Zero-Sum Games between Mean-Field Teams: Reachability-based Analysis under Mean-Field Sharing

Yue Guan, Mohammad Afshari, Panagiotis Tsiotras

Georgia Institute of Technology
{yguan44, mafshari, tsiotras}@gatech.edu

Abstract

This work studies the behaviors of two large-population teams competing in a discrete environment. The team-level interactions are modeled as a zero-sum game while the agent dynamics within each team is formulated as a collaborative mean-field team problem. Drawing inspiration from the mean-field literature, we first approximate the large-population team game with its infinite-population limit. Subsequently, we introduce two fictitious coordinators and transform the infinite-population game to an equivalent zero-sum game between the two coordinators. Via a novel reachability analysis, we study the optimal coordination strategies, which induce decentralized strategies under the original information structure. We establish the ϵ-optimality of the resulting team strategies for the finite-population game, with the sub-optimality diminishing as the team size approaches infinity. The theoretical guarantees are verified by numerical examples.

Introduction

Multi-agent decision-making arises in many applications, ranging from warehouse robots (Li et al. 2021) to organizational economics (Gibbons, Roberts et al. 2013). While the majority of the literature formulates the problems within either the cooperative or competitive setting, results on mixed collaborative-competitive team behaviors are relatively sparse. In this work, we consider a competitive team game, where two teams, each comprising a large number of intelligent agents, compete at the team level, while agents collaborate within each team. Such hierarchical interactions hold significant relevance in domains such as military operations (Tyler et al. 2020) and other multi-agent systems operating in adversarial environments.

There are two major challenges when trying to solve such competitive team problems:

1. Large-population team problems are *computationally* challenging since the solution complexity increases exponentially with the number of agents, and, in general, the team optimal control problems belong to the NEXP complexity class (Bernstein et al. 2002).
2. Competitive team problems are *conceptually* challenging due to the elusive nature of the opponent team, and thus one cannot directly deploy approximation techniques available in the large-population game literature.

The scalability challenge in large-population multi-agent systems has been addressed for a specific class of games known as the mean-field games (Huang, Malhamé, and Caines 2006). The salient feature of a mean-field game is that agents are weakly-coupled in their dynamics and rewards through their state distribution (the so-called mean-field). The intractable interactions among agents can then be approximated as the interaction between a typical agent and the "mass" of infinitely many other agents. This approximation technique has been extended to *single*-team settings known as the mean-field team problem (Arabneydi and Mahajan 2014). A dynamic programming decomposition is developed for this problem, where all agents within the team deploy the same strategy prescribed by a fictitious coordinator. However, in the competitive team setting, although one may restrict the strategies used by his/her team to be identical, extending the same assumption to the opponent team may lead to a substantial underestimation of the opponent's capabilities and thus requires further justification.

Main Contributions

We address the two previous challenges by introducing a class of discrete zero-sum mean-field team games (ZS-MFTGs) as an extension to the mean-field team problems. Importantly, ZS-MFTG models competitive team behaviors and draws focus to the approximation of the opponent team strategies.

We develop a dynamic program that constructs ϵ-optimal strategies to the proposed large-population team problem. Notably, our approach finds an optimal solution at the infinite-population limit and considers only *identical* team strategies. This avoids both the so-called "curse of dimensionality" issue in multi-agent systems and the book-keeping of individual strategies. Our main results provide a sub-optimality bound on the exploitability for our proposed solution in the original finite-population game, even when the opponent's team strategy is non-identical. Specifically, we show that the sub-optimality decreases at the rate of $\mathcal{O}(\underline{N}^{-0.5})$, where $\underline{N}$ is the size of the smaller team.

Our results stem from a novel reachability-based analysis of the mean-field approximation. In particular, we show that any finite-population team behavior can be effectively

approximated by an infinite-population team that uses identical team strategies. This result allows us to approximate the original problem with two competing infinite-population teams and transform the resulting infinite-population problem into a zero-sum game between two *fictitious* coordinators. Such transformation leads to a dynamic program based on the common-information technique (Nayyar, Mahajan, and Teneketzis 2013) that efficiently constructs the optimal team strategies.

Related Literature

Mean-Field Games The mean-field game (MFG) model was introduced in (Huang, Caines, and Malhamé 2007; Lasry and Lions 2007) to address scalability issues in large-population games. The salient feature of MFG is that *selfish* agents are weakly-coupled in their dynamics and rewards through the mean-field (state distribution). If the population is sufficiently large, then an approximately optimal solution can be obtained by solving the infinite-population limit which is known as the mean-field equilibrium (MFE). See (Laurière et al. 2022) for an overview of the results in the MFG literature. The main differences between our setup and the MFG are the following: (a) we seek team optimal strategies while MFG seeks a Nash equilibrium. In particular, we provide performance guarantees when the entire opponent team deviates, while MFG only considers single-agent deviations; (b) The MFE assumes that all agents apply the same strategy and solves the mean-field flow offline. Hence, the MFE strategy is open-loop to the MF. However, under the ZS-MFTG setting, different opponent team strategies lead to different mean-field trajectories. Consequently, we require feedback on the MFs to respond to the strategies deployed by the opponent team.

Mean-Field Teams The single-team problem was explored in (Arabneydi and Mahajan 2014), where agents share a common team reward, resulting in a collaborative problem. The work of (Arabneydi and Mahajan 2015) directly assumes that all agents within the team apply the same strategy and the optimality for the finite-population game is *only* assured in the LQG setting (Mahajan and Nayyar 2015). Our work encompasses a more intricate two-team zero-sum scenario and justifies the identical team strategy assumptions.

The concurrent work of (Sanjari, Saldi, and Yüksel 2023) studies a similar team-against-team problem but in a continuous state and action setting. The authors analyze the existence of equilibria by modeling randomized strategies as Borel probability measures. Our work differs in the following aspects: (a) The work of Sanjari, Saldi, and Yüksel relies on the Kakutani fixed point theorem to establish the existence of a Nash equilibrium. In contrast, the discrete nature of our formulation renders the convexity of the best-response correspondence invalid, as exemplified in Numