mu'_t, \nu_t).
$$

Hence,

$$
\mathcal{R}_{\mu,t}(\lambda\mu_t + (1-\lambda)\mu'_t, \nu_t) \supseteq \lambda\mathcal{R}_{\mu,t}\big(\mu_t, \nu_t\big) + (1-\lambda)\mathcal{R}_{\mu,t}\big(\mu'_t, \nu_t\big).
$$

□

Proof. We can now prove Theorem C.1. The game value at time T is directly the mean-field reward, which is assumed to be concave-convex. Hence, one can provide an inductive proof leveraging Lemma 1 and Lemma 2 to show that both the lower and upper game values are concave-convex at each time step. Note that the lower game value in (C.8) has the form of a max-min marginal function. Finally, as the reachable sets are compact and convex, one can apply the minimax theorem [Owen, 2013], which guarantees that the max-min optimal value is the same as the min-max optimal value, and thus ensures the existence of the game value. □

D Results on Minimally-Informed Critics

Proposition 5. *Let $\mu_t^{N_1}$ and $\nu_t^{N_2}$ denote the EDs induced by identical Blue and Red team policies. Then, 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 $i \in [N_1]$, the individual critic satisfies $V_{i,t}^{N_1,\phi_t}(x_{i,t}, \mu_t^{N_1}, \nu_t^{N_2}) = V_{\mathrm{Blue},t}^{N_1,\phi_t}(\mu_t^{N_1}, \nu_t^{N_2})$, where $V_{\mathrm{Blue},t}^{N_1,\phi_t}$ is the team-level value function.*

We present the proof for a single team (Blue). The proof can be directly extended to the opponent team (Red). Let the critic value function at time t for each Blue team agent i, following policy ϕ_t be

[4] Since the condition $\bar{\mu}_t(x) = 0$ together with $\lambda \in (0,1)$ implies that $\mu_t(x) = 0$ and $\mu'_t(x) = 0$.

given by $V_{i,t}^{N_1,\phi_t}(x_t^i, \mu_t, \nu_t)$ where ζ^i represents its parameterization, namely,

$$V_{i,\zeta_i,t}^{N_1,(\phi,\psi)}(x_t^i, \mu_t, \nu_t) = \mathbb{E}_{\phi,\psi}\left[\sum_{t'=t}^{\infty} \gamma^{t'-t} r^{t'}(\mu_{t'}, \nu_{t'})\right].$$

Note that RHS of the above expression is the closed form representation for the value function and is independent of x_t^i. Thus, for all $i \in \{1, 2, \ldots, N_1\}$ we can define

$$V_{\zeta_{\text{Blue}},t}^{N_1,(\phi,\psi)}(\mu_t, \nu_t) \triangleq V_{i,\zeta_i,t}^{N_1,(\phi,\psi)}(x_t^i, \mu_t, \nu_t),$$

where $V_{\zeta_{\text{Blue}},t}^{N_1,(\phi,\psi)}(\mu_t, \nu_t) \triangleq \mathbb{E}_{\phi,\psi}\left[\sum_{t'=t}^{\infty} \gamma^{t'-t} r^{t'}(\mu_{t'}, \nu_{t'}) | \mu_t, \nu_t\right]$ and the result follows.

E Proof of Theorem 3.1

Theorem E.1 (Guan et al. [2024a])**.** *The optimal identical Blue team policy $\phi_\infty^* \in \Phi$ obtained from the equivalent zero-sum coordinator game is ϵ-optimal Blue team policy. Formally, for all joint states $\mathbf{x}^{N_1}$ and $\mathbf{y}^{N_2}$,*

$$\min_{\psi^{N_2} \in \Psi^{N_2}} J^{N,\phi_\infty^*,\psi^{N_2}}(\mathbf{x}^{N_1}, \mathbf{y}^{N_2}) \geq \underline{J}^{N*}(\mathbf{x}^{N_1}, \mathbf{y}^{N_2}) - \mathcal{O}\Big(\frac{1}{\sqrt{\underline{N}}}\Big), \text{ where } \underline{N} = \min\{N_1, N_2\}. \tag{E.14}$$

Theorem 3.1. *Let $\tilde{\phi}^*$ be a solution to the finite-population optimization restricted to identical policies, i.e., $\max_{\phi\in\Phi} \min_{\psi\in\Psi} J^{N,\phi,\psi}(\mathbf{x}_0^{N_1}, \mathbf{y}_0^{N_2})$. Then, the exploitability of $\tilde{\phi}^*$ satisfies*

$$\min_{\psi^{N_2} \in \Psi^{N_2}} J^{N,\tilde{\phi}^*,\psi^{N_2}}(\mathbf{x}^{N_1}, \mathbf{y}^{N_2}) \geq \underline{J}^{N*}(\mathbf{x}^{N_1}, \mathbf{y}^{N_2}) - \mathcal{O}(1/\sqrt{\underline{N}}), \quad \forall \mathbf{x}^{N_1} \in \mathcal{X}^{N_1}, \mathbf{y}^{N_2} \in \mathcal{Y}^{N_2}.$$

Sketch of Proof. The proof proceeds in two steps: (i) we first show that the optimal identical Blue policy obtained from the *finite-N system* is ϵ-optimal against identical Red exploitation in the *infinite-population limit*; (ii) we then invoke the approximation result in Guan et al. [2024b] to show that this identical Blue policy remains ϵ-optimal against non-identical Red policies in the original finite-population game. □

Proof. We begin by noting two key lemmas that will be used to prove the theorem. The first lemma follows directly from analysis and convex optimization.

Lemma 3. *Let $f(x, y)$ and $g(x, y)$ be two functions defined for all x, y in some compact convex set. Suppose that for all x, y, the functions are ϵ-close: $|f(x, y) - g(x, y)| < \epsilon$, which is equivalent to,*

$$f(x, y) - \epsilon < g(x, y) < f(x, y) + \epsilon$$

Consider the max-min values of both functions $f^* = \max_x \min_y f(x, y)$ and $g^* = \max_x \min_y g(x, y)$. Let x_f^* be the maximizer that attains f^*. Then, the performance of the optimal policy x_f^* for f, when evaluated under g, is 2ϵ-optimal. Specifically, $g^* - 2\epsilon \leq \min_y g(x_f^*, y) \leq g^*$.

Lemma 4. *For all identical team policies (ϕ, ψ) we have the following inequality for all joint states $\mathbf{x}^{N_1} \in \mathcal{X}^{N_1}$ and $\mathbf{y}^{N_2} \in \mathcal{Y}^{N_2}$,*

$$|J^{N,\phi,\psi}(\mathbf{x}^{N_1}, \mathbf{y}^{N_2}) - J_\infty^{\phi,\psi}(\mu^{N_1}, \nu^{N_2})| \leq \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big), \tag{E.15}$$

where $\mu^{N_1} = \text{Emp}_\mu(\mathbf{x}^{N_1})$, $\nu^{N_2} = \text{Emp}_\nu(\mathbf{y}^{N_2})$.

Proof. We can prove this by mathematical induction. We show that for any pair of identical policies (ϕ, ψ) we have the following inequality for all joint states $\mathbf{x}^{N_1} \in \mathcal{X}^{N_1}$ and $\mathbf{y}^{N_2} \in \mathcal{Y}^{N_2}$,

$$J^{N,\phi,\psi}(\mathbf{x}^{N_1}, \mathbf{y}^{N_2}) \geq J_\infty^{\phi,\psi}(\mu^{N_1}, \nu^{N_2}) - \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big), \tag{E.16}$$

where $\mu^{N_1} = \text{Emp}_\mu(\mathbf{x}^{N_1})$, $\nu^{N_2} = \text{Emp}_\nu(\mathbf{y}^{N_2})$. The proof is constructed based on induction.

Base case: At the terminal timestep T, since there is no decision to be made, both value functions are equal to the terminal reward and are thus the same. Formally, for all $\mathbf{x}_T^{N_1} \in \mathbf{X}^{N_1}$ and $\mathbf{y}_T^{N_2} \in \mathbf{Y}^{N_2}$,

$$J_T^{N,\phi,\psi}(\mathbf{x}_T^{N_1}, \mathbf{y}_T^{N_2}) = J_{\infty,T}^{\phi,\psi}(\mu_T^{N_1}, \nu_T^{N_2}).$$

For simplicity, we do not emphasize the correspondence between the joint states and the EDs for the rest of the proof, as it is clear from the context.

Inductive hypothesis: Assume that at $t+1$, the following holds for all joint states $\mathbf{x}_{t+1}^{N_1} \in \mathbf{X}^{N_1}$ and $\mathbf{y}_{t+1}^{N_2} \in \mathbf{Y}^{N_2}$:

$$J_{t+1}^{N,\phi,\psi}(\mathbf{x}_{t+1}^{N_1}, \mathbf{y}_{t+1}^{N_2}) \geq J_{\infty,t+1}^{\phi,\psi}(\mu_{t+1}^{N_1}, \nu_{t+1}^{N_2}) - \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big). \tag{E.17}$$

Induction: At timestep t, consider an arbitrary pair of joint states $\mathbf{x}_t^{N_1} \in \mathbf{X}^{N_1}$ and $\mathbf{y}_t^{N_2} \in \mathbf{Y}^{N_2}$, and their corresponding EDs $\mu_t^{N_1}$ and $\nu_t^{N_2}$. Then, for all joint states $\mathbf{x}_t^{N_1} \in \mathbf{X}^{N_1}$ and $\mathbf{y}_t^{N_2} \in \mathbf{Y}^{N_2}$, we have

$$\begin{aligned}
&J_t^{N,\phi,\psi}(\mathbf{x}_t^{N_1}, \mathbf{y}_t^{N_2}) \\
&= r_t(\mu_t^{N_1}, \nu_t^{N_2}) + \mathbb{E}_{\phi,\psi}\Big[J_{t+1}^{N,\phi,\psi}(\mathbf{X}_{t+1}^{N_1}, \mathbf{Y}_{t+1}^{N_2})\Big] \\
&\overset{\text{(i)}}{\geq} r_t(\mu_t^{N_1}, \nu_t^{N_2}) + \mathbb{E}_{\phi,\psi}\Big[J_{\infty,t+1}^{\phi,\psi}(\mathcal{M}_{t+1}^{N_1}, \mathcal{N}_{t+1}^{N_2})\Big] - \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big) \\
&= r_t(\mu_t^{N_1}, \nu_t^{N_2}) - \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big) + \mathbb{E}_{\phi,\psi}\Big[J_{\infty,t+1}^{\phi,\psi}(\mathcal{M}_{t+1}^{N_1}, \mathcal{N}_{t+1}^{N_2}) \\
&\qquad\qquad - J_{\infty,t+1}^{\phi,\psi}(\mu_{t+1}^{\phi}, \nu_{t+1}^{\psi}) + J_{\infty,t+1}^{\phi,\psi}(\mu_{t+1}^{\phi}, \nu_{t+1}^{\psi})\Big] \\
&\overset{\text{(ii)}}{\geq} r_t(\mu_t^{N_1}, \nu_t^{N_2}) + J_{\infty,t+1}^{\phi,\psi}(\mu_{t+1}^{\phi}, \nu_{t+1}^{\psi}) - \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big) - L_{J,t+1} \underbrace{\mathbb{E}_{\phi}\Big[\mathrm{d_{TV}}\big(\mathcal{M}_{t+1}^{N_1}, \mu_{t+1}^{\phi}\big)\Big]}_{=\mathcal{O}(\frac{1}{\sqrt{N_1}})} \\
&\qquad\qquad - \underbrace{L_{J,t+1}\mathbb{E}_{\psi}\Big[\mathrm{d_{TV}}\big(\mathcal{N}_{t+1}^{N_2}, \nu_{t+1}^{\psi}\big)\Big]}_{=\mathcal{O}(\frac{1}{\sqrt{N_2}})} \\
&\overset{\text{(iii)}}{\geq} r_t(\mu_t^{N_1}, \nu_t^{N_2}) + J_{\infty,t+1}^{\phi,\psi}(\mu_{t+1}^{\phi}, \nu_{t+1}^{\psi}) - \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big) \\
&\overset{\text{(iv)}}{=} J_{\infty,t}^{\phi,\psi}(\mu_t^{N_1}, \nu_t^{N_2}) - \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big).
\end{aligned}$$

For inequality (i), we used the inductive hypothesis; for inequality (ii), we utilized the Lipschitz continuity of the coordinator value function where $\mu_{t+1}^{\phi} = \mu_t^{N_1} F(\mu_t^{N_1}, \nu_t^{N_2}, \phi)$ and $\nu_{t+1}^{\psi} = \nu_t^{N_2} G(\mu_t^{N_1}, \nu_t^{N_2}, \psi)$; inequality (iii) follows from Guan et al. [2024b] and equality (iv) follows from the definition of $J_{\infty,t}^{\phi,\psi}(\mu_t^{N_1}, \nu_t^{N_2})$ which completes the induction. We have that, for all joint states $\mathbf{x}^{N_1} \in \mathbf{X}^{N_1}$ and $\mathbf{y}^{N_2} \in \mathbf{Y}^{N_2}$,

$$J^{N,\phi,\psi}(\mathbf{x}^{N_1}, \mathbf{y}^{N_2}) = J_0^{N,\phi,\psi}(\mathbf{x}^{N_1}, \mathbf{y}^{N_2}) \geq J_{\infty}^{\phi,\psi}(\mu^{N_1}, \nu^{N_2}) - \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big).$$

Similarly, using induction we can show that for (ϕ, ψ) we have the following inequality for all joint states $\mathbf{x}^{N_1} \in \mathcal{X}^{N_1}$ and $\mathbf{y}^{N_2} \in \mathcal{Y}^{N_2}$,

$$J^{N,\phi,\psi}(\mathbf{x}^{N_1}, \mathbf{y}^{N_2}) \leq J_{\infty}^{\phi,\psi}(\mu^{N_1}, \nu^{N_2}) + \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big), \tag{E.18}$$

where $\mu^{N_1} = \mathrm{Emp}_{\mu}(\mathbf{x}^{N_1})$, $\nu^{N_2} = \mathrm{Emp}_{\nu}(\mathbf{y}^{N_2})$ and the result follows. □

We begin by presenting the key steps of the proof of Theorem 3.1:

Step 1: We have the following definition of the lower game value for the finite-population ZS-MFTG:

$$\underline{J}^{N*}(\mathbf{x}^{N_1}, \mathbf{y}^{N_2}) = \max_{\phi^{N_1} \in \Phi^{N_1}} \min_{\psi^{N_2} \in \Psi^{N_2}} J^{N,\phi^{N_1},\psi^{N_2}}(\mathbf{x}^{N_1}, \mathbf{y}^{N_2}). \tag{E.19}$$

If we restrict ourselves to the set of identical team policies $\Phi \subseteq \Phi^{N_1}$ it follows immediately that

$$\begin{aligned}\underline{J}^{N*}(\mathbf{x}^{N_1}, \mathbf{y}^{N_2}) &\geq \max_{\phi^{N_1} \in \Phi} \min_{\psi^{N_2} \in \Psi^{N_2}} J^{N,\phi^{N_1},\psi^{N_2}}(\mathbf{x}^{N_1}, \mathbf{y}^{N_2}) \\ &\geq \min_{\psi^{N_2} \in \Psi^{N_2}} J^{N,\tilde{\phi}^*,\psi^{N_2}}(\mathbf{x}^{N_1}, \mathbf{y}^{N_2}), \end{aligned} \tag{E.20}$$

where $\tilde{\phi}^*$ follows the definition from the statement of the theorem.

Step 2: We next define, $\tilde{\psi}^*_\infty = \arg\min_{\psi \in \Psi} J^{\tilde{\phi}^*,\psi}_\infty(\mu_0, \nu_0)$ which determines the optimal Red team response in the *infinite-population* domain, when the Blue team plays $\tilde{\phi}^*$ derived from the finite population optimization over identical policies. For $\tilde{\phi}^*$ we have the following inequality for all joint states $\mathbf{x}^{N_1} \in \mathcal{X}^{N_1}$ and $\mathbf{y}^{N_2} \in \mathcal{Y}^{N_2}$,

$$\begin{aligned}\min_{\psi^{N_2} \in \Psi^{N_2}} J^{N,\tilde{\phi}^*,\psi^{N_2}}(\mathbf{x}^{N_1}, \mathbf{y}^{N_2}) &\geq J^{\tilde{\phi}^*,\tilde{\psi}^*_\infty}_\infty(\mu^{N_1}, \nu^{N_2}) - \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big) \\ &= \min_{\psi \in \Psi} J^{\tilde{\phi}^*,\psi}_\infty(\mu^{N_1}, \nu^{N_2}) - \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big) \end{aligned} \tag{E.21}$$

where $\mu^{N_1} = \mathrm{Emp}_\mu(\mathbf{x}^{N_1})$, $\nu^{N_2} = \mathrm{Emp}_\nu(\mathbf{y}^{N_2})$. The proof of inequality (E.21) statement is identical to Step 2 in Appendix G.

Step 3: Now, applying Lemma 3 to $f = J^{N,\phi,\psi}(\mathbf{x}^{N_1}, \mathbf{y}^{N_2})$ and $g = J^{\phi,\psi}_\infty(\mu^{N_1}, \nu^{N_2})$ where $\epsilon = \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big)$ (by Lemma 4) and $x^*_f = \tilde{\phi}^*$, it follows that

$$\begin{aligned}\min_{\psi \in \Psi} J^{\tilde{\phi}^*,\psi}_\infty(\mu^{N_1}, \nu^{N_2}) &\geq \max_{\phi \in \Phi} \min_{\psi \in \Psi} J^{\phi,\psi}_\infty(\mu^{N_1}, \nu^{N_2}) - \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big) \\ &\geq \underline{J}^{N*}(\mathbf{x}^{N_1}, \mathbf{y}^{N_2}) - \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big), \end{aligned} \tag{E.22}$$

where the second inequality follows from Lemma 6 in Guan et al. [2024b]. Combining (E.21) and (E.22), we obtain the desired result.

□

F Proof of Theorem 3.2

Theorem 3.2 (*Informal*)**.** *The approximate policy gradient of the infinite-population Blue (Red) team policy computed using identical local policies* $(\phi_{\theta_{\mathrm{Blue}}}, \psi_{\theta_{\mathrm{Red}}})$ *via MF-MAPPO from the* (N_1, N_2)*-agent simulator converges to the infinite-population policy gradient, i.e.,* $\|\nabla_{\theta_{\mathrm{Blue}}} J_\infty(\phi_{\theta_{\mathrm{Blue}}}, \psi) - \nabla_{\theta_{\mathrm{Blue}}} \hat{J}^{N_1,N_2}(\phi_{\theta_{\mathrm{Blue}}}, \psi)\|_2 \to 0$ *as* $(N_1, N_2) \to \infty$.

We first discuss the infinite-population solution in detail. To reduce the complexity of team policy optimization domains in (2), the authors of Guan et al. [2024a] examined 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* referred to as the mean-fields (μ_t and ν_t for the Blue and Red teams, respectively) and the infinite-population game can be transformed to an equivalent zero-sum game between two coordinators. Furthermore, there is a one-to-one correspondence between infinite-population *coordination* policies (α, β) and local *identical* team policies $(\phi, \psi) \in \Phi \times \Psi$. The performance of (ϕ, ψ) in the equivalent zero-sum coordinator game is measured by

$$J^{\alpha,\beta}_\infty(\mu_0, \nu_0) \equiv J^{\phi,\psi}_\infty(\mu_0, \nu_0) = \sum_{t=0}^{T} r_t(\mu_t, \nu_t), \tag{F.23}$$

where μ_t and ν_t follow a *deterministic* dynamics [Guan et al., 2024a] similar to the state distribution propagation of a controlled Markov chain. Guan et al. [2024a] establishes guarantees that identical team policies resulting from the solution of this equivalent zero-sum *coordinator* game are still ϵ-optimal for the original max-min optimization problem in (2) where $\epsilon = \mathcal{O}(1/\sqrt{\underline{N}})$ and $\underline{N} = \min\{N_1, N_2\}$.

We extend this formulation to potentially *stochastic* infinite-population coordinator policies and the local finite-population policies as follows: a stochastic Blue coordination policy $\alpha \in \mathcal{A}$ induces an identical team policy $\phi \in \Phi$ according to the rule

$$\phi_t(u_t|x_t, \mu_t^\rho, \nu_t^\rho) = \int_{\pi \in \Pi} \pi_t(u_t|x_t)\alpha_t(\pi_t|\mu_t^\rho, \nu_t^\rho)d\pi \quad \forall \mu_t \in \mathcal{P}(\mathcal{X}),\ \nu_t \in \mathcal{P}(\mathcal{Y}),\ x_t \in \mathcal{X} \text{ and } u_t \in \mathcal{U}, \tag{F.24}$$

where $\pi_t(u_t|x_t)$ corresponds to the identical local policies for all states $x \in \mathcal{X}$ prescribed by the coordinator, i.e., $\pi_t \sim \alpha(\mu_t^\rho, \nu_t^\rho)$. Thus, we can now formally state Theorem 3.2 using the notion of coordinator policies.

Theorem F.1. *The approximate policy gradient of the infinite-population Blue (Red) team coordinator policy α (β) computed using local policies from the finite-population ZS-MFTG via MF-MAPPO ($\hat{J}^{N_1}(\alpha_\theta)$) uniformly tends to the true policy gradient as the population size increases, i.e., $\|\nabla_\theta J_\infty(\alpha_\theta) - \nabla_\theta \hat{J}^{N_1}(\alpha_\theta)\|_2 \to 0$ as $(N_1, N_2) \to \infty$.*

Sketch of Proof. In MF-MAPPO, we directly train the finite-population local policies (ϕ_t, ψ_t). As a result, the gradient of the coordinator policy $\nabla_\theta J(\alpha_\theta)$ only has action samples $u_{i,t}^{N_1}, i \in [N_1]$ from the finite-population local policies $\phi_t(\cdot|x_{i,t}^{N_1}, \mu_t^{N_1}, \nu_t^{N_2})$ and not $\pi_t^{N_1}(\cdot|x_{i,t}^{N_1}) \sim \alpha(\mu_t^{N_1}, \nu_t^{N_2})$ itself. Thus, our first step is to construct an approximation of $\pi_t^{N_1}$ using the obtained finite-population action samples $u_{i,t}^{N_1}, i \in [N_1]$, resulting in an approximate finite-population policy gradient expression given by $\nabla_\theta \hat{J}^{N_1}(\alpha_\theta)$. Coupled with existing mean-field approximation results [Cui et al., 2023, Guan et al., 2024a, Shao et al., 2024b] for the infinite-population limit $N_1, N_2 \to \infty$ (LLN, etc.) we show convergence. □

Proof. It is important to note that in MF-MAPPO we directly work with the finite-population local policies (ϕ, ψ), unlike prior works that train on the infinite-population coordinator policy (see Section 1). Consequently, the gradient of the coordinator policy $\nabla_\theta J(\alpha_\theta)$ computed using this finite-population approximation (denoted $\nabla_\theta \hat{J}^{N_1}(\alpha_\theta)$) only has action samples $u_{i,t}^{N_1}, i \in [N_1]$ from the finite-population local policies $\phi_t(\cdot|x_{i,t}^{N_1}, \mu_t^{N_1}, \nu_t^{N_2})$ and not direct access to $\pi_t^{N_1} \sim \alpha(\mu_t^{N_1}, \nu_t^{N_2})$. Thus, our first step is to construct an approximation of the policy based on the obtained action samples. We define this approximate policy from the obtained samples as follows,

$$\hat{\pi}_t^{N_1}(u|x) = \begin{cases} \frac{\sum_{i=1}^{N_1} \mathbb{1}_x\left(X_{i,t}^{N_1}\right)\mathbb{1}_u\left(U_{i,t}^{N_1}\right)}{N_1 \mathcal{M}_t^{N_1}(x)} & \text{if } \ \mathcal{M}_t^{N_1}(x) > 0, \\ 1/|\mathcal{U}| & \text{if } \ \mathcal{M}_t^{N_1}(x) = 0. \end{cases} \tag{F.25}$$

We can similarly define $\hat{\sigma}_t^{N_2}$ for the Red team. Using this constructed empirical policy $\hat{\pi}_t^{N_1}$ from the sampled actions $\mathbf{u}_t^{N_1}$, the policy gradient $\nabla_\theta \hat{J}^{N_1}(\alpha_\theta)$ for the finite-population ZS-MFTG is,

$$\nabla_\theta \hat{J}^{N_1}(\alpha_\theta) = \sum_{t=0}^{T} \mathbb{E}_{\mathbf{u}^{N_1} \sim \phi, \mathbf{v}^{N_2} \sim \psi}\left[Q_t^{N,\alpha,\beta}(\mu_t^{N_1}, \nu_t^{N_2}, \hat{\pi}_t^{N_1}, \hat{\sigma}_t^{N_2}) \nabla_\theta \log \alpha_\theta(\hat{\pi}_t^{N_1}|\mu_t^{N_1}, \nu_t^{N_2})\right], \tag{F.26}$$

where,

$$\begin{aligned} &Q_t^{N,\alpha,\beta}(\mu_t^{N_1}, \nu_t^{N_2}, \hat{\pi}_t^{N_1}, \hat{\sigma}_t^{N_2}) \\ &= \mathbb{E}_{\mathbf{u}^{N_1} \sim \phi, \mathbf{v}^{N_2} \sim \psi}\left[\sum_{\tau=0}^{T} r(\mu_\tau^{N_1}, \nu_\tau^{N_2})\middle|\mu_{\tau=0}^{N_1} = \mu_t^{N_1}, \nu_{\tau=0}^{N_2} = \nu_t^{N_2}, \pi_{\tau=0}^{N_1} = \hat{\pi}_t^{N_1}, \sigma_{\tau=0}^{N_2} = \hat{\sigma}_t^{N_2}\right]. \end{aligned} \tag{F.27}$$

Recall that $\mathbf{u}^{N_1}$ and $\mathbf{v}^{N_2}$ enter (F.26) and (F.27) through the construction of $\hat{\pi}_t^{N_1}$ (Equation F.31) and $\hat{\sigma}_t^{N_2}$ respectively. Furthermore, we have the following expression for the gradient of the infinite-population coordinator game:

$$\nabla_\theta J(\alpha_\theta) = \sum_{t=0}^{T} \mathbb{E}_{\pi \sim \alpha, \sigma \sim \beta}\left[Q_t^{\alpha,\beta}(\mu_t, \nu_t, \pi_t, \sigma_t) \nabla_\theta \log \alpha_\theta(\pi_t|\mu_t, \nu_t)\right], \tag{F.28}$$

where,

$$Q_t^{\alpha,\beta}(\mu,\nu,\pi,\sigma) = \mathbb{E}_{\pi\sim\alpha,\sigma\sim\beta}\left[\sum_{\tau=0}^{T} r(\mu_\tau,\nu_\tau)\big|\mu_{\tau=0}=\mu,\nu_{\tau=0}=\nu,\pi_{\tau=0}=\pi,\sigma_{\tau=0}=\sigma\right]. \tag{F.29}$$

While (F.27) and (F.29) utilize the joint coordinator policies (α,β), the policy gradients (F.26) and (F.28) are taken for each team *individually*. Consequently, the policy gradient analysis is done on a per-team basis. We focus on the policy gradient with respect to the Blue team but the analysis of the Red team's policy gradient is symmetric and can be proved by an identical approach.

The following lemma relates the expectation in (F.26) to (α,β) using (F.24). It shows that a stochastic coordinator policy, under shared randomness, caninduce a closed-loop policy ϕ that achieves the same expected performance.

Lemma 5. *Assume common randomness among all agents in a team when sampling the local policy π. Then, given a function $g:\mathcal{U}\to\mathbb{R}$ of the joint actions $\mathbf{u}_t^{N_1}$ such that $u_{i,t}^{N_1}\sim\phi(\cdot|x_{i,t}^{N_1},\mu_t^{N_1},\nu_t^{N_2})$, or equivalently, $\mathbf{u}_t^{N_1}\sim\phi(\cdot|\mathbf{x}_t^{N_1},\mu_t^{N_1},\nu_t^{N_2})$, it holds that*

$$\mathbb{E}_{\mathbf{u}_t^{N_1}\sim\phi}\left[g(\mathbf{u}_t^{N_1})\right] = \mathbb{E}_{\pi\sim\alpha}\left[\mathbb{E}_{\mathbf{u}_t^{N_1}\sim\pi}\left[g(\mathbf{u}_t^{N_1})\right]\right].$$

Proof. By the definition of expectation,

$$\mathbb{E}_{\mathbf{u}_t^{N_1}\sim\phi}\left[g(\mathbf{u}_t^{N_1})\right] = \sum_{\mathbf{u}_t^{N_1}} g(\mathbf{u}_t^{N_1})\phi(\mathbf{u}_t^{N_1}|\mathbf{x}_t^{N_1},\mu_t^{N_1},\nu_t^{N_2}),$$

where the sum is taken over all the possible joint actions that can be sampled from ϕ. By using the definition of ϕ from (F.24),

$$\begin{aligned}
\sum_{\mathbf{u}_t^{N_1}} g(\mathbf{u}_t^{N_1})\phi(\mathbf{u}_t^{N_1}|\mathbf{x}_t^{N_1},\mu_t^{N_1},\nu_t^{N_2}) &= \sum_{\mathbf{u}_t^{N_1}} g(\mathbf{u}_t^{N_1}) \underbrace{\int_{\pi\in\Pi}\pi_t(\mathbf{u}_t^{N_1}|\mathbf{x}_t^{N_1})\alpha_t(\mu_t^{N_1},\nu_t^{N_2})d\pi}_{=\mathbb{E}_{\pi\sim\alpha}\left[\pi_t(\mathbf{u}_t^{N_1}|\mathbf{x}_t^{N_1})\right]} \\
&= \sum_{\mathbf{u}_t^{N_1}} g(\mathbf{u}_t^{N_1})\mathbb{E}_{\pi\sim\alpha}\left[\pi_t(\mathbf{u}_t^{N_1}|\mathbf{x}_t^{N_1})\right] \\
&= \sum_{\mathbf{u}_t^{N_1}} \mathbb{E}_{\pi\sim\alpha}\left[g(\mathbf{u}_t^{N_1})\pi_t(\mathbf{u}_t^{N_1}|\mathbf{x}_t^{N_1})\right] \\
&= \mathbb{E}_{\pi\sim\alpha}\left[\sum_{\mathbf{u}_t^{N_1}} g(\mathbf{u}_t^{N_1})\pi_t(\mathbf{u}_t^{N_1}|\mathbf{x}_t^{N_1})\right],
\end{aligned}$$

where the last two expressions follow from linearity (and finiteness) of expectation. It follows then,

$$\mathbb{E}_{\mathbf{u}_t^{N_1}\sim\phi}\left[g(\mathbf{u}_t^{N_1})\right] = \mathbb{E}_{\pi\sim\alpha}\left[\mathbb{E}_{\mathbf{u}_t^{N_1}\sim\pi}\left[g(\mathbf{u}_t^{N_1})\right]\right].$$

□

Remark 3. This shared-randomness assumption is only required for this theoretical equivalence and is not required by MF-MAPPO itself. In practice, MF-MAPPO directly learns the local policy ϕ from finite-population interactions and therefore does not require a common random seed to be shared across agents.

We further assume continuity with respect to the local policies.

Assumption 6. The Q-functions $Q^{\alpha,\beta}$ and $Q^{N,\alpha,\beta}$ from (F.29) and (F.27) are continuous in inputs (π,σ) and $(\hat{\pi}^{N_1},\hat{\sigma}^{N_2})$.

It is important to note that the action-space Π and Σ are continuous. Thus, although Assumption 6 seems restrictive, the use of function approximators to train $Q^{\alpha,\beta}$ and $Q^{N,\alpha,\beta}$ like neural networks (endowed with continuous activation functions), ensures that the resultant functions are continuous in their inputs [Lillicrap et al., 2019].

The following assumption, as done in many works [Cui et al., 2024, 2023], states that the gradient of the coordinator policy is continuous and uniformly bounded.

Assumption 7. The log-gradient $\nabla \log \alpha_\theta(\pi_t|\mu_t,\nu_t)$ is L_∇-Lipschitz continuous and uniformly bounded.

Furthermore, we have the following result from Guan et al. [2024a]:

Corollary 1. *Let $\mathbf{X}_t^{N_1}$, $\mathbf{Y}_t^{N_2}$, $\mathcal{M}_t^{N_1}$, and $\mathcal{N}_t^{N_2}$ be the joint states and the corresponding EDs of a finite-population game. Denote the next Blue ED induced by an identical Blue team policy $\phi_t \in \Phi_t$ as $\mathcal{M}_{t+1}^{N_1}$. Then, the following holds:*

$$\mathbb{E}_{\phi_t}\left[\mathrm{d_{TV}}\big(\mathcal{M}_{t+1}^{N_1},\mu_{t+1}\big)\big|\,\mathbf{X}_t^{N_1},\mathbf{Y}_t^{N_2}\right] \leq \frac{|\mathcal{X}|}{2}\sqrt{\frac{1}{N_1}},$$

where $\mu_{t+1} = \mathcal{M}_t^{N_1} F_t(\mathcal{M}_t^{N_1},\mathcal{N}_t^{N_2},\phi_t)$.

As mentioned in the main text, we assume that $r_t \in [-R_{\max}, R_{\max}]$. Consequently, the Q-functions $Q_t^{\alpha,\beta}$ and $Q_t^{N,\alpha,\beta}$ are uniformly bounded over the finite-horizon. We are interested in deriving an explicit result for $\|\nabla_\theta J(\alpha_\theta) - \nabla_\theta \hat{J}^{N_1}(\alpha_\theta)\|_2$ as the population sizes $(N_1,N_2)\to\infty$. In particular, using shorthand notation $\mathbb{E}_{\alpha,\beta}[\cdot] \triangleq \mathbb{E}_{\pi\sim\alpha,\sigma\sim\beta}[\cdot]$ and $\mathbb{E}_{\phi,\psi}[\cdot] \triangleq \mathbb{E}_{\mathbf{u}^{N_1}\sim\phi,\mathbf{v}^{N_2}\sim\psi}[\cdot]$)

$$\begin{aligned}\|\nabla_\theta J(\alpha_\theta) - \nabla_\theta \hat{J}^{N_1}(\alpha_\theta)\|_2 = \Big\| \sum_{t=0}^{T} &\mathbb{E}_{\alpha,\beta}\left[Q_t^{\alpha,\beta}(\mu_t,\nu_t,\pi_t,\sigma_t)\nabla_\theta \log \alpha_\theta(\pi_t|\mu_t,\nu_t)\right] \\ &- \mathbb{E}_{\phi,\psi}\left[Q_t^{N,\alpha,\beta}(\mu_t^{N_1},\nu_t^{N_2},\hat{\pi}_t^{N_1},\hat{\sigma}_t^{N_2})\nabla_\theta \log \alpha_\theta(\hat{\pi}_t^{N_1}|\mu_t^{N_1},\nu_t^{N_2})\right]\Big\|_2.\end{aligned}$$

For the rest of the proof, we implicitly assume $\|\cdot\| \equiv \|\cdot\|_2$ to denote the 2-norm. By Lemma 5, under the common randomness assumption, we can rewrite the above as

$$\begin{aligned}\|\nabla_\theta J(\alpha_\theta) - \nabla_\theta \hat{J}^{N_1}(\alpha_\theta)\| = \Big\| \sum_{t=0}^{T} &\mathbb{E}_{\alpha,\beta}\Big[Q_t^{\alpha,\beta}(\mu_t,\nu_t,\pi_t,\sigma_t)\nabla_\theta \log \alpha_\theta(\pi_t|\mu_t,\nu_t) \\ &- \mathbb{E}_{\mathbf{u}\sim\pi,\mathbf{v}\sim\sigma}\left[Q_t^{N,\alpha,\beta}(\mu_t^{N_1},\nu_t^{N_2},\hat{\pi}_t^{N_1},\hat{\sigma}_t^{N_2})\nabla_\theta \log \alpha_\theta(\hat{\pi}_t^{N_1}|\mu_t^{N_1},\nu_t^{N_2})\right]\Big]\Big\|,\end{aligned}$$

where $\hat{\pi}_t^{N_1},\hat{\sigma}_t^{N_2}$ depend on the random variables $\mathbf{u}^{N_1}\sim\pi, \mathbf{v}^{N_2}\sim\sigma$ respectively. We add and subtract the term $\mathbb{E}_{\mathbf{u}^{N_1}\sim\pi,\mathbf{v}^{N_2}\sim\sigma}\left[Q_t^{\alpha,\beta}(\mu_t^{N_1},\nu_t^{N_2},\hat{\pi}_t^{N_1},\hat{\sigma}_t^{N_2})\nabla_\theta \log \alpha_\theta(\hat{\pi}_t^{N_1}|\mu_t^{N_1},\nu_t^{N_2})\right]$ and split the gradient terms and analyze the two terms individually.

$$\begin{aligned}&\|\nabla_\theta J(\alpha_\theta) - \nabla_\theta \hat{J}^{N_1}(\alpha_\theta)\| \\ &\leq \Big\| \sum_{t=0}^{T} \mathbb{E}_{\alpha,\beta}\Big[Q_t^{\alpha,\beta}(\mu_t,\nu_t,\pi_t,\sigma_t)\nabla_\theta \log \alpha_\theta(\pi_t|\mu_t,\nu_t) \\ &- \mathbb{E}_{\mathbf{u}^{N_1}\sim\pi,\mathbf{v}^{N_2}\sim\sigma}\left[Q_t^{\alpha,\beta}(\mu_t^{N_1},\nu_t^{N_2},\hat{\pi}_t^{N_1},\hat{\sigma}_t^{N_2})\nabla_\theta \log \alpha_\theta(\hat{\pi}_t^{N_1}|\mu_t^{N_1},\nu_t^{N_2})\right]\Big]\Big\| \\ &+ \Big\| \sum_{t=0}^{T} \mathbb{E}_{\alpha,\beta,\mathbf{u}^{N_1}\sim\pi,\mathbf{v}^{N_2}\sim\sigma}\left[\Big(Q_t^{\alpha,\beta}(\mu_t^{N_1},\nu_t^{N_2},\hat{\pi}_t^{N_1},\hat{\sigma}_t^{N_2}) - Q_t^{N,\alpha,\beta}(\mu_t^{N_1},\nu_t^{N_2},\hat{\pi}_t^{N_1},\hat{\sigma}_t^{N_2})\Big)\nabla_\theta \log \alpha_\theta(\hat{\pi}_t^{N_1}|\mu_t^{N_1},\nu_t^{N_2})\right]\Big]\Big\|.\end{aligned} \tag{F.30}$$

Since the first term in the above expression (F.30) is a constant with respect to the random variables $\mathbf{u}^{N_1}, \mathbf{v}^{N_2}$ and the expectation of a constant is the value itself, we can equivalently write it as

$$\Big\| \sum_{t=0}^{T} \mathbb{E}_{\alpha,\beta,\mathbf{u}^{N_1}\sim\pi,\mathbf{v}^{N_2}\sim\sigma} \Big[Q_t^{\alpha,\beta}(\mu_t,\nu_t,\pi_t,\sigma_t) \nabla_\theta \log \alpha_\theta(\pi_t|\mu_t,\nu_t) \\ - Q_t^{\alpha,\beta}(\mu_t^{N_1},\nu_t^{N_2},\hat{\pi}_t^{N_1},\hat{\sigma}_t^{N_2}) \nabla_\theta \log \alpha_\theta(\hat{\pi}_t^{N_1}|\mu_t^{N_1},\nu_t^{N_2}) \Big] \Big\|, \tag{F.31}$$

Define $h_t(\mu,\nu,\pi,\sigma) \triangleq Q_t^{\alpha,\beta}(\mu,\nu,\pi,\sigma)\nabla_\theta \log \alpha_\theta(\pi|\mu,\nu)$. Assumptions 6 and 7 ensure continuity of the function $h_t(\mu,\nu,\pi,\sigma)$ with respect to all its inputs. Using this definition of h, rewrite (F.31) as

$$\Big\| \sum_{t=0}^{T} \mathbb{E}_{\alpha,\beta,\mathbf{u}\sim\pi,\mathbf{v}\sim\sigma} \Big[h_t(\mu_t,\nu_t,\pi_t,\sigma_t) - h_t(\mu_t^{N_1},\nu_t^{N_2},\hat{\pi}_t^{N_1},\hat{\sigma}_t^{N_2}) \Big] \Big\|. \tag{F.32}$$

Using the definition of the expectation operators,

$$\Big\| \sum_{t=0}^{T} \gamma^t \mathbb{E}_{\alpha,\beta,\mathbf{u}^{N_1}\sim\pi,\mathbf{v}^{N_2}\sim\sigma} \Big[h_t(\mu_t,\nu_t,\pi_t,\sigma_t) - h_t(\mu_t^{N_1},\nu_t^{N_2},\hat{\pi}_t^{N_1},\hat{\sigma}_t^{N_2}) \Big] \Big\| \\ = \Big\| \sum_{t=0}^{T} \gamma^t \int_\pi \int_\sigma h_t(\mu_t,\nu_t,\pi_t,\sigma_t)\beta_t(\mu_t,\nu_t)\alpha_t(\mu_t,\nu_t) d\sigma d\pi \\ - \int_\pi \int_\sigma \Big[\sum_{\mathbf{u}^{N_1}} \sum_{\mathbf{v}^{N_2}} h_t(\mu_t^{N_1},\nu_t^{N_2},\hat{\pi}_t^{N_1},\hat{\sigma}_t^{N_2}) \pi(\mathbf{u}^{N_1}|\mathbf{x}_t^{N_1}) \sigma(\mathbf{v}^{N_2}|\mathbf{y}_t^{N_1}) \Big] \beta_t(\mu_t,\nu_t)\alpha_t(\mu_t,\nu_t) d\sigma d\pi \Big\|$$

We already have, by weak LLN that $\mathrm{d_{TV}}\big(\mu_t^{N_1},\mu_t\big) \leq \mathcal{O}(1/\sqrt{N_1})$ and $\mathrm{d_{TV}}\big(\nu_t^{N_2},\nu_t\big) \leq \mathcal{O}(1/\sqrt{N_2})$, with the error going to zero as the population size becomes large, i.e., $N_1, N_2 \to \infty$. A similar argument using weak LLN can be made for $(\pi_t^{N_1},\sigma_t^{N_2}) \to (\pi_t,\sigma_t)$ as $N_1, N_2 \to \infty$. Furthermore, as the population size becomes large, the empirical policies constructed $(\hat{\pi}_t^{N_1},\hat{\sigma}_t^{N_2}) \to (\pi_t^{N_1},\sigma_t^{N_2})$ because the number of samples of $\mathbf{u}_t^{N_1}$ and $\mathbf{v}_t^{N_2}$ scale with (N_1,N_2).[5] Thus, it follows that $(\hat{\pi}_t^{N_1},\hat{\sigma}_t^{N_2}) \to (\pi_t,\sigma_t)$ as the team sizes become large and, coupled with the continuity of h with respect to $\hat{\pi}_t^{N_1},\hat{\sigma}_t^{N_2}$, the following simplification holds:

$$\sum_{\mathbf{u}} \sum_{\mathbf{v}} h_t(\mu_t^{N_1},\nu_t^{N_2},\hat{\pi}_t^{N_1},\hat{\sigma}_t^{N_2}) \pi(\mathbf{u}|\mathbf{x}_t^{N_1}) \sigma(\mathbf{v}|\mathbf{y}_t^{N_1}) \to \sum_{\mathbf{u}} \sum_{\mathbf{v}} h_t(\mu_t,\nu_t,\pi_t,\sigma_t) \pi(\mathbf{u}|\mathbf{x}_t^{N_1}) \sigma(\mathbf{v}|\mathbf{y}_t^{N_1}) \\ = h_t(\mu_t,\nu_t,\pi_t,\sigma_t) \sum_{\mathbf{u}} \pi(\mathbf{u}^{N_1}|\mathbf{x}_t^{N_1}) \sum_{\mathbf{v}} \sigma(\mathbf{v}^{N_2}|\mathbf{y}_t^{N_1}) \\ = h_t(\mu_t,\nu_t,\pi_t,\sigma_t).$$

For the second term in (F.30), we first focus on the term

$$\Big(Q_t^{\alpha,\beta}(\mu_t^{N_1},\nu_t^{N_2},\hat{\pi}_t^{N_1},\hat{\sigma}_t^{N_2}) - Q_t^{N,\alpha,\beta}(\mu_t^{N_1},\nu_t^{N_2},\hat{\pi}_t^{N_1},\hat{\sigma}_t^{N_2}) \Big).$$

The expectation in (F.27) is with respect to $\mathbf{u}^{N_1} \sim \phi, \mathbf{v}^{N_2} \sim \psi$. It follows from Lemma 5, we can rewrite (F.27)as

$$Q_t^{N,\alpha,\beta}(\mu_t^{N_1},\nu_t^{N_2},\hat{\pi}_t^{N_1},\hat{\sigma}_t^{N_2}) \\ = \mathbb{E}_{\pi\sim\alpha,\sigma\sim\beta,\mathbf{u}^{N_1}\sim\pi,\mathbf{v}^{N_2}\sim\sigma} \left[\sum_{\tau=0}^{T} r(\mu_\tau^{N_1},\nu_\tau^{N_2}) \big| \mu_0^{N_1} = \mu_t^{N_1}, \nu_0^{N_2} = \nu_t^{N_2}, \pi_0^{N_1} = \hat{\pi}_t^{N_1}, \sigma_0^{N_2} = \hat{\sigma}_t^{N_2} \right]$$

On the other hand, the expectation in (F.29) is with respect to α,β. Using a similar argument as (F.31), we can write (F.29) as

$$Q_t^{\alpha,\beta}(\mu_t^{N_1},\nu_t^{N_2},\hat{\pi}_t^{N_1},\hat{\sigma}_t^{N_2}) \\ = \mathbb{E}_{\pi\sim\alpha,\sigma\sim\beta,\mathbf{u}^{N_1}\sim\pi,\mathbf{v}^{N_2}\sim\sigma} \left[\sum_{\tau=0}^{T} r(\mu_\tau,\nu_\tau) \big| \mu_0 = \mu_t^{N_1}, \nu_0 = \nu_t^{N_2}, \pi_0 = \hat{\pi}_t^{N_1}, \sigma_0 = \hat{\sigma}_t^{N_2} \right].$$

[5] Note that as $N_1 \to \infty$, $\hat{\pi}_t^{N_1}(\cdot|x) \neq \pi_t^{N_1}(\cdot|x)$ for states x such that $\mu(x) = 0$ by the very definition of the approximation (F.25). However, it is easy to check that the mean-field trajectories remain the same and the unoccupied states have no role to play in the evolution.

We apply the following lemma along with Assumption 7 to establish convergence of the second term in (F.30).

Lemma 6. *Under Assumption 1 and Lipschitz policies 7, $Q^{N,\alpha,\beta} \to Q^{\alpha,\beta}$ as $N_1, N_2 \to \infty$.*

Proof. Assumption 1 allows the $Q_t^{N,\alpha,\beta}, Q_t^{\alpha,\beta}$ to have compact support. Thus, it suffices to prove pointwise convergence [Cui et al., 2023] for each τ, i.e.,

$$\mathbb{E}\left[r(\mu_\tau^{N_1}, \nu_\tau^{N_2}) \middle| \mu_0^{N_1} = \mu_t^{N_1}, \nu_0^{N_2} = \nu_t^{N_2}, \pi_0^{N_1} = \hat{\pi}_t^{N_1}, \sigma_0^{N_2} = \hat{\sigma}_t^{N_2}\right] \to$$
$$\mathbb{E}\left[r(\mu_\tau, \nu_\tau) \middle| \mu_0 = \mu_t^{N_1}, \nu_0 = \nu_t^{N_2}, \pi_0 = \hat{\pi}_t^{N_1}, \sigma_0 = \hat{\sigma}_t^{N_2}\right] \tag{F.33}$$

as $N_1, N_2 \to \infty$, where the expectation is taken with respect to $\alpha, \beta, \mathbf{u} \sim \pi, \mathbf{v} \sim \sigma$ (dropped for readability). We follow an induction approach similar to the one presented in Cui et al. [2023]. It trivially holds at $\tau = 0$. Let us assume that (F.33) holds at time step τ. For subsequent $\tau + 1$,

$$\begin{aligned}
&\Big\|\mathbb{E}\left[r(\mu_{\tau+1}, \nu_{\tau+1}) \middle| \mu_0 = \mu_t^{N_1}, \nu_0 = \nu_t^{N_2}, \pi_0 = \hat{\pi}_t^{N_1}, \sigma_0 = \hat{\sigma}_t^{N_2}\right] \\
&- \mathbb{E}\left[r(\mu_{\tau+1}^{N_1}, \nu_{\tau+1}^{N_2}) \middle| \mu_0^{N_1} = \mu_t^{N_1}, \nu_0^{N_2} = \nu_t^{N_2}, \pi_0^{N_1} = \hat{\pi}_t^{N_1}, \sigma_0^{N_2} = \hat{\sigma}_t^{N_2}\right] \Big\| \\
&\leq \Big\|\mathbb{E}\left[r\Big(\mu_\tau F(\mu_\tau, \nu_\tau, \pi_\tau), \nu_\tau G(\mu_\tau, \nu_\tau, \sigma_\tau)\Big) \middle| \mu_0 = \mu_t^{N_1}, \nu_0 = \nu_t^{N_2}, \pi_0 = \hat{\pi}_t^{N_1}, \sigma_0 = \hat{\sigma}_t^{N_2}\right] \\
&- \mathbb{E}\left[r\Big(\mu_\tau^{N_1} F(\mu_\tau^{N_1}, \nu_\tau^{N_2}, \phi_\tau), \nu_\tau^{N_2} G(\mu_\tau^{N_1}, \nu_\tau^{N_2}, \psi_\tau)\Big) \middle| \mu_0^{N_1} = \mu_t^{N_1}, \nu_0^{N_2} = \nu_t^{N_2}, \pi_0^{N_1} = \hat{\pi}_t^{N_1}, \sigma_0^{N_2} = \hat{\sigma}_t^{N_2}\right] \Big\| \\
&+ \Big\|\mathbb{E}\left[r\Big(\mu_\tau^{N_1} F(\mu_\tau^{N_1}, \nu_\tau^{N_2}, \phi_\tau), \nu_\tau^{N_2} G(\mu_\tau^{N_1}, \nu_\tau^{N_2}, \psi_\tau)\Big) \middle| \mu_0^{N_1} = \mu_t^{N_1}, \nu_0^{N_2} = \nu_t^{N_2}, \pi_0^{N_1} = \hat{\pi}_t^{N_1}, \sigma_0^{N_2} = \hat{\sigma}_t^{N_2}\right] \\
&- \mathbb{E}\left[r(\mu_{\tau+1}^{N_1}, \nu_{\tau+1}^{N_2}) \middle| \mu_0^{N_1} = \mu_t^{N_1}, \nu_0^{N_2} = \nu_t^{N_2}, \pi_0^{N_1} = \hat{\pi}_t^{N_1}, \sigma_0^{N_2} = \hat{\sigma}_t^{N_2}\right] \Big\|
\end{aligned}$$

The convergence of the first term follows from the weak LLN [Cui et al., 2023] and a similar argument as used for (F.32). By using the Lipshcitzness of the reward function (Assumption 1) and Corollary 1, the second term in the expression is bounded by $\mathcal{O}\Big(\frac{1}{\sqrt{\underline{N}}}\Big)$, where $\underline{N} = \min(N_1, N_2)$ and goes to zero as $(N_1, N_2) \to \infty$. □

It follows from the convergence of the two individual terms in F.30 that $\|\nabla_\theta J(\alpha_\theta) - \nabla_\theta \hat{J}^{N_1}(\alpha_\theta)\| \to 0$ as $(N_1, N_2) \to \infty$ □

G Proof of Theorem 3.3

Theorem 3.3. *Let $\mathcal{G}_N$ denote the finite-population game in which agents use identical team policies (ϕ_t^*, ψ_t^*) with an (N_1, N_2)-agent simulator. Consider another game $\mathcal{G}_{\tilde{N}}$ with the same setup but population sizes $(\tilde{N}_1, \tilde{N}_2)$ satisfying $\tilde{N}_1/\tilde{N}_2 = N_1/N_2$ and $\tilde{N}_1 \geq N_1$. If (ϕ_t^*, ψ_t^*) is ϵ-optimal in $\mathcal{G}_N$, then it remains ϵ-optimal in $\mathcal{G}_{\tilde{N}}$.*

Sketch of Proof. The key idea of the proof is to exploit the fact that the equivalent infinite-population games for $\mathcal{G}_1$ and $\mathcal{G}_2$ are the same as $N_1/N_2 = \tilde{N}_1/\tilde{N}_2$ [Guan et al., 2024a]. We first compute the performance of ϕ^*, the optimal identical team policy derived for $\mathcal{G}_1$ using MF-MAPPO (finite-population training), when applied to the *infinite-population* coordinator setting. We then compare the performance of the aforementioned ϕ^* for $\mathcal{G}_2$ with the *infinite-population* coordinator. Then, equating the common equivalent coordinators for $\mathcal{G}_1$ and $\mathcal{G}_2$, the result follows. □

Proof. Consider game $\mathcal{G}_1$. Let us restrict ourselves to the set of identical team policies $\Phi \subseteq \Phi^{N_1}$ and suppose that ϕ^* is the optimal identical Blue team policy obtained from the (N_1, N_2) finite population game via MF-MAPPO. From Theorem 3.1, we already know that

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

Denote $J^{N,\phi^*,\psi^{N_2*}}(\mathbf{x}^{N_1},\mathbf{y}^{N_2}) \triangleq \min_{\psi^{N_2}\in\Psi^{N_2}} J^{N,\phi^*,\psi^{N_2}}(\mathbf{x}^{N_1},\mathbf{y}^{N_2})$, where ψ^{N_2*} is the potentially non-identical policy that solves the optimization problem.

Now, for $\mathcal{G}_2$ we want to show that

$$\min_{\psi^{\bar{N}_2}\in\Psi^{\bar{N}_2}} J^{\bar{N},\phi^*,\psi^{\bar{N}_2}}(\mathbf{x}^{\bar{N}_1},\mathbf{y}^{\bar{N}_2}) \geq \underline{J}^{\bar{N}*}(\mathbf{x}^{\bar{N}_1},\mathbf{y}^{\bar{N}_2}) - \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big), \tag{G.34}$$

where ϕ^* corresponds to the identical policy derived from $\mathcal{G}_1$.

We know that the $\frac{N_1}{N_2} = \frac{\bar{N}_1}{\bar{N}_2}$. Thus, the infinite-population coordinator games remain identical for both $\mathcal{G}_1$ and $\mathcal{G}_2$. We prove the theorem in the following two steps.

Step 1. We first show that for ϕ^* from $\mathcal{G}_1$ we have the following inequality for all joint states $\mathbf{x}^{N_1}\in\mathcal{X}^{N_1}$ and $\mathbf{y}^{N_2}\in\mathcal{Y}^{N_2}$,

$$J_\infty^{\phi^*,\tilde{\psi}^*_\infty}(\mu^{N_1},\nu^{N_2}) \geq J^{N,\phi^*,\psi^{N_2*}}(\mathbf{x}^{N_1},\mathbf{y}^{N_2}) - \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big), \tag{G.35}$$

where $\mu^{N_1} = \mathrm{Emp}_\mu(\mathbf{x}^{N_1})$ and $\nu^{N_2} = \mathrm{Emp}_\nu(\mathbf{y}^{N_2})$ and $\tilde{\psi}^*_\infty\in\Psi$ is given by,

$$J_\infty^{\phi^*,\tilde{\psi}^*_\infty}(\mu_0,\nu_0) = \min_{\psi\in\Psi} J_\infty^{\phi^*,\psi}(\mu_0,\nu_0). \tag{G.36}$$

The above equation (G.36) determines the optimal Red team response in the *infinite-population* domain, when the Blue team plays ϕ^* derived from $\mathcal{G}_1$. We prove this through an inductive argument.

Base case: At the terminal timestep T, the two value functions are the same. Thus, we have, for all joint states $\mathbf{x}_T^{N_1}\in\mathbf{X}^{N_1}$ and $\mathbf{y}_T^{N_2}\in\mathbf{Y}^{N_2}$, that

$$J_{\infty,T}^{\phi^*,\tilde{\psi}^*_\infty}(\mu_T^{N_1},\nu_T^{N_1}) = J_T^{N,\phi^*,\psi^{N_2*}}(\mathbf{x}_T^{N_1},\mathbf{y}_T^{N_2}) = r_t(\mu_T^{N_1},\nu_T^{N_1}).$$

Inductive hypothesis: Assume that, at time step $t+1$, the following holds for all $\mathbf{x}_{t+1}^{N_1}\in\mathbf{X}^{N_1}$ and $\mathbf{y}_{t+1}^{N_2}\in\mathbf{Y}^{N_2}$,

$$J_{\infty,t+1}^{\phi^*,\tilde{\psi}^*_\infty}(\mu_{t+1}^{N_1},\nu_{t+1}^{N_1}) \geq J_{t+1}^{N,\phi^*,\psi^{N_2*}}(\mathbf{x}_{t+1}^{N_1},\mathbf{y}_{t+1}^{N_2}) - \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big). \tag{G.37}$$

Induction: Consider arbitrary $\mathbf{x}_t^{N_1}\in\mathbf{X}^{N_1}$ and $\mathbf{y}_t^{N_2}\in\mathbf{Y}^{N_2}$.

For simplicity, we do not emphasize the correspondence between the joint states and the EDs for the rest of the proof, as it is clear from the context. For the identical team policies $(\phi_t^*,\psi_t)\in\Phi\times\Psi$, at time step t, denote

$$\mu_{t+1}^{\phi^*} \triangleq \mu_t^{N_1} F(\mu_t^{N_1},\nu_t^{N_2},\phi_t^*), \qquad \nu_{t+1}^{\psi} \triangleq \nu_t^{N_2} G(\mu_t^{N_1},\nu_t^{N_2},\psi_t). \tag{G.38}$$

For notational simplicity, we drop the conditions $\mathbf{X}_t^{N_1} = \mathbf{x}_t^{N_1}$ and $\mathbf{Y}_t^{N_2} = \mathbf{y}_t^{N_2}$ in the following derivations. Then, we have

$$
\begin{aligned}
&J_t^{N,\phi^*,\psi^{N_2*}}(\mathbf{x}_t^{N_1}, \mathbf{y}_t^{N_2}) \\
&= \min_{\psi_t^{N_2} \in \Psi_t^{N_2}} r_t(\mu_t^{N_1}, \nu_t^{N_2}) + \mathbb{E}_{\phi_t^*, \psi_t^{N_2}}\Big[J_{t+1}^{N,\phi^*,\psi^{N_2*}}(\mathbf{X}_{t+1}^{N_1}, \mathbf{Y}_{t+1}^{N_2})\Big] \\
&\overset{\text{(i)}}{\leq} r_t(\mu_t^{N_1}, \nu_t^{N_2}) + \min_{\psi_t^{N_2} \in \Psi_t^{N_2}} \mathbb{E}_{\phi_t^*, \psi_t^{N_2}}\Big[J_{\infty,t+1}^{\phi^*,\tilde{\psi}_\infty^*}(\mathcal{M}_{t+1}^{N_1}, \mathcal{N}_{t+1}^{N_2})\Big] + \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big) \\
&\overset{\text{(ii)}}{\leq} r_t(\mu_t^{N_1}, \nu_t^{N_2}) + \min_{\psi_t \in \Psi_t} \mathbb{E}_{\phi_t^*, \psi_t}\Big[J_{\infty,t+1}^{\phi^*,\tilde{\psi}_\infty^*}(\mathcal{M}_{t+1}^{N_1}, \mathcal{N}_{t+1}^{N_2})\Big] + \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big) \\
&\overset{\text{(iii)}}{=} r_t(\mu_t^{N_1}, \nu_t^{N_2}) + \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big) + \min_{\psi_t \in \Psi_t} \mathbb{E}_{\phi_t^*, \psi_t}\Big[J_{\infty,t+1}^{\phi^*,\tilde{\psi}_\infty^*}(\mathcal{M}_{t+1}^{N_1}, \mathcal{N}_{t+1}^{N_2}) - J_{\infty,t+1}^{\phi^*,\tilde{\psi}_\infty^*}(\mu_{t+1}^{\phi^*}, \nu_{t+1}^{\psi_t}) \\
&\qquad\qquad + J_{\infty,t+1}^{\phi^*,\tilde{\psi}_\infty^*}(\mu_{t+1}^{\phi^*}, \nu_{t+1}^{\psi_t})\Big] \\
&\overset{\text{(iv)}}{\leq} r_t(\mu_t^{N_1}, \nu_t^{N_2}) + \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big) + \min_{\psi_t \in \Psi_t} \mathbb{E}_{\phi_t^*, \psi_t}\Big[L_{J,t+1}\mathrm{d_{TV}}\big(\mathcal{M}_{t+1}^{N_1}, \mu_{t+1}^{\phi^*}\big) \\
&\qquad\qquad + L_{J,t+1}\mathrm{d_{TV}}\big(\mathcal{N}_{t+1}^{N_2}, \nu_{t+1}^{\psi_t}\big) + J_{\infty,t+1}^{\phi^*,\tilde{\psi}_\infty^*}(\mu_{t+1}^{\phi^*}, \nu_{t+1}^{\psi_t})\Big] \\
&\overset{\text{(v)}}{=} r_t(\mu_t^{N_1}, \nu_t^{N_2}) + \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big) + \min_{\psi_t \in \Psi_t} J_{\infty,t+1}^{\phi^*,\tilde{\psi}_\infty^*}(\mu_{t+1}^{\phi^*}, \nu_{t+1}^{\psi_t}) + L_{J,t+1}\mathbb{E}_{\psi_t}\Big[\mathrm{d_{TV}}\big(\mathcal{N}_{t+1}^{N_2}, \nu_{t+1}^{\psi_t}\big)\Big] \\
&\qquad\qquad + L_{J,t+1}\mathbb{E}_{\phi_t^*}\Big[\mathrm{d_{TV}}\big(\mathcal{M}_{t+1}^{N_1}, \mu_{t+1}^{\phi^*}\big)\Big] \\
&\overset{\text{(vi)}}{\leq} r_t(\mu_t^{N_1}, \nu_t^{N_2}) + \min_{\psi_t \in \Psi_t} J_{\infty,t+1}^{\phi^*,\tilde{\psi}_\infty^*}(\mu_{t+1}^{\phi^*}, \nu_{t+1}^{\psi_t}) + \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big) \\
&\overset{\text{(vii)}}{=} r_t(\mu_t^{N_1}, \nu_t^{N_2}) + \min_{\nu_{t+1} \in \mathbb{R}_{\nu,t}(\mu_t^{N_1}, \nu_t^{N_2})} J_{\infty,t+1}^{\phi^*,\tilde{\psi}_\infty^*}(\mu_{t+1}^{\phi^*}, \nu_{t+1}^{\psi_t}) + \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big) \\
&\overset{\text{(viii)}}{=} J_{\infty,t}^{\phi^*,\tilde{\psi}_\infty^*}(\mu_t^{N_1}, \nu_t^{N_2}) + \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big)
\end{aligned}
$$

For inequality (i), we used the inductive hypothesis; for inequality (ii), we reduced the optimization domain of the Red team to the set of *identical* team policies; inequality (iv) is a result of the Lipschitz continuity of the value function [Guan et al., 2024a]; for equality (v), we use the fact that the mean-fields $(\mu_{t+1}^{\phi^*}, \nu_{t+1}^{\psi_t})$ are induced deterministically from the distributions at time t; inequality (vi) holds as a consequence of Corollary 1; (vii) converts the optimization domain from the policy space $\psi_t \in \Psi_t$ to the corresponding reachable set $\mathbb{R}_{\nu,t}(\mu_t^{N_1}, \nu_t^{N_2}) \triangleq \{\nu_{t+1} | \exists \psi_t \in \Psi_t \text{ s.t.} \nu_{t+1} = \nu_t G_t(\mu_t, \nu_t, \psi_t)\}$ following [Guan et al., 2024a]; (vii) follows from the definition of $J_{\infty,t}^{\phi^*,\tilde{\psi}_\infty^*}(\mu_t^{N_1}, \nu_t^{N_2})$. Thus, $J^{N,\phi^*,\psi^{N_2*}}(\mathbf{x}^{N_1}, \mathbf{y}^{N_2}) = J_0^{N,\phi^*,\psi^{N_2*}}(\mathbf{x}^{N_1}, \mathbf{y}^{N_2})$ and (G.35) follows.

Step 2. We show that for ϕ^* from $\mathcal{G}_1$ we have the following inequality for all joint states $\mathbf{x}^{\bar{N}_1} \in \mathcal{X}^{\bar{N}_1}$ and $\mathbf{y}^{\bar{N}_2} \in \mathcal{Y}^{\bar{N}_2}$ in $\mathcal{G}_2$,

$$
\min_{\psi^{\bar{N}_2} \in \Psi^{\bar{N}_2}} J^{\bar{N},\phi^*,\psi^{\bar{N}_2}}(\mathbf{x}^{\bar{N}_1}, \mathbf{y}^{\bar{N}_2}) \geq J_\infty^{\phi^*,\tilde{\psi}_\infty^*}(\mu^{\bar{N}_1}, \nu^{\bar{N}_2}) - \mathcal{O}\Big(\frac{1}{\sqrt{\underline{\bar{N}}}}\Big), \tag{G.39}
$$

where $\mu^{\bar{N}_1} = \mathrm{Emp}_\mu(\mathbf{x}^{\bar{N}_1})$, $\nu^{\bar{N}_2} = \mathrm{Emp}_\nu(\mathbf{y}^{\bar{N}_2})$ and $\underline{\bar{N}} = \min(\bar{N}_1, \bar{N}_2)$. The proof is constructed based on induction. Fix an arbitrary Red team policy $\psi^{\bar{N}_2} \in \Psi^{\bar{N}_2}$.

Base case: At the terminal timestep T, since there is no decision to be made, both value functions are equal to the terminal reward and are thus the same. Formally, for all $\mathbf{x}_T^{\bar{N}_1} \in \mathbf{X}^{\bar{N}_1}$ and $\mathbf{y}_T^{\bar{N}_2} \in \mathbf{Y}^{\bar{N}_2}$,

$$
J_T^{\bar{N},\phi^*,\psi^{\bar{N}_2}}(\mathbf{x}_T^{\bar{N}_1}, \mathbf{y}_T^{\bar{N}_2}) = J_{\infty,T}^{\phi^*,\tilde{\psi}_\infty^*}(\mu_T^{N_1}, \nu_T^{N_1}) = r_t(\mu_T^{\bar{N}_1}, \nu_T^{\bar{N}_1}),
$$

where $\mu_T^{\bar{N}_1} = \mathrm{Emp}_\mu(\mathbf{x}_T^{\bar{N}_1})$ and $\nu_T^{\bar{N}_2} = \mathrm{Emp}_\nu(\mathbf{y}_T^{\bar{N}_2})$. For simplicity, we do not emphasize the correspondence between the joint states and the EDs for the rest of the proof, as it is clear from the context.

Inductive hypothesis: Assume that at $t+1$, the following holds for all joint states $\mathbf{x}_{t+1}^{\bar{N}_1} \in \mathbf{X}^{\bar{N}_1}$ and $\mathbf{y}_{t+1}^{\bar{N}_2} \in \mathbf{Y}^{\bar{N}_2}$:

$$J_{t+1}^{\bar{N},\phi^*,\psi^{\bar{N}_2}}(\mathbf{x}_{t+1}^{\bar{N}_1}, \mathbf{y}_{t+1}^{N_2}) \geq J_{\infty,t+1}^{\phi^*,\tilde{\psi}_\infty^*}(\mu_{t+1}^{\bar{N}_1}, \nu_{t+1}^{\bar{N}_1}) - \mathcal{O}\Big(\frac{1}{\sqrt{\underline{\bar{N}}}}\Big). \tag{G.40}$$

Induction: At timestep t, consider an arbitrary pair of joint states $\mathbf{x}_t^{\bar{N}_1} \in \mathbf{X}^{\bar{N}_1}$ and $\mathbf{y}_t^{\bar{N}_2} \in \mathbf{Y}^{\bar{N}_2}$, and their corresponding EDs $\mu_t^{\bar{N}_1}$ and $\nu_t^{\bar{N}_2}$.

For the identical team policy $\phi^* \in \Phi$, denote $\mu_{t+1}^{\phi^*})$ from (G.38). Furthermore, from Theorem 1 in Guan et al. [2024a] there exists a $\nu_{\text{apprx},t+1}^{\psi_t^{\bar{N}_2}}$ for the Red team policy $\psi_t^{N_2}$ such that

$$\mathbb{E}_{\psi_t^{\bar{N}_2}}\left[\mathrm{d_{TV}}\big(\mathcal{N}_{t+1}^{\bar{N}_2}, \nu_{\text{apprx},t+1}^{\psi_t^{\bar{N}_2}}\big)\big|\mathbf{X}_t^{\bar{N}_1} = \mathbf{x}_t^{\bar{N}_1}, \mathbf{Y}_t^{\bar{N}_2} = \mathbf{y}_t^{\bar{N}_2}\right] \leq \frac{|\mathcal{Y}|}{2}\sqrt{\frac{1}{\bar{N}_2}}. \tag{G.41}$$

Then, for all joint states $\mathbf{x}_t^{\bar{N}_1} \in \mathbf{X}^{\bar{N}_1}$ and $\mathbf{y}_t^{\bar{N}_2} \in \mathbf{Y}^{\bar{N}_2}$, we have

$$\begin{aligned}
&J_t^{\bar{N},\phi^*,\psi^{\bar{N}_2}}(\mathbf{x}_t^{\bar{N}_1}, \mathbf{y}_t^{\bar{N}_2}) \\
&= r_t(\mu_t^{\bar{N}_1}, \nu_t^{\bar{N}_2}) + \mathbb{E}_{\phi^*,\psi^{\bar{N}_2}}\Big[J_{t+1}^{\bar{N},\phi^*,\psi^{\bar{N}_2}}(\mathbf{X}_{t+1}^{\bar{N}_1}, \mathbf{Y}_{t+1}^{\bar{N}_2})\Big] \\
&\overset{\text{(i)}}{\geq} r_t(\mu_t^{\bar{N}_1}, \nu_t^{\bar{N}_2}) + \mathbb{E}_{\phi^*,\psi^{\bar{N}_2}}\Big[J_{\infty,t+1}^{\phi^*,\tilde{\psi}_\infty^*}(\mathcal{M}_{t+1}^{\bar{N}_1}, \mathcal{N}_{t+1}^{\bar{N}_2})\Big] - \mathcal{O}\Big(\frac{1}{\sqrt{\underline{\bar{N}}}}\Big) \\
&= r_t(\mu_t^{\bar{N}_1}, \nu_t^{\bar{N}_2}) - \mathcal{O}\Big(\frac{1}{\sqrt{\underline{\bar{N}}}}\Big) + \mathbb{E}_{\phi^*,\psi^{\bar{N}_2}}\Big[J_{\infty,t+1}^{\phi^*,\tilde{\psi}_\infty^*}(\mathcal{M}_{t+1}^{\bar{N}_1}, \mathcal{N}_{t+1}^{\bar{N}_2}) \\
&\qquad\qquad - J_{\infty,t+1}^{\phi^*,\tilde{\psi}_\infty^*}(\mu_{t+1}^{\phi^*}, \nu_{\text{apprx},t+1}^{\psi_t^{\bar{N}_2}}) + J_{\infty,t+1}^{\phi^*,\tilde{\psi}_\infty^*}(\mu_{t+1}^{\phi^*}, \nu_{\text{apprx},t+1}^{\psi_t^{\bar{N}_2}})\Big] \\
&\overset{\text{(ii)}}{\geq} r_t(\mu_t^{\bar{N}_1}, \nu_t^{\bar{N}_2}) + J_{\infty,t+1}^{\phi^*,\tilde{\psi}_\infty^*}(\mu_{t+1}^{\phi^*}, \nu_{\text{apprx},t+1}^{\psi_t^{\bar{N}_2}}) - \mathcal{O}\Big(\frac{1}{\sqrt{\underline{\bar{N}}}}\Big) - L_{J,t+1}\underbrace{\mathbb{E}_{\phi^*}\Big[\mathrm{d_{TV}}\big(\mathcal{M}_{t+1}^{\bar{N}_1}, \mu_{t+1}^{\phi^*}\big)\Big]}_{=\mathcal{O}(\frac{1}{\sqrt{\bar{N}_1}}) \text{ due to Corollary 1}} \\
&\qquad\qquad - \underbrace{L_{J,t+1}\mathbb{E}_{\psi^{N_2}}\Big[\mathrm{d_{TV}}\big(\mathcal{N}_{t+1}^{N_2}, \nu_{\text{apprx},t+1}^{\psi_t^{\bar{N}_2}}\big)\Big]}_{=\mathcal{O}(\frac{1}{\sqrt{\bar{N}_2}}) \text{ due to(G.41)}} \\
&\overset{\text{(iii)}}{\geq} r_t(\mu_t^{\bar{N}_1}, \nu_t^{\bar{N}_2}) + J_{\infty,t+1}^{\phi^*,\tilde{\psi}_\infty^*}(\mu_{t+1}^{\phi^*}, \nu_{\text{apprx},t+1}^{\psi_t^{\bar{N}_2}}) - \mathcal{O}\Big(\frac{1}{\sqrt{\underline{\bar{N}}}}\Big) \\
&\overset{\text{(iv)}}{\geq} r_t(\mu_t^{\bar{N}_1}, \nu_t^{\bar{N}_2}) + \min_{\nu_{t+1}\in\mathbb{R}_{\nu,t}(\mu_t^{\bar{N}_1}, \nu_t^{\bar{N}_2})} J_{\infty,t+1}^{\phi^*,\tilde{\psi}_\infty^*}(\mu_{t+1}^{\phi^*}, \nu_{t+1}) - \mathcal{O}\Big(\frac{1}{\sqrt{\underline{\bar{N}}}}\Big) \\
&\overset{\text{(v)}}{=} J_{\infty,t}^{\phi^*,\tilde{\psi}_\infty^*}(\mu_t^{\bar{N}_1}, \nu_t^{\bar{N}_2}) - \mathcal{O}\Big(\frac{1}{\sqrt{\underline{\bar{N}}}}\Big).
\end{aligned}$$

For inequality (i), we used the inductive hypothesis; for inequality (ii), we utilized the Lipschitz continuity of the coordinator value function [Guan et al., 2024a] using mean-field distributions $(\mu_{t+1}^{\phi^*}, \nu_{\text{apprx},t+1}^{\psi_t^{\bar{N}_2}})$ from (G.38) and the deterministic transitions from t to $t+1$; inequality (iv) follows from the definition of the $\min$ operator; and equality (v) follows from the definition of $J_{\infty,t}^{\phi^*,\tilde{\psi}_\infty^*}(\mu_t^{N_1}, \nu_t^{N_2})$ which completes the induction.

Since the Red team policy $\psi^{\bar{N}_2} \in \Psi^{\bar{N}_2}$ is arbitrary, we have that, for all joint states $\mathbf{x}^{\bar{N}_1} \in \mathbf{X}^{\bar{N}_1}$ and $\mathbf{y}^{\bar{N}_2} \in \mathbf{Y}^{\bar{N}_2}$,

$$\begin{aligned}
\min_{\psi^{\bar{N}_2}\in\Psi^{\bar{N}_2}} J^{\bar{N},\phi^*,\psi^{\bar{N}_2}}(\mathbf{x}^{\bar{N}_1}, \mathbf{y}^{\bar{N}_2}) &= \min_{\psi^{\bar{N}_2}\in\Psi^{\bar{N}_2}} J_0^{\bar{N},\phi^*,\psi^{\bar{N}_2}}(\mathbf{x}^{\bar{N}_1}, \mathbf{y}^{\bar{N}_2}) \\
&\geq \underline{J}_\infty^{\phi^*,\tilde{\psi}_\infty^*}(\mu^{\bar{N}_1}, \nu^{\bar{N}_2}) - \mathcal{O}\Big(\frac{1}{\sqrt{\underline{\bar{N}}}}\Big).
\end{aligned}$$

Step 3. Combining (G.35) from Step 1 with Theorem 3.1, we have for all joint states $\mathbf{x}^{N_1} \in \mathcal{X}^{N_1}$ and $\mathbf{y}^{N_2} \in \mathcal{Y}^{N_2}$,

$$J_\infty^{\phi^*,\tilde{\psi}_\infty^*}(\mu^{N_1},\nu^{N_1}) \geq J^{N,\phi^*,\psi^{N_2*}}(\mathbf{x}^{N_1},\mathbf{y}^{N_2}) - \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big)$$
$$\geq \underline{J}^{N*}(\mathbf{x}^{N_1},\mathbf{y}^{N_2}) - \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big), \tag{G.42}$$

where $\mu^{N_1} = \mathrm{Emp}_\mu(\mathbf{x}^{N_1})$ and $\nu^{N_2} = \mathrm{Emp}_\nu(\mathbf{y}^{N_2})$.

Assuming that the initial distributions are the same for $\mathcal{G}_1$ and $\mathcal{G}_2$, i.e., $\mu^{N_1} = \mu^{\bar{N}_1}$ and $\nu^{N_2} = \nu^{\bar{N}_2}$ we have the following sequence of inequalities

$$\min_{\psi^{\bar{N}_2}\in\Psi^{\bar{N}_2}} J^{\bar{N},\phi^*,\psi^{\bar{N}_2}}(\mathbf{x}^{\bar{N}_1},\mathbf{y}^{\bar{N}_2}) \overset{\text{(i)}}{\geq} J_\infty^{\phi^*,\tilde{\psi}_\infty^*}(\mu^{\bar{N}_1},\nu^{\bar{N}_2}) - \mathcal{O}\Big(\frac{1}{\sqrt{\underline{\bar{N}}}}\Big)$$
$$\overset{\text{(ii)}}{\geq} \underline{J}^{N*}(\mathbf{x}^{N_1},\mathbf{y}^{N_2}) - \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big) - \mathcal{O}\Big(\frac{1}{\sqrt{\underline{\bar{N}}}}\Big)$$
$$\overset{\text{(iii)}}{\geq} \underline{J}_\infty^{\phi_\infty^*,\psi_\infty^*}(\mu^{N_1},\nu^{N_1}) - \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big) - \mathcal{O}\Big(\frac{1}{\sqrt{\underline{\bar{N}}}}\Big)$$
$$\overset{\text{(iv)}}{\geq} \underline{J}^{\bar{N}*}(\mathbf{x}^{\bar{N}_1},\mathbf{y}^{\bar{N}_2}) - \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big) - \mathcal{O}\Big(\frac{1}{\sqrt{\underline{\bar{N}}}}\Big)$$

where $J_\infty^{\phi_\infty^*,\psi_\infty^*}(\mu_0,\nu_0) = \max_{\phi\in\Phi}\min_{\psi\in\Psi}\ J_\infty^{\phi,\psi}(\mu_0,\nu_0)$. Note that (i) follows from Step 2 (G.39); (ii) is a consequence of (G.42) and inequalities (iii) and (iv) follow from Lemmas 5 (applied on game $\mathcal{G}_1$) and 6 (applied on game $\mathcal{G}_2$) in Guan et al. [2024b] respectively.

By expanding the terms $\mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big)$ and $\mathcal{O}\Big(\frac{1}{\sqrt{\underline{\bar{N}}}}\Big)$ from Guan et al. [2024b], we can fully characterize the transfer bound across population sizes. In particular, for the $(\bar{N}_1,\bar{N}_2)$-population game, ϕ^* satisfies

$$\min_{\psi^{\bar{N}_2}\in\Psi^{\bar{N}_2}} J^{\bar{N},\phi^*,\psi^{\bar{N}_2}}(\mathbf{x}^{\bar{N}_1},\mathbf{y}^{\bar{N}_2}) \geq \underline{J}^{\bar{N}*}(\mathbf{x}^{\bar{N}_1},\mathbf{y}^{\bar{N}_2}) \tag{G.43}$$
$$- \sum_{t=0}^{T} L_{J,t}\Big(\frac{|\mathcal{X}|}{2}\sqrt{\frac{1}{N_1}} + \frac{|\mathcal{Y}|}{2}\sqrt{\frac{1}{N_2}}\Big) - \sum_{t=0}^{T} L_{J,t}\Big(\frac{|\mathcal{X}|}{2}\sqrt{\frac{1}{\bar{N}_1}} + \frac{|\mathcal{Y}|}{2}\sqrt{\frac{1}{\bar{N}_2}}\Big), \tag{G.44}$$

where $L_{J,t}$ denotes the Lipschitz constant of the time-t value function. Note that $\sum_{t=0}^{T} L_{J,t}\Big(\frac{|\mathcal{X}|}{2}\sqrt{\frac{1}{N_1}} + \frac{|\mathcal{Y}|}{2}\sqrt{\frac{1}{N_2}}\Big) = \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big) = \epsilon$, where $\underline{N} = \min(N_1,N_2)$ and similarly, we can define $\mathcal{O}\Big(\frac{1}{\sqrt{\underline{\bar{N}}}}\Big) = \bar{\epsilon}$. Thus, the identical policy pair learned in the smaller finite-population game transfers to the larger game with a vanishing finite-population error, and the corresponding optimality gap is bounded by $\epsilon + \bar{\epsilon}$. Combined with the fact that $\min(\bar{N}_1,\bar{N}_2) > \min(N_1,N_2)$, we obtain ϵ-optimality for the game $\mathcal{G}_{\bar{N}}$.

Remark 4. Theorem 3.3 applies to the *exact optimizers* of the finite-population optimization restricted to identical team policies. The suboptimality bound comprises two components: the first captures the approximation error of the identical policy ϕ^* trained in the (N_1,N_2)-population game (Theorem 3.1); the second captures the stochasticity mismatch between training and deployment due to differing population sizes. Thus, for a fixed deployment population, increasing the training population size can improve the guarantee by reducing the first component, while training beyond the target size offers diminishing returns. For simplicity, we assume matching initial empirical distributions. The result can be extended to mismatched initial distributions using a similar argument to that in the original mean-field paper [Huang et al., 2006].

□

H Proof of Proposition 1

Proposition 1. *If the log-probability of the Blue team policy is uniformly bounded such that for all $x \in \mathcal{X}, u \in \mathcal{U}, \mu \in \mathcal{P}(\mathcal{X}), \nu \in \mathcal{P}(\mathcal{Y})$, $\|\nabla_\eta \log \phi(u \mid x, \mu, \nu)\|_2 \leq L_\phi/2|\mathcal{U}|$, where the gradient is taken with respect to $\eta \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})$, then $\phi(u \mid x, \mu, \nu)$ is L_ϕ-Lipschitz continuous, i.e., $\sum_u |\phi_t(u|x,\mu,\nu) - \phi_t(u|x,\mu',\nu')| \leq L_\phi \left(\mathrm{d_{TV}}(\mu,\mu') + \mathrm{d_{TV}}(\nu,\nu')\right)$ for all $x \in \mathcal{X}$.*

Proof. We prove the proposition in two parts.

Step 1. We first show that $\|\nabla_\eta \log \phi(u \mid x, \mu, \nu)\|_2 \leq \frac{L_\phi}{2|\mathcal{U}|}$ implies

$$\sum_u |\log \phi_t(u|x,\hat{\mu},\hat{\nu}) - \log \phi_t(u|x,\hat{\mu}',\hat{\nu}')| \leq L_\phi \left(\mathrm{d_{TV}}(\hat{\mu},\hat{\mu}') + \mathrm{d_{TV}}(\hat{\nu},\hat{\nu}')\right) \quad \forall\ x \in \mathcal{X}. \tag{H.45}$$

By the Definition of η, we can denote $\log \phi_t(u|x,\eta) \triangleq \log \phi_t(u|x,\hat{\mu},\hat{\nu})$. By the Mean-Value Theorem and the bounded gradient from Step 1 we have for all $x \in \mathcal{X}$, $u \in \mathcal{U}$, $\eta \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})$,

$$\begin{aligned}
|\log \phi_t(u|x,\eta) - \log \phi_t(u|x,\eta')| &\leq \sup_{\eta \in \mathcal{P}(\mathcal{X})\times\mathcal{P}(\mathcal{Y})} \|\nabla_\eta \log \phi(u \mid x,\eta)\|_2 \, \|\eta - \eta'\|_2 \\
&\leq \sup_{\eta \in \mathcal{P}(\mathcal{X})\times\mathcal{P}(\mathcal{Y})} \|\nabla_\eta \log \phi(u \mid x,\eta)\|_2 \, \|\eta - \eta'\|_1 \\
&\leq \frac{L_\phi}{2|\mathcal{U}|} \cdot 2\left(\mathrm{d_{TV}}(\hat{\mu},\hat{\mu}') + \mathrm{d_{TV}}(\hat{\nu},\hat{\nu}')\right) \\
&= \frac{L_\phi}{|\mathcal{U}|}\left(\mathrm{d_{TV}}(\hat{\mu},\hat{\mu}') + \mathrm{d_{TV}}(\hat{\nu},\hat{\nu}')\right).
\end{aligned}$$

Summing over all actions for any given $x \in \mathcal{X}$,

$$\begin{aligned}
\sum_u |\log \phi_t(u|x,\eta) - \log \phi_t(u|x,\eta')| &\leq \sum_u \frac{L_\phi}{|\mathcal{U}|}\left(\mathrm{d_{TV}}(\hat{\mu},\hat{\mu}') + \mathrm{d_{TV}}(\hat{\nu},\hat{\nu}')\right) \\
&\leq L_\phi \left(\mathrm{d_{TV}}(\hat{\mu},\hat{\mu}') + \mathrm{d_{TV}}(\hat{\nu},\hat{\nu}')\right) \sum_u \frac{1}{|\mathcal{U}|} \\
&\leq L_\phi \left(\mathrm{d_{TV}}(\hat{\mu},\hat{\mu}') + \mathrm{d_{TV}}(\hat{\nu},\hat{\nu}')\right),
\end{aligned}$$

completing the proof of Part 1.

Step 2. Using Step 1, we show that (H.45) implies

$$\sum_u |\phi_t(u|x,\hat{\mu},\hat{\nu}) - \phi_t(u|x,\hat{\mu}',\hat{\nu}')| \leq L_\phi \left(\mathrm{d_{TV}}(\hat{\mu},\hat{\mu}') + \mathrm{d_{TV}}(\hat{\nu},\hat{\nu}')\right) \quad \forall\ x \in \mathcal{X}. \tag{H.46}$$

Define $h(z) \triangleq e^z$ and values $a \triangleq \log \phi_t(u|x,\hat{\mu}',\hat{\nu}')$ and $b \triangleq \log \phi_t(u|x,\hat{\mu},\hat{\nu})$. It follows from the Mean-Value Theorem that for all $x \in \mathcal{X}$ and $u \in \mathcal{U}$,

$$\begin{aligned}
|h(b) - h(a)| &= |h'(c)||b-a|,\ \min(a,b) \leq c \leq \max(a,b) \\
|\phi_t(u|x,\hat{\mu},\hat{\nu}) - \phi_t(u|x,\hat{\mu}',\hat{\nu}')| &= e^c |\log \phi_t(u|x,\hat{\mu},\hat{\nu}) - \log \phi_t(u|x,\hat{\mu}',\hat{\nu}')| \\
&\leq e^{\max(a,b)} |\log \phi_t(u|x,\hat{\mu},\hat{\nu}) - \log \phi_t(u|x,\hat{\mu}',\hat{\nu}')|.
\end{aligned}$$

Note that $e^a = \phi_t(u|x,\hat{\mu}',\hat{\nu}') \leq 1$ and $e^b = \phi_t(u|x,\hat{\mu},\hat{\nu}) \leq 1$. Consequently $e^{\max(a,b)} \leq 1$ and we have for all $x \in \mathcal{X}$,

$$\begin{aligned}
|\phi_t(u|x,\hat{\mu},\hat{\nu}) - \phi_t(u|x,\hat{\mu}',\hat{\nu}')| &\leq |\log \phi_t(u|x,\hat{\mu},\hat{\nu}) - \log \phi_t(u|x,\hat{\mu}',\hat{\nu}')| \\
\sum_u |\phi_t(u|x,\hat{\mu},\hat{\nu}) - \phi_t(u|x,\hat{\mu}',\hat{\nu}')| &\leq \sum_u |\log \phi_t(u|x,\hat{\mu},\hat{\nu}) - \log \phi_t(u|x,\hat{\mu}',\hat{\nu}')| \\
\sum_u |\phi_t(u|x,\hat{\mu},\hat{\nu}) - \phi_t(u|x,\hat{\mu}',\hat{\nu}')| &\leq L_\phi \left(\mathrm{d_{TV}}(\hat{\mu},\hat{\mu}') + \mathrm{d_{TV}}(\hat{\nu},\hat{\nu}')\right),
\end{aligned}$$

where the last inequality follows from (H.45) □

I Proof of Proposition 2

Proposition 2. *When a policy pair (ϕ_t^*, ψ_t^*) learned via GR-MF-MAPPO is deployed with an ϵ-accurate estimator and satisfies a uniform Lipschitz bound, the cumulative regret satisfies $\Delta J(\phi_t^*, \psi_t^*) \leq K\epsilon + \mathcal{O}(1/\sqrt{N})$.*

Sketch of Proof. MFs induced by identical team policies in an infinite-population game closely approximate the EDs induced in the corresponding finite-population game. Combining this consequence with Lipschitz continuity in the dynamics, rewards (Assumption 1) and policies (Proposition 1), we prove the statement. □

Proof. For the performance guarantee,

$$\Delta J(\phi_t^*, \psi_t^*) = \mathbb{E}_{\phi^*, \psi^*}\left[\left|\sum_{t=0}^{T} r_t(\mathcal{M}_t^{N_1}, \mathcal{N}_t^{N_2}) - \sum_{t=0}^{T} r_t(\hat{\mathcal{M}}_t^{N_1}, \hat{\mathcal{N}}_t^{N_2})\right|\right], \tag{I.47}$$

we have,

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

From Lipschitz rewards (Assumption 1), we have,

$$\begin{aligned}
&\Delta J(\phi_t^*, \psi_t^*; \Gamma) \\
&= \mathbb{E}_{\phi^*, \psi^*}\Big[\Big|\sum_{t=0}^{T} r_t(\mathcal{M}_t^{N_1}, \mathcal{N}_t^{N_2}) - \sum_{t=0}^{T} r_t(\hat{\mathcal{M}}_t^{N_1}, \hat{\mathcal{N}}_t^{N_2})\Big| \Big| \mathbf{x}_0^{N_1} = \hat{\mathbf{x}}_0^{N_1}, \mathbf{y}_0^{N_2} = \hat{\mathbf{y}}_0^{N_2}\Big] \\
&\leq \mathbb{E}_{\phi^*, \psi^*}\Big[\sum_{t=0}^{T} L_r\left(\mathrm{d_{TV}}\big(\mathcal{M}_t^{N_1}, \hat{\mathcal{M}}_t^{N_1}\big) + \mathrm{d_{TV}}\big(\mathcal{N}_t^{N_2}, \hat{\mathcal{N}}_t^{N_2}\big)\right) \Big| \mathbf{x}_0^{N_1} = \hat{\mathbf{x}}_0^{N_1}, \mathbf{y}_0^{N_2} = \hat{\mathbf{y}}_0^{N_2}\Big] \\
&\leq L_r \sum_{t=0}^{T} \mathbb{E}_{\phi^*, \psi^*}\Big[\mathrm{d_{TV}}\big(\mathcal{M}_t^{N_1}, \hat{\mathcal{M}}_t^{N_1}\big) + \mathrm{d_{TV}}\big(\mathcal{N}_t^{N_2}, \hat{\mathcal{N}}_t^{N_2}\big) \Big| \mathbf{x}_0^{N_1} = \hat{\mathbf{x}}_0^{N_1}, \mathbf{y}_0^{N_2} = \hat{\mathbf{y}}_0^{N_2}\Big],
\end{aligned} \tag{I.48}$$

where the last inequality follows from the linearity of expectations.

Step 1. We first show the following for all time steps t,

$$\begin{aligned}
&\mathbb{E}_{\phi^*, \psi^*}\Big[\mathrm{d_{TV}}\big(\mathcal{M}_{t+1}^{N_1}, \hat{\mathcal{M}}_{t+1}^{N_1}\big) + \mathrm{d_{TV}}\big(\mathcal{N}_{t+1}^{N_2}, \hat{\mathcal{N}}_{t+1}^{N_2}\big) \Big| \mathbf{x}_t^{N_1}, \hat{\mathbf{x}}_t^{N_1}, \mathbf{y}_t^{N_2}, \hat{\mathbf{y}}_t^{N_2}\Big] \\
&\leq \kappa_1 + \kappa_2\left(\mathrm{d_{TV}}\big(\mathcal{M}_t^{N_1}, \hat{\mathcal{M}}_t^{N_1}\big) + \mathrm{d_{TV}}\big(\mathcal{N}_t^{N_2}, \hat{\mathcal{N}}_t^{N_2}\big)\right),
\end{aligned}$$

where $\kappa_1 \triangleq \frac{1}{2}(L_\phi + L_\psi)\epsilon + \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big)$ and $\kappa_2 \triangleq (1 + L_f + \frac{1}{2}L_\phi + \frac{1}{2}L_\psi)$.

By the linearity property of expectations,

$$\begin{aligned}
&\mathbb{E}_{\phi^*, \psi^*}\Big[\mathrm{d_{TV}}\big(\mathcal{M}_{t+1}^{N_1}, \hat{\mathcal{M}}_{t+1}^{N_1}\big) + \mathrm{d_{TV}}\big(\mathcal{N}_{t+1}^{N_2}, \hat{\mathcal{N}}_{t+1}^{N_2}\big) \Big| \mathbf{x}_t^{N_1}, \hat{\mathbf{x}}_t^{N_1}, \mathbf{y}_t^{N_2}, \hat{\mathbf{y}}_t^{N_2}\Big] \\
&= \mathbb{E}_{\phi^*}\Big[\mathrm{d_{TV}}\big(\mathcal{M}_{t+1}^{N_1}, \hat{\mathcal{M}}_{t+1}^{N_1}\big) \Big| \mathbf{x}_t^{N_1}, \hat{\mathbf{x}}_t^{N_1}, \mathbf{y}_t^{N_2}, \hat{\mathbf{y}}_t^{N_2}\Big] + \mathbb{E}_{\psi^*}\Big[\mathrm{d_{TV}}\big(\mathcal{N}_{t+1}^{N_2}, \hat{\mathcal{N}}_{t+1}^{N_2}\big) \Big| \mathbf{x}_t^{N_1}, \hat{\mathbf{x}}_t^{N_1}, \mathbf{y}_t^{N_2}, \hat{\mathbf{y}}_t^{N_2}\Big]
\end{aligned}$$

We now bound the Blue team's mean-field, i.e., the first term on the RHS. The Red team can be bounded similarly (second term).

$$\begin{aligned}
&\mathbb{E}_{\phi^*}\Big[\mathrm{d_{TV}}\big(\mathcal{M}_{t+1}^{N_1}, \hat{\mathcal{M}}_{t+1}^{N_1}\big) \Big| \mathbf{x}_t^{N_1}, \hat{\mathbf{x}}_t^{N_1}, \mathbf{y}_t^{N_2}, \hat{\mathbf{y}}_t^{N_2}\Big] \\
&\leq \mathbb{E}_{\phi^*}\Big[\mathrm{d_{TV}}\big(\mathcal{M}_{t+1}^{N_1}, \mathcal{M}_{t+1}\big) \Big| \mathbf{x}_t^{N_1}, \hat{\mathbf{x}}_t^{N_1}, \mathbf{y}_t^{N_2}, \hat{\mathbf{y}}_t^{N_2}\Big] \\
&+ \mathbb{E}_{\phi^*}\Big[\mathrm{d_{TV}}\big(\mathcal{M}_{t+1}, \hat{\mathcal{M}}_{t+1}\big) \Big| \mathbf{x}_t^{N_1}, \hat{\mathbf{x}}_t^{N_1}, \mathbf{y}_t^{N_2}, \hat{\mathbf{y}}_t^{N_2}\Big] \\
&+ \mathbb{E}_{\phi^*}\Big[\mathrm{d_{TV}}\big(\hat{\mathcal{M}}_{t+1}^{N_1}, \hat{\mathcal{M}}_{t+1}\big) \Big| \mathbf{x}_t^{N_1}, \hat{\mathbf{x}}_t^{N_1}, \mathbf{y}_t^{N_2}, \hat{\mathbf{y}}_t^{N_2}\Big],
\end{aligned} \tag{I.49}$$

where $\mathcal{M}_{t+1} = \mathcal{M}_t^{N_1} F(\mathcal{M}_t^{N_1}, \mathcal{N}_t^{N_2}, \phi^*)$ and $\hat{\mathcal{M}}_{t+1} = \hat{\mathcal{M}}_t^{N_1} F(\hat{\mathcal{M}}_t^{N_1}, \hat{\mathcal{N}}_t^{N_2}, \phi^*, \Gamma_{\text{D-PC}})$ are shorthand notations for the next induced mean-field from the *infinite*-population *deterministic* dynamics [Guan et al., 2024a] with and without the estimator $\Gamma_{\text{D-PC}}$. From Corollary 1 in Guan et al. [2024a], we have

$$\mathbb{E}_{\phi^*}\Big[\mathrm{d_{TV}}\big(\mathcal{M}_{t+1}^{N_1}, \mathcal{M}_{t+1}\big)\Big|\mathbf{x}_t^{N_1}, \hat{\mathbf{x}}_t^{N_1}, \mathbf{y}_t^{N_2}, \hat{\mathbf{y}}_t^{N_2}\Big] \leq \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big),$$
$$\mathbb{E}_{\phi^*}\Big[\mathrm{d_{TV}}\big(\hat{\mathcal{M}}_{t+1}^{N_1}, \hat{\mathcal{M}}_{t+1}\big)\Big|\mathbf{x}_t^{N_1}, \hat{\mathbf{x}}_t^{N_1}, \mathbf{y}_t^{N_2}, \hat{\mathbf{y}}_t^{N_2}\Big] \leq \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big).$$

Thus, we are left to simplify

$$\begin{aligned}
&\mathbb{E}_{\phi^*}\Big[\mathrm{d_{TV}}\big(\mathcal{M}_{t+1}, \hat{\mathcal{M}}_{t+1}\big)\Big|\mathbf{x}_t^{N_1}, \hat{\mathbf{x}}_t^{N_1}, \mathbf{y}_t^{N_2}, \hat{\mathbf{y}}_t^{N_2}\Big] \\
&= \mathrm{d_{TV}}\big(\mathcal{M}_t^{N_1} F(\mathcal{M}_t^{N_1}, \mathcal{N}_t^{N_2}, \phi^*), \hat{\mathcal{M}}_t^{N_1} F(\hat{\mathcal{M}}_t^{N_1}, \hat{\mathcal{N}}_t^{N_2}, \phi^*, \Gamma_{\text{D-PC}})\big),
\end{aligned}$$

where the expectation vanishes due to *deterministic* transitions under identical policies (ϕ^*, ψ^*). Now,

$$\begin{aligned}
&2\mathrm{d_{TV}}\big(\mathcal{M}_t^{N_1} F(\mathcal{M}_t^{N_1}, \mathcal{N}_t^{N_2}, \phi^*), \hat{\mathcal{M}}_t^{N_1} F(\hat{\mathcal{M}}_t^{N_1}, \hat{\mathcal{N}}_t^{N_2}, \phi^*, \Gamma_{\text{D-PC}})\big) \\
&= \sum_{x' \in \mathcal{X}} \Big| \sum_{x \in \mathcal{X}} \sum_{u \in \mathcal{U}} f_t(x'|x, u, \mathcal{M}_t^{N_1}, \mathcal{N}_t^{N_2}) \phi_t^*(u|x, \mathcal{M}_t^{N_1}, \mathcal{N}_t^{N_2}) \mathcal{M}_t^{N_1}(x) \\
&- \sum_{x \in \mathcal{X}} \sum_{u \in \mathcal{U}} f_t(x'|x, u, \hat{\mathcal{M}}_t^{N_1}, \hat{\mathcal{N}}_t^{N_2}) \phi_t^*(u|x, \hat{\mathcal{M}}_t^{N_1}, \hat{\mathcal{N}}_{x,t}^{N_2}) \hat{\mathcal{M}}_t^{N_1}(x) \Big| \\
&\leq \sum_{x' \in \mathcal{X}} \sum_{x \in \mathcal{X}} \sum_{u \in \mathcal{U}} \Big| f_t(x'|x, u, \mathcal{M}_t^{N_1}, \mathcal{N}_t^{N_2}) \phi_t^*(u|x, \mathcal{M}_t^{N_1}, \mathcal{N}_t^{N_2}) \mathcal{M}_t^{N_1}(x) \\
&- f_t(x'|x, u, \hat{\mathcal{M}}_t^{N_1}, \hat{\mathcal{N}}_t^{N_2}) \phi_t^*(u|x, \hat{\mathcal{M}}_t^{N_1}, \hat{\mathcal{N}}_{x,t}^{N_2}) \hat{\mathcal{M}}_t^{N_1}(x) \Big|.
\end{aligned} \tag{I.50}$$

Firstly, we add and subtract $f_t(x'|x, u, \hat{\mathcal{M}}_t^{N_1}, \hat{\mathcal{N}}_t^{N_2}) \phi_t^*(u|x, \hat{\mathcal{M}}_t^{N_1}, \hat{\mathcal{N}}_t^{N_2}) \hat{\mathcal{M}}_t^{N_1}(x)$ to split absolute value as

$$2\mathrm{d_{TV}}\big(\mathcal{M}_t^{N_1} F(\mathcal{M}_t^{N_1}, \mathcal{N}_t^{N_2}, \phi^*), \hat{\mathcal{M}}_t^{N_1} F(\hat{\mathcal{M}}_t^{N_1}, \hat{\mathcal{N}}_t^{N_2}, \phi^*, \Gamma_{\text{D-PC}})\big) \leq \sum_{x' \in \mathcal{X}} \sum_{x \in \mathcal{X}} \sum_{u \in \mathcal{U}} (I_1 + I_2),$$

where,

$$I_1 \triangleq \Big| f_t(x'|x, u, \mathcal{M}_t^{N_1}, \mathcal{N}_t^{N_2}) \phi_t^*(u|x, \mathcal{M}_t^{N_1}, \mathcal{N}_t^{N_2}) \mathcal{M}_t^{N_1}(x) - f_t(x'|x, u, \hat{\mathcal{M}}_t^{N_1}, \hat{\mathcal{N}}_t^{N_2}) \phi_t^*(u|x, \hat{\mathcal{M}}_t^{N_1}, \hat{\mathcal{N}}_t^{N_2}) \hat{\mathcal{M}}_t^{N_1}(x) \Big|$$
$$I_2 \triangleq \Big| f_t(x'|x, u, \hat{\mathcal{M}}_t^{N_1}, \hat{\mathcal{N}}_t^{N_2}) \phi_t^*(u|x, \hat{\mathcal{M}}_t^{N_1}, \hat{\mathcal{N}}_t^{N_2}) \hat{\mathcal{M}}_t^{N_1}(x) - f_t(x'|x, u, \hat{\mathcal{M}}_t^{N_1}, \hat{\mathcal{N}}_t^{N_2}) \phi_t^*(u|x, \hat{\mathcal{M}}_t^{N_1}, \hat{\mathcal{N}}_{x,t}^{N_2}) \hat{\mathcal{M}}_t^{N_1}(x) \Big|.$$

Similar to (H.46), we can simplify I_2 using Proposition 1 and obtain

$$\sum_{x' \in \mathcal{X}} \sum_{x \in \mathcal{X}} \sum_{u \in \mathcal{U}} I_2 < L_\phi \epsilon.$$

We now individually bound the terms in I_1 using the definition of total variation distance, the identity

$$|abc - a'b'c'| \leq |a - a'|bc + a'|b - b'|c + a'b'|c - c'|, \; a, b, c \geq 0,$$

Assumption 1 and Proposition 1 as follows:

$$\sum_{x' \in \mathcal{X}} \sum_{x \in \mathcal{X}} \sum_{u \in \mathcal{U}} I_1 \leq 2\mathrm{d_{TV}}\big(\mathcal{M}_t^{N_1}, \hat{\mathcal{M}}_t^{N_1}\big) + (L_\phi + L_f)\Big(\mathrm{d_{TV}}\big(\mathcal{M}_t^{N_1}, \hat{\mathcal{M}}_t^{N_1}\big) + \mathrm{d_{TV}}\big(\mathcal{N}_t^{N_2}, \hat{\mathcal{N}}_t^{N_2}\big)\Big)$$

Thus we can substitute the expressions for I_1 and I_2 back in (I.50) and rewrite (I.49) to get

$$\begin{aligned}
&\mathbb{E}_{\phi^*}\Big[\mathrm{d_{TV}}\big(\mathcal{M}_{t+1}^{N_1},\hat{\mathcal{M}}_{t+1}^{N_1}\big)\Big|\mathbf{x}_t^{N_1},\hat{\mathbf{x}}_t^{N_1},\mathbf{y}_t^{N_2},\hat{\mathbf{y}}_t^{N_2}\Big]\\
&<\mathrm{d_{TV}}\big(\mathcal{M}_t^{N_1},\hat{\mathcal{M}}_t^{N_1}\big)\\
&+\frac{1}{2}(L_\phi+L_f)\left(\mathrm{d_{TV}}\big(\mathcal{M}_t^{N_1},\hat{\mathcal{M}}_t^{N_1}\big)+\mathrm{d_{TV}}\big(\mathcal{N}_t^{N_2},\hat{\mathcal{N}}_t^{N_2}\big)\right)+\frac{1}{2}L_\phi\epsilon+\mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big)
\end{aligned}\tag{I.51}$$

By symmetry,

$$\begin{aligned}
&\mathbb{E}_{\psi^*}\Big[\mathrm{d_{TV}}\big(\mathcal{N}_{t+1}^{N_2},\hat{\mathcal{N}}_{t+1}^{N_2}\big)\Big|\mathbf{x}_t^{N_1},\hat{\mathbf{x}}_t^{N_1},\mathbf{y}_t^{N_2},\hat{\mathbf{y}}_t^{N_2}\Big]\\
&<\mathrm{d_{TV}}\big(\mathcal{N}_t^{N_2},\hat{\mathcal{N}}_t^{N_2}\big)\\
&+\frac{1}{2}(L_\psi+L_f)\left(\mathrm{d_{TV}}\big(\mathcal{M}_t^{N_1},\hat{\mathcal{M}}_t^{N_1}\big)+\mathrm{d_{TV}}\big(\mathcal{N}_t^{N_2},\hat{\mathcal{N}}_t^{N_2}\big)\right)+\frac{1}{2}L_\psi\epsilon+\mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big)
\end{aligned}\tag{I.52}$$

Adding (I.51) and (I.52) we obtain,

$$\begin{aligned}
&\mathbb{E}_{\phi^*,\psi^*}\Big[\mathrm{d_{TV}}\big(\mathcal{M}_{t+1}^{N_1},\hat{\mathcal{M}}_{t+1}^{N_1}\big)+\mathrm{d_{TV}}\big(\mathcal{N}_{t+1}^{N_2},\hat{\mathcal{N}}_{t+1}^{N_2}\big)\Big|\mathbf{x}_t^{N_1},\hat{\mathbf{x}}_t^{N_1},\mathbf{y}_t^{N_2},\hat{\mathbf{y}}_t^{N_2}\Big]\\
&<(1+L_f+\frac{1}{2}L_\phi+\frac{1}{2}L_\psi)\left(\mathrm{d_{TV}}\big(\mathcal{M}_t^{N_1},\hat{\mathcal{M}}_t^{N_1}\big)+\mathrm{d_{TV}}\big(\mathcal{N}_t^{N_2},\hat{\mathcal{N}}_t^{N_2}\big)\right)+\frac{1}{2}(L_\phi+L_\psi)\epsilon+\mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big)\\
&=\kappa_1+\kappa_2\left(\mathrm{d_{TV}}\big(\mathcal{M}_t^{N_1}(x),\hat{\mathcal{M}}_t^{N_1}(x)\big)+\mathrm{d_{TV}}\big(\mathcal{N}_t^{N_2},\hat{\mathcal{N}}_t^{N_2}\big)\right).
\end{aligned}$$

Step 2. Define

$$a_t\triangleq\mathbb{E}_{\phi^*,\psi^*}\Big[\mathrm{d_{TV}}\big(\mathcal{M}_t^{N_1},\hat{\mathcal{M}}_t^{N_1}\big)+\mathrm{d_{TV}}\big(\mathcal{N}_t^{N_2},\hat{\mathcal{N}}_t^{N_2}\big)\Big|\mathbf{x}_0^{N_1}=\hat{\mathbf{x}}_0^{N_1},\mathbf{y}_0^{N_2}=\hat{\mathbf{y}}_0^{N_2}\Big]$$

At $t=0$, $\mathcal{M}_0^{N_1}=\hat{\mathcal{M}}_0^{N_1}$ and $\mathcal{N}_0^{N_2}=\hat{\mathcal{N}}_0^{N_2}$ as we begin the fully observable and estimated scenarios under the same initial joint states, i.e., $\mathbf{x}_0^{N_1}=\hat{\mathbf{x}}_0^{N_1},\mathbf{y}_0^{N_2}=\hat{\mathbf{y}}_0^{N_2}$. Thus, $a_0=0$. We proceed to show that $a_{t+1}\leq\kappa_1+\kappa_2 a_t$ where κ_1 and κ_2 are the same as Step 1. By the law of iterated expectations,

$$\begin{aligned}
&\mathbb{E}_{\phi^*,\psi^*}\Big[\mathrm{d_{TV}}\big(\mathcal{M}_{t+1}^{N_1},\hat{\mathcal{M}}_{t+1}^{N_1}\big)+\mathrm{d_{TV}}\big(\mathcal{N}_{t+1}^{N_2},\hat{\mathcal{N}}_{t+1}^{N_2}\big)\Big|\mathbf{x}_0^{N_1}=\hat{\mathbf{x}}_0^{N_1},\mathbf{y}_0^{N_2}=\hat{\mathbf{y}}_0^{N_2}\Big]\\
&=\mathbb{E}_{\phi^*,\psi^*}\left[\Big[\mathbb{E}_{\phi^*,\psi^*}\mathrm{d_{TV}}\big(\mathcal{M}_{t+1}^{N_1},\hat{\mathcal{M}}_{t+1}^{N_1}\big)+\mathrm{d_{TV}}\big(\mathcal{N}_{t+1}^{N_2},\hat{\mathcal{N}}_{t+1}^{N_2}\big)\Big|\mathbf{x}_t^{N_1},\hat{\mathbf{x}}_t^{N_1},\mathbf{y}_t^{N_2},\hat{\mathbf{y}}_t^{N_2}\Big]\Big|\mathbf{x}_0^{N_1}=\hat{\mathbf{x}}_0^{N_1},\mathbf{y}_0^{N_2}=\hat{\mathbf{y}}_0^{N_2}\right]\\
&<\mathbb{E}_{\phi^*,\psi^*}\left[\kappa_1+\kappa_2\left(\mathrm{d_{TV}}\big(\mathcal{M}_t^{N_1},\hat{\mathcal{M}}_t^{N_1}\big)+\mathrm{d_{TV}}\big(\mathcal{N}_t^{N_2},\hat{\mathcal{N}}_t^{N_2}\big)\right)|\mathbf{x}_0^{N_1}=\hat{\mathbf{x}}_0^{N_1},\mathbf{y}_0^{N_2}=\hat{\mathbf{y}}_0^{N_2}\right]\\
&=\kappa_1+\kappa_2\mathbb{E}_{\phi^*,\psi^*}\Big[\mathrm{d_{TV}}\big(\mathcal{M}_t^{N_1},\hat{\mathcal{M}}_t^{N_1}\big)+\mathrm{d_{TV}}\big(\mathcal{N}_t^{N_2},\hat{\mathcal{N}}_t^{N_2}\big)|\mathbf{x}_0^{N_1}=\hat{\mathbf{x}}_0^{N_1},\mathbf{y}_0^{N_2}=\hat{\mathbf{y}}_0^{N_2}\Big]\\
&=\kappa_1+\kappa_2 a_t,
\end{aligned}$$

where we use the result obtained from Step 1 along with the linearity property of expectations. Using this new notation, (I.48) simplifies to

$$\Delta J(\phi_t^*,\psi_t^*)\leq L_r\sum_{t=0}^{T}a_t$$

We have $a_0=0$, $a_1<\kappa_1$, $a_2<\kappa_1+\kappa_2 a_1$ or $a_2<\kappa_1(\kappa_2+1)$, $a_3<\kappa_2 a_2+\kappa_1$ or $a_3<\kappa_1(\kappa_2^2+\kappa_2+1)$ and so on. This can be written compactly for all $t>0$ as

$$a_t<\kappa_1\sum_{\tau=0}^{t-1}\kappa_2^\tau=\kappa_1\frac{\kappa_2^t-1}{\kappa_2-1},$$

as $\kappa_2 = (1 + L_f + \frac{1}{2}L_\phi + \frac{1}{2}L_\psi) \neq 1$. This is a geometric sum and thus we have

$$\begin{aligned}
\Delta J(\phi_t^*, \psi_t^*) &\leq L_r \sum_{t=0}^{T} a_t \\
&= L_r \sum_{t=1}^{T} a_t \\
&< L_r \sum_{t=1}^{T} \kappa_1 \frac{\kappa_2^t - 1}{\kappa_2 - 1} \\
&= \frac{L_r \kappa_1}{\kappa_2 - 1} \sum_{t=1}^{T} \kappa_2^t - 1 \\
&= \frac{L_r \kappa_1}{\kappa_2 - 1} \left(\kappa_2 \frac{\kappa_2^T - 1}{\kappa_2 - 1} - T \right) \\
&= \underbrace{\frac{1}{2} \frac{L_r (L_\phi + L_\psi)}{\kappa_2 - 1} \left(\kappa_2 \frac{\kappa_2^T - 1}{\kappa_2 - 1} - T \right)}_{K} \epsilon + \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big) \frac{L_r}{\kappa_2 - 1} \left(\kappa_2 \frac{\kappa_2^T - 1}{\kappa_2 - 1} - T \right) \\
&= K\epsilon + \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big).
\end{aligned}$$

□

J RPS and cRPS Setup

J.1 State Space

We have three states in this representation of the game: rock, paper and scissors. We denote this state space as $\mathcal{S} = \{\texttt{R}, \texttt{P}, \texttt{S}\}$. The empirical distribution of the Blue team is denoted by $\mu \in \mathcal{P}(\mathcal{S})$ and that of the Red team is denoted by $\nu \in \mathcal{P}(\mathcal{S})$. Since we have three states for each team, both EDs lie in a three-dimensional simplex denoted by $\mathcal{P}(\mathcal{S})$.

J.2 Action Space

J.2.1 RPS

At each state, we define three actions denoted by $\mathcal{A} = \{\texttt{CW}, \texttt{CCW}, \texttt{Stay}\}$. These actions represent the ability of the agents to move from one state to the other in the following fashion:

1. `CW` denotes a clockwise cyclic action from one state to the other, i.e., from $\texttt{R} \to \texttt{P}$, $\texttt{P} \to \texttt{S}$, $\texttt{S} \to \texttt{R}$.
2. `CCW` denotes a counterclockwise cyclic movement, i.e., from $\texttt{R} \to \texttt{S}$, $\texttt{S} \to \texttt{P}$, $\texttt{P} \to \texttt{R}$.
3. `Stay` denotes the idle action (remain in the same state as before).

J.2.2 cRPS

At each state we have two actions denoted by $\mathcal{A} = \{\texttt{CW}, \texttt{Stay}\}$. These actions represent the ability of the agents to move from one state to the other in the following fashion:

1. `CW` denotes a clockwise cyclic action from one state to the other, i.e., from $\texttt{R} \to \texttt{P}$, $\texttt{P} \to \texttt{S}$, $\texttt{S} \to \texttt{R}$.
2. `Stay` denotes the idle action (remain in the same state as before).

Thus, we cannot directly jump from `R` to `S` within a single step, but must go via `P`. Mathematically, `S` does not lie in the reachable set of `R`. The reachable set $\mathcal{R}(s)$ for each state s at a given time step under this modified action space is as follows

$$\mathcal{R}(\texttt{R}) = \{\texttt{R}, \texttt{P}\}, \quad \mathcal{R}(\texttt{P}) = \{\texttt{P}, \texttt{S}\}, \quad \mathcal{R}(\texttt{S}) = \{\texttt{S}, \texttt{R}\}.$$

The state-action space of both RPS and cRPS are presented in Figure J.1.

J.3 Dynamics and Transition Probabilities

For both RPS and cRPS, we consider deterministic transitions $\mathcal{T}(s, a, s')$, which implies that given a state-action pair (s, a), the agent reaches a unique next state s' with certainty (no distribution over the reachable states). Thus, for state $\texttt{R}$ and action $\texttt{Stay}$, the transition function $\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \{0, 1\}$ is given as:

$$\mathcal{T}(\texttt{R}, \texttt{Stay}, \texttt{R}) = 1, \quad \mathcal{T}(\texttt{R}, \texttt{Stay}, \texttt{P}) = 0, \quad \mathcal{T}(\texttt{R}, \texttt{Stay}, \texttt{S}) = 0.$$

This implies that an agent in the state $\texttt{R}$ upon taking the $\texttt{Stay}$ action remains in state $\texttt{R}$. Similarly,

$$\mathcal{T}(\texttt{R}, \texttt{CW}, \texttt{R}) = 0, \quad \mathcal{T}(\texttt{R}, \texttt{CW}, \texttt{P}) = 1, \quad \mathcal{T}(\texttt{R}, \texttt{CW}, \texttt{S}) = 0.$$

represent the method to transition from $\texttt{R}$ to $\texttt{P}$.

J.4 Reward Structure

In a two player RPS game, the reward matrix for Player 1 is defined as:

$$A = \begin{bmatrix} 0 & -1 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{bmatrix}.$$

We extend this two-player framework to the multi-agent team game formulation. Define the pairwise reward for an agent at state $x \in \mathcal{S}$ within the Blue team and at state $y \in \mathcal{S}$ from the Red team as

$$r(x, y) \triangleq A_{xy},$$

where A_{xy} represents the element from the reward matrix A corresponding to the states x (row player) and y (column player). In lieu of the zero-sum structure, the reward for the agent at y with respect to x becomes $-r(x, y)$. Thus, for each player $x_i \in \mathcal{S}$ and $i = 1, 2, \ldots, N_1$ in the Blue team and $y_j \in \mathcal{S}$ and $j = 1, 2, \ldots, N_2$ in the Red team, the reward for the Blue team can be defined as

$$R_{\text{Blue}}(\mathbf{x}, \mathbf{y}) = \frac{1}{N_1} \sum_{i=1}^{N_1} \underbrace{\left[\frac{1}{N_2} \sum_{j=1}^{N_2} r(x_i, y_j) \right]}_{\text{reward for agent } i}.$$

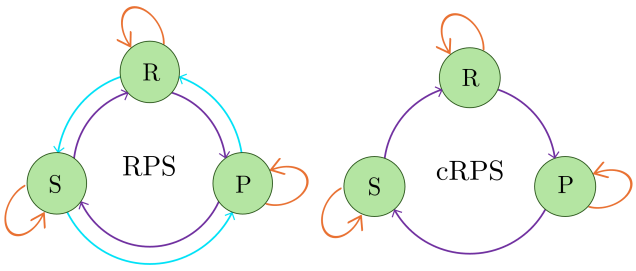

Figure J.1: State-action spaces for RPS and cRPS.

Rewriting the term inside the square brackets as

$$\begin{aligned}\frac{1}{N_2}\sum_{j=1}^{N_2} r(x_i, y_j) &= \frac{1}{N_2}\sum_{y\in\mathcal{S}}\sum_{j=1}^{N_2} r(x_i, y)\mathbf{1}_{y_j=y} \\ &= A_{x_i s_0}\sum_{j=1}^{N_2}\frac{1}{N_2}\mathbf{1}_{y_j=s_0} + A_{x_i s_1}\sum_{j=1}^{N_2}\frac{1}{N_2}\mathbf{1}_{y_j=s_1} \\ &+ A_{x_i s_2}\sum_{j=1}^{N_2}\frac{1}{N_2}\mathbf{1}_{y_j=s_2} \\ &= A_{x_i s_0}\nu(s_0) + A_{x_i s_1}\nu(s_1) + A_{x_i s_2}\nu(s_2) \\ &= A(x_i)\nu, \end{aligned} \tag{J.53}$$

where $A(x_i)$ is the row of the reward matrix corresponding to state x_i. Using (J.53), the total Blue reward can be expressed as

$$\begin{aligned} R_{\text{Blue}}(\mathbf{x}, \mathbf{y}) &= \frac{1}{N_1}\sum_{i=1}^{N_1} A(x_i)\nu \\ &= \Big(\frac{1}{N_1}\sum_{x\in\mathcal{S}}\sum_{i=1}^{N_1} A(x_i)\mathbf{1}_{x_i=x}\Big)\nu \\ &= \Big(A(s_0)\mu(s_0) + A(s_1)\mu(s_1) + A(s_2)\mu(s_2)\Big)\nu \\ &= \mu^\intercal A\nu. \end{aligned}$$

Proposition 6. *With initial conditions $\mu_0 = [1, 0, 0]^\intercal$ (equivalent to $[\mu(\texttt{R}), \mu(\texttt{P}), \mu(\texttt{S})]$) and $\nu_0 = [0, 1, 0]^\intercal$, all mean-field optimal trajectories satisfy $\mu_t^* = \nu_t^* = [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]^\intercal$ for all $t \geq 2$, and $\mu_1^* = [0, 1-\delta, \delta]^\intercal$ where $\delta \in [\frac{1}{3}, \frac{2}{3}]$ and $\nu_1^* = [0, \frac{2}{3}, \frac{1}{3}]^\intercal$. Furthermore, the unique game value is given by $-1/3$.*

Proof. For the constrained RPS game under the stated initial condition, we cannot obtain the target distribution $\begin{bmatrix}\frac{1}{3} & \frac{1}{3} & \frac{1}{3}\end{bmatrix}^\intercal$ after a single time step but this may be possible for $t \geq 2$. To this end, consider the following candidate trajectory respecting the transition dynamics:

$$\begin{aligned} \mu_0^\intercal &= [1 \quad 0 \quad 0]\,, \\ \mu_1^\intercal &= [1-x_1 \quad x_1 \quad 0]\,, \\ \mu_2^\intercal &= [1-x_1-x_2 \quad x_1+x_2-x_3 \quad x_3]\,, \\ \mu_t^\intercal &= \begin{bmatrix}\frac{1}{3} & \frac{1}{3} & \frac{1}{3}\end{bmatrix}, \quad \forall t > 2. \\ \nu_0^\intercal &= [0 \quad 1 \quad 0]\,, \\ \nu_1^\intercal &= [0 \quad 1-y_1 \quad y_1]\,, \\ \nu_2^\intercal &= [y_3 \quad 1-y_1-y_2 \quad y_1+y_2-y_3]\,, \\ \nu_t^\intercal &= \begin{bmatrix}\frac{1}{3} & \frac{1}{3} & \frac{1}{3}\end{bmatrix}, \quad \forall t > 2. \end{aligned}$$

In order to respect the simplex structure for μ_t and ν_t, we have the following constraints at all times:

$$0 \leq x_1, x_2, x_3 \leq 1, \quad x_2 \leq 1 - x_1, \quad x_3 \leq x_1.$$

Similarly,

$$0 \leq y_1, y_2, y_3 \leq 1, \quad y_2 \leq 1 - y_1, \quad y_3 \leq y_1.$$

For the distribution at $t = 2$ to be $\begin{bmatrix}\frac{1}{3} & \frac{1}{3} & \frac{1}{3}\end{bmatrix}^\intercal$ for both teams, we get the additional constraints

$$\begin{aligned} x_3 = y_3 &= \frac{1}{3}, \\ x_1 + x_2 - x_3 = y_1 + y_2 - y_3 &= \frac{1}{3}, \\ \Rightarrow x_1 + x_2 = y_1 + y_2 &= \frac{2}{3}, \end{aligned}$$

which implies that

$$x_1, y_1 \leq \frac{2}{3},$$

since $x_2, y_2 \geq 0$. The constraints now take the form

$$\frac{1}{3} \leq x_1 \leq \frac{2}{3}, \tag{J.54}$$

and similarly,

$$\frac{1}{3} \leq y_1 \leq \frac{2}{3}. \tag{J.55}$$

The objective function for cRPS is given by

$$J^{N,\phi,\psi}\big(\mu_0, \nu_0\big) = \mathbb{E}_{\phi,\psi}\Big[\sum_{t=1}^{T} \mu_t^\intercal A \nu_t \Big| \mu_0, \nu_0\Big], \tag{J.56}$$

which leads to the optimization problem

$$\max_{\phi^t} \min_{\psi^t} \quad J^{N_1,N_2,\phi^t,\psi^t}(\mu_0, \nu_0) = \mu_1^\intercal A \nu_1 \ . \tag{J.57}$$

Substituting $\mu_1^\intercal = [1 - x_1, x_1, 0]$ and $\nu_1^\intercal = [0, 1 - y_1, y_1]$ results in the following expression for the maximizing Blue team:

$$\begin{aligned}\max_{\phi^t} \quad J^{N_1,N_2,\phi^t,\psi^t}(\mu_0, \nu_0) &= x_1 + 2y_1 - 3x_1y_1 - 1 \\ &= x_1(1 - 3y_1) + (2y_1 - 1).\end{aligned}$$

Since this equation is linear in x_1, the solution to the maximization problem subject to the constraint (J.54) is

$$x_1 = \frac{1}{3}, \quad y_1 > \frac{1}{3}, \tag{J.58}$$

$$x_1 = \frac{2}{3}, \quad y_1 < \frac{1}{3}, \tag{J.59}$$

$$x_1 \in [\frac{1}{3}, \frac{2}{3}], \quad y_1 = \frac{1}{3}. \tag{J.60}$$

Following the same approach for the minimizing Red team, we get the following objective,

$$\begin{aligned}\min_{\psi^t} \quad J^{N_1,N_2,\phi^t,\psi^t}(\mu_0, \nu_0) &= x_1 + 2y_1 - 3x_1y_1 - 1 \\ &= y_1(2 - 3x_1) + (x_1 - 1),\end{aligned}$$

subject to the constraint (J.55), with the solution being:

$$y_1 = \frac{1}{3}, \quad x_1 < \frac{2}{3}, \tag{J.61}$$

$$y_1 = \frac{2}{3}, \quad x_1 > \frac{2}{3}, \tag{J.62}$$

$$y_1 \in [\frac{1}{3}, \frac{2}{3}], \quad x_1 = \frac{2}{3}. \tag{J.63}$$

The Constraint (J.54) ensures that (J.62) cannot hold, while constraint (J.55) similarly prevents (J.59) from holding.

Consider now the case when $y_1 > \frac{1}{3}$. From (J.58) it follows that $x_1 = \frac{1}{3}$. Conversely, if the Blue team commits to a distribution with $x_1 = \frac{1}{3}$, the Red team's best response given by (J.61) gives $y_1 = \frac{1}{3}$, resulting in an incentive for the Red team to deviate from $y_1 > \frac{1}{3}$. Thus, (J.58) does not

constitute an optimal solution. Following a similar argument, it can be shown that (J.63) is not an optimal solution either, as illustrated below.

Assume that $x_1 = \frac{2}{3}$. From (J.63), $y_1 \in [\frac{1}{3}, \frac{2}{3}]$. Now, if the Red team announces that it will deploy the distribution $y_1 \in [\frac{1}{3}, \frac{2}{3}]$, the Blue team's response for x_1 follows from (J.58) and (J.60). We have already established that (J.58) is not an optimal solution. This implies that $x_1 \in [\frac{1}{3}, \frac{2}{3}]$ can be a possible response to the Red team. However, it violates (J.63), where $x_1 = \frac{2}{3}$ follows from strict equality. Thus, (J.63) does not constitute an optimal solution as the Blue team has an incentive to deviate.

Now, suppose the Blue team announces a distribution where $x_1 \in [\frac{1}{3}, \frac{2}{3}]$. In this case, the Red team's optimal response, derived from (J.61) and (J.60), is $y_1 = \frac{1}{3}$. Conversely, if the Red team announces that its distribution will be $y_1 = \frac{1}{3}$, the Blue team will still follow $x_1 \in [\frac{1}{3}, \frac{2}{3}]$. Since neither team has an incentive to deviate from these distributions, they form an optimal trajectory. Thus, the solution to the bilinear optimization problem for two-time step convergence takes the form:

$$\mu_1^* = \begin{bmatrix} 1 - x_1 \\ x_1 \\ 0 \end{bmatrix} \quad \text{and} \quad \nu_1^* = \begin{bmatrix} 0 \\ \frac{2}{3} \\ \frac{1}{3} \end{bmatrix}, \tag{J.64}$$

such that $x_1 \in [\frac{1}{3}, \frac{2}{3}]$, leading to a game value of $-\frac{1}{3}$. This establishes the distribution at $t = 1$ and confirms the existence of a two-time step optimal trajectory, thereby proving the first part of the proposition.

Now note the following:

1. The original objective function (J.56) can be expressed in a bilinear form (similar to the expressions for μ_0, μ_1, μ_2 using x_1, x_2, x_3). This makes it concave in the first argument and convex in the second argument.
2. The mean-fields μ and ν lie on a simplex and are hence, compact and convex.

Thus, by the generalized version of von Neumann's minimax theorem [Sorin, 2002], we conclude that the game value is unique, proving the second part of the proposition [6]. □

J.5 Implementation Details and Hyperparameters

The state distributions are represented as arrays that are concatenated together to form the global observation. This becomes the input to the critic network, which consists of a single hidden layer of 64 neurons and two tanh activation functions. The output is a single value that is equal to the estimated value function. On the other hand, the actor-network consists of a single MLP layer of 64 neurons that is concatenated with the local agent observation. Additionally, the logits are converted to a probability distribution through a softmax layer. The dimension scales with $|\mathcal{A}|$. Both the actor and critic networks are initialized using orthogonal initialization [Huang et al., 2022].

The single-stage RPS game is trained for 5,000 time steps with the actor and critic learning rates set to 0.0005 and 0.001, respectively, which remain constant throughout training. The networks are updated using the ADAM optimizer [Kingma and Ba, 2017] every 50 time steps for 10 epochs and a PPO clip value of 0.1. The entropy is decayed from 0.01 to 0.001 geometrically. We use an episode length of 1 after which the rewards are bootstrapped.

Moreover, since we have a single "team" buffer and the input/output dimensions are small, we do not use a mini-batch based update. For cRPS we use an episode length of 10 after which the rewards are bootstrapped. cRPS is trained using 200,000 time steps (=20,000 episodes) and is updated every 100 time steps. The algorithm was trained on a single NVIDIA GeForce RTX 3070 GPU and the training times are given in Tables M.1 and 2.

[6] Note: The optimal infinite horizon trajectory itself need not be unique (we have shown that x_1 can take a range of values).

K Battlefield Setup

K.1 State and Action Space

We consider a large-scale two-team (Blue and Red) ZS-MFTG on an $n \times n$ grid world. The state of the ith 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 (μ) and Red (ν) 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. In the following subsections, we elaborate on the weakly coupled transition dynamics and reward structure introduced in the game, followed by a detailed discussion of the training procedure and network architecture for MF-MAPPO in this example.

K.2 Interaction Between Agents

The transitions between states for agents belonging to both teams are characterized by their dynamics. These dynamics are probabilistic and depend on interactions among agents and are weakly coupled through their mean-field distributions. The weak coupling dynamics is keeping in line with the assumption in Guan et al. [2024a].

An agent at a given grid cell can be deactivated by the opponent team with a nonzero probability if the empirical mean-field of the opponent team at the grid cell supersedes that of the agent's own team. Similarly, a deactivated agent can be revived if the empirical mean-field of the agent's team is greater than the opponent's. This is referred to as numerical advantage. The total transition probability from state (p, s) to state (p', s') by taking an action a is given by

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

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

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

and

$$\begin{aligned}
\mathbb{P}\big(1 \,|\, (p, 0), \mu, \nu\big) &= \text{clip}_{[0,1]}\left(\beta_x(\mu(p) - \nu(p))\right), \\
\mathbb{P}\big(0 \,|\, (p, 0), \mu, \nu\big) &= 1 - \mathbb{P}\big(1 \,|\, (p, 0), \mu, \nu\big),
\end{aligned}$$

where $\nu(p) - \mu(p)$ is the Red team's numerical advantage over the Blue team at p. Similarly, the Blue team's numerical advantage over Red is given by $\mu(p) - \nu(p)$. α_x and β_x are tuning parameters to control the Red team's deactivation power and Blue team's reactivation power respectively. The Red team, being the defending team, is given a slight advantage in terms of higher deactivation power. This enables the possibility of capturing Blue team agents. However, to avoid degeneracy, the Red team agents are not allowed to enter the target. For our experiments, we assume $\alpha_x = 15$, $\alpha_y = 5$, and $\beta_x = \beta_y = 0$.

K.3 Reward Structure

The team rewards only depend on the mean-fields of the two teams. For the battlefield scenario, the Blue team agents receive a positive reward corresponding to the fraction of agents that reach the target alive. This is a one-time reward that depends on the change in the fraction of the population of the agents at the target, i.e., if $\mu_t|_{\text{target}} = \mu_{t+1}|_{\text{target}}$, then the team does not receive any positive reward. Each agent in the team receives an identical "team reward." The reward function is mathematically formulated as

$$R_{\text{Blue},t+1}(\mu, \nu) = \kappa\, \Delta\mu_{t+1}|_{\text{target}},$$

where,

$$\Delta\mu_{t+1}|_{\text{target}} = \mu_{t+1}(p^x = \text{Target}, s^x = 1) - \mu_t(p^x = \text{Target}, s^x = 1).$$

We have chosen $\kappa = 100$ in our simulations (heavier emphasis on reaching the target). The Red team's reward is the negative of the Blue team since we have a zero-sum game. Each team aims to maximize its own expected reward.

K.4 Implementation and Hyperparameters

The state distribution for a grid world of size $n \times n$ is represented as a three-dimensional array of size $(2, n, n)$ for each team. The first layer depicts the mean-field of the agents over an $n \times n$ grid that are alive and active, while the second layer gives information about the team's deactivated population. Each team's distribution is then concatenated together to form the global observation. This becomes the common information that is the input to the critic network which in our case is of size $(4, n, n)$ as we have two teams. Both neural networks consist of two main parts: a convolutional block and a fully connected block.

For the critic, the first CNN layer is the input layer that takes the 4 channels and outputs 32 channels, with a kernel size of 3x3, stride of 1, and padding of 1. Followed by ReLU activation, we have a hidden layer that takes 32 channels and outputs 64 channels, with the same kernel size, stride, and padding. Lastly, after another ReLU activation, we have the output layer that takes 64 channels and outputs 64 channels, again with the same kernel size, stride, and padding. After another ReLU layer, the output of the CNN is passed through an MLP. Namely, a fully connected (dense) layer takes the flattened output of the convolutional block and reduces it to 128 units. Between the input and the output layers, we have a single tanh activation function.

On the other hand, the input to the actor-network is split into two CNN blocks: one to process the common information and one to process the local information. The local information channel is an array of size $(1, n, n)$ that locates the position of the agent with value +1 if it is active and -1 if it has been deactivated. This local information is passed through a single CNN layer that outputs 16 channels with a kernel size of 3x3, stride of 1, and padding of 1 while the common information is passed through two such layers with the output of 32 channels. Both outputs are then followed by a ReLU activation function and the latent representation of the common information combined with the local agent observation is then passed through an MLP architecture.

A fully connected (dense) layer takes the flattened output of the convolutional block and reduces it to 512 units. We have a single hidden layer that reduces the dimension further to 128 and then the output logits. The layers are separated by the tanh activation functions. Finally, the logits are converted to a probability distribution through a softmax layer. Both the actor and critic networks are initialized using orthogonal initialization [Huang et al., 2022]. The architectures of the shared-team actor and minimally-informed critic networks for this example are shown in Figures K.2 and K.3 respectively.

All maps are trained using a single NVIDIA GeForce RTX 3070 GPU. The actor and critic learning rates are set to 0.0005 and 0.001 and both decay geometrically by a factor of 0.999. The networks are updated using the ADAM optimizer [Kingma and Ba, 2017] with two mini-batches for 10 epochs and a PPO clip value of 0.1. The entropy coefficient is initialized to 0.01 and decays with a factor of 0.995.

Maps 1 and 2 which are 4×4 grid worlds are trained for 5×10^6 and 4.5×10^6 time steps, respectively, and in both cases, the episode length is 20 time steps and the update frequency is every 500 time steps. The total training period is about one day. On the other hand, Map 3 being 8×8 in dimension, has an episode length of 64, is trained for 9×10^6 time steps and its network is updated every 1,000 time steps. The total training period is approximately three days.

For evaluating D-PC, we employ MF-MAPPO to learn Lipschitz policies under a fully observable mean-field regime, and use these policies to evaluate the performance of our dynamic estimation algorithm, D-PC. We consider a time-invariant, edgeless visibility graph, wherein agents have access only to the mean-field distribution of their opponent corresponding to their current cell. For the communication graph, we assume a subgrid based topology. In the subgrid communication topology, an agent located at state x can exchange information with its neighboring states arranged in a 3×3 grid centered around x. The communication graph is naturally time-varying since the number of states occupied by the agents need not be fixed. In accordance with Assumption 3 and the theoretical

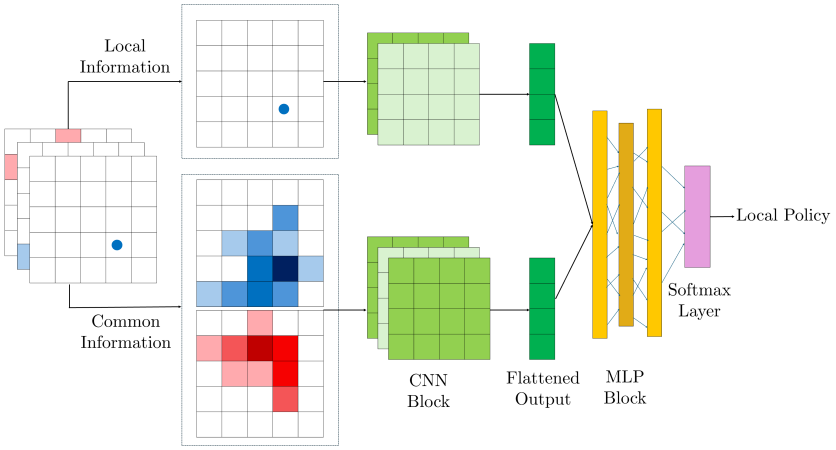

Figure K.2: MF-MAPPO: Shared-team actor for battlefield

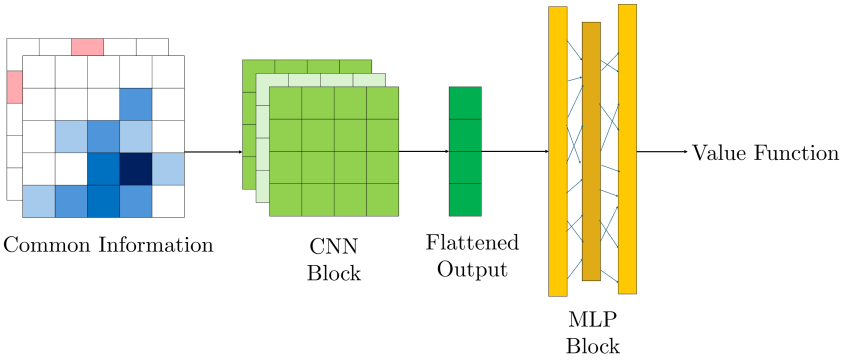

Figure K.3: MF-MAPPO: Minimally-informed critic for battlefield

requirements of both D-PC and benchmark algorithms, we explicitly impose the constraint that the communication graph is connected at all time steps.

At each time step t, the initial Red team's estimate for the D-PC algorithm by the Blue team (i.e., at $\tau = 0$) for every state x is taken as the projection of the estimate from the previous time step $t-1$ onto the constraint set $\mathcal{R}(x)$, that is,

$$\hat{\nu}_{x,t}^{N_2,\tau=0} = \Omega_{\mathcal{R}(x)}\left[\hat{\nu}_{x,t-1}^{N_2,R_{\text{com}}}\right], \quad x \in \mathcal{X}_t^o.$$

At $t = 0$ and $\tau = 0$, we assume $\hat{\nu}_{x,t=0}^{N_2,\tau=0} \sim \text{unif}\left(\mathcal{R}(x)\right)$ for all $x \in \mathcal{X}_t^o$.

L Epidemiology Setup

L.1 State and Action Space

We consider a large-scale two-team (Blue and Red) ZS-MFTG on an $n \times n$ grid world where the Blue team represents a population trying to prevent itself from a spreading virus, represented by the Red team. The state of the ith 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 unhealthy and 1 being healthy. The state of the Red agent as $y_i = p_i^y$, denoting the position of the Red agent. The state spaces for the Blue and Red teams are denoted by $\mathcal{X} = \mathcal{S}_{\text{position}} \times \mathcal{S}_{\text{status}}$ and $\mathcal{Y} = \mathcal{S}_{\text{position}}$, respectively. The mean-fields of the Blue (μ) and Red (ν) teams are distributions over the joint position and status space, i.e., $\mu \in \mathcal{P}(\mathcal{S}_{\text{position}} \times \mathcal{S}_{\text{status}})$ while $\nu \in \mathcal{P}(\mathcal{S}_{\text{position}})$. 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

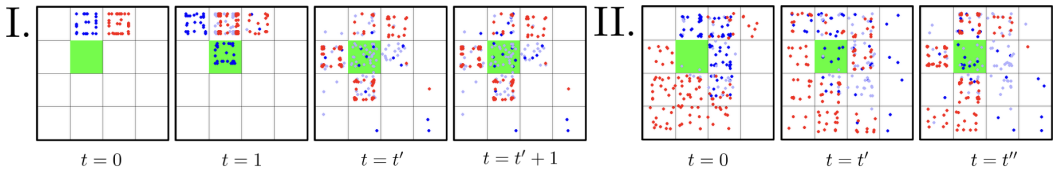

Figure L.4: Epidemiology via MF-MAPPO: In both cases, Red team concentrates around the hospital making it difficult for healing Blue team agents to decongest while some Blue agents successfully maintain distance.

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.

L.2 Interaction Between Agents and Reward Structure

We consider weakly coupled dynamics where the Blue team agent at a cell can be infected by a non-zero probability that is proportional to the Red team population as well as the infected Blue team population at that cell. For the Blue agents, the total transition probability from state (p, s) to state (p', s') by taking an action a is given by

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

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

$$\begin{aligned}\mathbb{P}\big(0 \mid (p, 1), \mu, \nu\big) &= \alpha(\nu(p) + \mu(p, s = 0))\\ \mathbb{P}\big(1 \mid (p, 0), \mu, \nu\big) &= \beta_x(\mu(p) - \nu(p)),\end{aligned}$$

where α is a tuning parameter to control the rate of infection. The Blue team agents can go to the hospital to heal and become healthy with probability β. The Red team on the other hand moves around the grid trying to infect as much of the population. However, to avoid degeneracy, the Red team agents are not allowed to enter the hospital. The reward depends on the proportion of healthy Blue agents, i.e., $\sum_{s=1} \mu(p, s)$, thereby incentivizing Blue agents to decongest and if infected, heal at the hospital and Red agents to surround the hospital, ready for contaminating the Blue team. Thus, this example, unlike the battlefield, exhibits asymmetry in the dynamics between the Red and Blue teams. For our experiments, we assume $\alpha = 15$, and $\beta = 5$.

L.3 Implementation and Hyperparameters

The MF-MAPPO architecture resembles the Battlefield networks introduced in K. However, the Red team mean-field has a size $(1, n, n)$, while Blue continues having $(2, n, n)$. The policy was trained using a single NVIDIA GeForce RTX 3070 GPU. Due to asymmetry in the dynamics, we separately tune the hyperparameters for both teams. For the Blue team the actor and critic learning rates are set to 0.0001 and 0.0005 with 3 epochs of training. On the other hand the Red team network learns with 0.0005 for the actor and 0.001 for the critic, with updates over 10 epochs. The entropy coefficient is initialized to 0.01 and both the entropy and the learning rates for both teams decay geometrically throughout the training period. The total training period is about half a day. While Table 3 quantitatively evaluates MF-MAPPO, a visualization of the team behaviors is presented in L.4.

M Additional Results

In this section, we present additional analyses and simulation results. In particular, we evaluate the exploitability of policies trained using MF-MAPPO and investigate the algorithm's sample efficiency in the battlefield scenario. Furthermore, we present results for the standard Rock-Paper-Scissors game and additional experiments on the zero-sum battlefield game, both using MF-MAPPO.

M.1 Exploitability Evaluation

For cRPS, the Blue and Red policies exhibit very low exploitability (0.002 and 0.004, respectively), closely matching the equilibrium performance of -0.331. Together with the learned mean-field trajectories matching the analytical equilibrium in Proposition 6, this provides empirical evidence that MF-MAPPO converges toward the equilibrium.

For Battlefield, the Blue policy has relatively low exploitability (4.735), while the Red policy is substantially more exploitable (26.114), compared with an equilibrium performance of 76.26. Although our training losses indicate convergence of the MF-MAPPO updates, this does not necessarily imply convergence to an equilibrium. One possible explanation is that, unlike cRPS, the Battlefield game may not admit a game value and therefore may not possess an equilibrium toward which MF-MAPPO can converge. In addition, the asymmetric structure of the environment where Red is prohibited from entering the target region may make the Red policy more constrained and sensitive to best-response exploitation. Approximation error, entropy regularization, and limited exploration may also contribute, and we therefore refrain from attributing the higher Red-team exploitability to any single cause, given the limited experiment we performed. Nevertheless, the Red policy consistently outperforms the baselines in cross-play (Table 3), providing complementary evidence that it is comparatively strong despite not being fully non-exploitable.

Establishing broader conditions (beyond independent-dynamics) for game-value existence in ZS-MFTGs and understanding practical learning convergence beyond the settings considered here are important directions for future work.

M.2 Rock-Paper-Scissors (RPS)

We first extend the two-player Rock-Paper-Scissors (RPS) game to a game played between two populations as described in Appendix J. The Nash equilibrium for this population-based RPS game is the uniform population distribution $[1/3, 1/3, 1/3]$ over the 3 states [Raghavan, 1994].

We compare MF-MAPPO with DDPG-MFTG [Shao et al., 2024a] based on the training time, average test rewards and attainment of the computed Nash distributions for $N_1 = N_2 = 1,000$ agents. We exclude MADDPG [Lowe et al., 2017] from our comparison, as it scales poorly to hundreds or thousands of agents due to its reliance on all agents' local and global observations and actions as inputs to its critic networks. We include the training curve for cRPS in Figure M.6 for reference.

From the learning curves in Figures M.5 and M.6 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. This corroborates with the results presented for cRPS (Table 2) in the main text. However, as shown in Table M.1, MF-MAPPO does take slightly longer to train since, unlike DDPG-MFTG, MF-MAPPO avoids mini-batch training, following Yu et al. [2022].

We tested the learned policy with a fixed initial distribution $\mu_{t=0} = [1, 0, 0]^\intercal$ and $\nu_{t=0} = [0, 1, 0]^\intercal$, and the resulting trajectories are visualized in Figure M.7. 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 M.7, DDPG-MFTG diverges from the equilibrium whereas MF-MAPPO converges immediately.

M.3 Battlefield

M.3.1 Validation Cases for MF-MAPPO

The following subsection qualitatively discusses the battlefield game for different map layouts. For these results, both teams are deploying policies trained using MF-MAPPO.

Table M.1: Performance comparison for RPS

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

Map A. The first map is a simple 4×4 grid world with a single target that we use to validate our algorithm. The target is partially blocked by an obstacle, see Figure M.8. For the initial condition in Figure M.8, the Blue team is initially split into two equal groups. The Blue team decides to merge the two sub-groups of agents into a single group. With this formation, the Red team has zero numerical advantage over the Blue team when they encounter in (g), resulting in all Blue agents safely arriving at the target. In comparison, if the two Blue subgroups do not merge but move toward the target one at a time, it will lead to 50% of the Blue team population being deactivated (first subgroup), followed by the remaining 50% (second subgroup). This demonstrates how the observation of mean-field distributions guides rational decision-making.

Map B. This map is identical to the one presented in Section 5. In the first scenario (Figure M.9), 70% of the Blue agents start at cell [2, 2], and 30% at cell [1, 1], while the Red team is evenly split between cells [0, 1] and [3, 3]. Half of the Red team at [0, 1] successfully blocks the 30% Blue agent group from entering the left corridor due to its numerical advantage, which forces the Blue agents to opt for the right corridor. At the same time, the larger Blue group with 70% of the population utilized their numerical advantage over the half Red team at the top right and deactivated all the Red agents as shown in (c) and reached the target at time step (d). This allowed the smaller 30% group to follow through the same corridor without losing agents. In the second scenario (Figure M.10), with the Red team evenly distributed at the corners, 30% of the Blue agents start at cell [2, 2] and 70% at cell [3, 1]. The Red team's numerical advantage at [3, 3] forces the Blue agents to move around and regroup (Figures M.10(b)-M.10(f)). Once united, the Blue team's numerical advantage forces the Red subgroup at [0,1] to disperse to avoid deactivation, allowing Blue to reach the target.

Map C. In Figure M.11, we deploy the MF-MAPPO policies trained with $N_1 = N_2 = 100$ (from the main text) to larger populations of $\bar{N}_1 = \bar{N}_2 = 1500$ and $\bar{N}_1 = \bar{N}_2 = 1000$ using the same initial condition as Figure 5I and II respectively. The teams continue to achieve their objectives as established in Theorem 3.3, exhibiting behaviors similar to $N_1 = N_2 = 100$ setting. Section 5 presented scenarios featuring structured initial configurations for both teams over an 8×8 grid. It is important to emphasize, however, that the algorithm is trained on a diverse set of initial conditions for a given map, ranging from agents concentrated within a few selected cells to agents distributed randomly across the grid world. The following examples demonstrate that the teams are able to accomplish their objectives even in scenarios where agents are dispersed across the environment

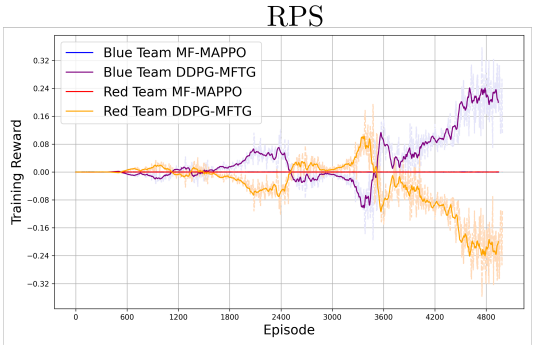

Figure M.5: Training curve for RPS.

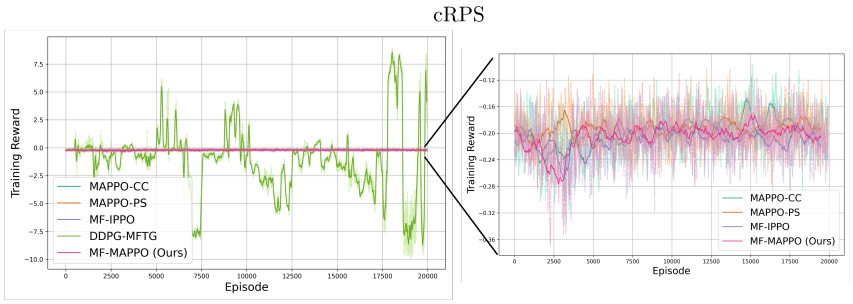

Figure M.6: Training curve for cRPS.

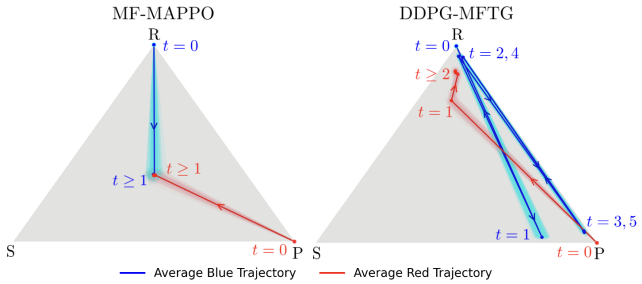

Figure M.7: ED trajectories induced by learned team policies on the state distribution simplex. Mean trajectories are averaged based on the 150 runs from fixed initialization $\mu_{t=0} = [1, 0, 0]^\intercal$ and $\nu_{t=0} = [0, 1, 0]^\intercal$; $N_1 = N_2 = 1,000$.

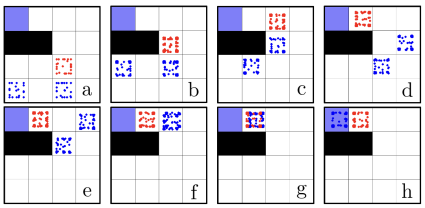

Figure M.8: Red is concentrated; Blue is evenly split.

rather than clustered into just 1-2 subgroups using the same policy from Section 5. In Figure M.12, the Blue team is initialized randomly, and local subgroups of agents emerge and coordinate to reach the target. This behavior is particularly pronounced near the upper target, where a greater numerical advantage facilitates successful coalition formation (Figures M.12(c)–(e)).

Turning to the randomly distributed Red team agents in Figure M.13, it is observed that they concentrate near the two target entrances and successfully neutralize most Blue team subgroups.

M.3.2 Efficiency of MF-MAPPO with Baselines

While Table 3 reports the performance of the learned policies when evaluated against each other, Table M.2 provides a complementary analysis of the sample efficiency of the baselines in the Battlefield environment. We define sample efficiency as the reward obtained per interaction with the environment:

$$\text{Sample Efficiency} = \frac{\text{Total rewards}}{\text{Total environment steps}}.$$

In addition, we report the average episode length, which serves as an indicator of how quickly episodes terminate. Shorter episodes suggest that the Blue agents increasingly learn to reach the target, while the Red agents simultaneously learn to defend the target and deactivate Blue agents—either outcome leading to episode termination.

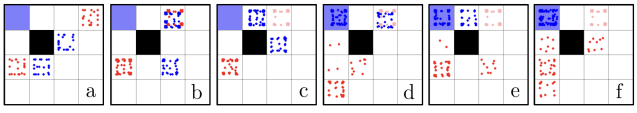

Figure M.9: Red is evenly split; 30% Blue are at $[1, 1]$ and 70% are at $[2, 2]$.

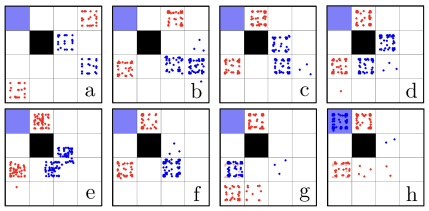

Figure M.10: Red is evenly split; 70% Blue are at $[3, 1]$ and 30% are at $[2, 2]$.

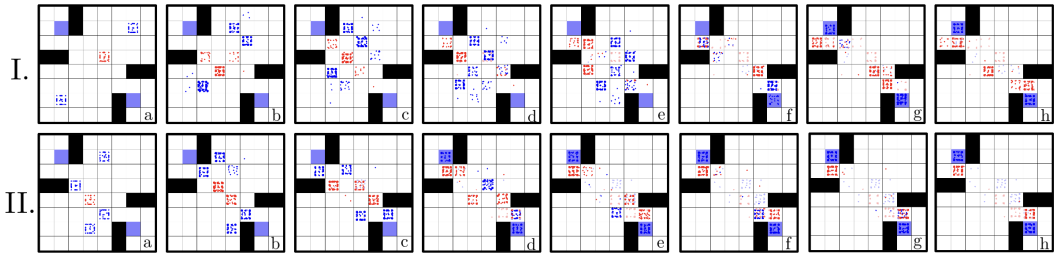

Figure M.11: I. 1500 agents in each team where Red is concentrated; 30% Blue are at the bottom, rest are at the top II. 1000 agents in each team with Blue evenly split; Red is concentrated.

Table M.2: Performance metrics for Battlefield

Algorithm	Critic Input	Sample Efficiency	Avg. Episode Length
MF-IPPO	$\{(x_{i,t}^{N_1}, \mu_t)\}_{i=1}^{N_1}$	6.3392	9.258
MAPPO-PS	$\{(x_{i,t}^{N_1}, \mu_t, \nu_t)\}_{i=1}^{N_1}$	5.4581	12.084
MAPPO-CC	$(\mathbf{x}_t^{N_1}, \mathbf{y}_t^{N_2}, \mu_t, \nu_t)$	5.7878	9.8808
DDPG-MFTG	(ϕ_t, μ_t, ν_t)	0.0548	20.0
MF-MAPPO	(μ_t, ν_t)	**14.6238**	**4.8448**

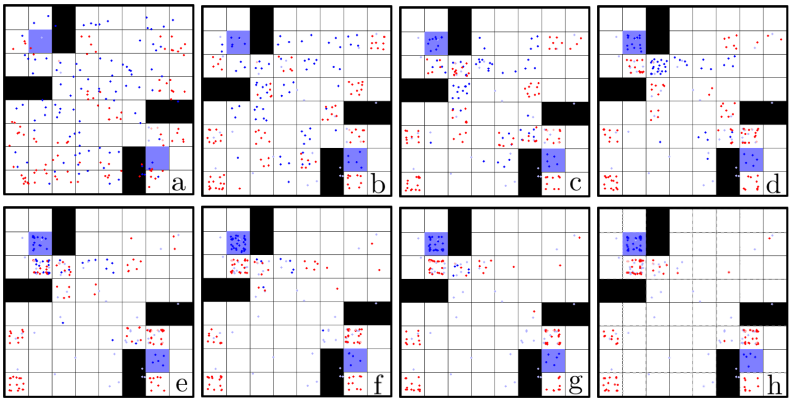

Figure M.12: Blue team is randomly spread around the map.

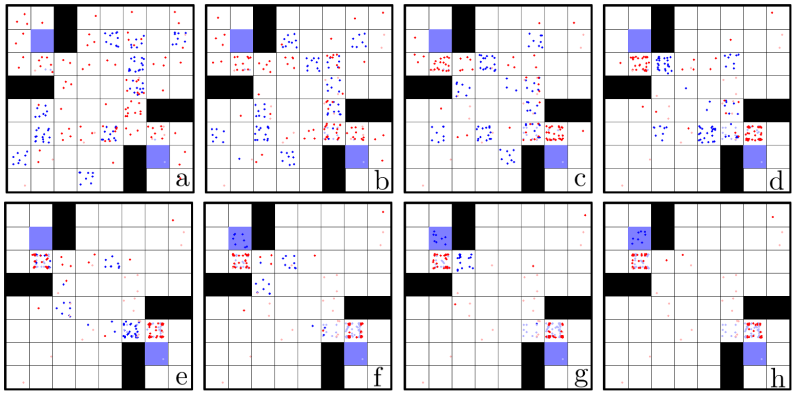

Figure M.13: Red team is randomly spread around the map.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and/or introduction (under "Our contributions", Section 1) explicitly state the claims made.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of this work in the conclusion (Section 6) and explicitly list key assumptions in the main text (Section 2).

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All theoretical results have been rigorously proven in the supplementary material and their implications and intuition have been provided in the main paper through experiments and general discussion (Sections 3 and 5).

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: All hyperparameters, neural network architectures and other information have been listed in the supplementary material, along with pseudocodes to the proposed algorithms.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide a github link to the repository containing the novel benchmark environments (`MFEnv`). Further, we provide detailed descriptions of the experimental setup, hyperparameters, and training procedures in the main text and supplementary material to facilitate reproducibility.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer:[Yes]

   Justification: We present the settings in the main text (Section 5) while the auxiliary details are deferred to the supplemental material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We present our results after running the proposed algorithm over multiple seeds and initial conditions to emphasize the statistical significance and generalizability of our results. The details are mentioned in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We discuss the details of the resources and computational complexity in the main text for the toy example (cRPS) in Section 5 and further elaborate it for the large scale battlefields and epidemiology examples in the supplemental material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The authors have reviewed the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

Justification: This paper presents work whose goal is to advance the field of game theory and reinforcement learning through foundational research.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper poses no such risks.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We have provided citations wherever necessary for the code packages.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes].

    Justification: We include scripts and a well documented repository for the newly introduced `MFEnv` benchmark.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components. The authors acknowledge the use of GPT-4 and GPT-5 for polishing the main text and for quick access to existing results in probability theory, in particular, for Theorem 3.2 and Proposition 2.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.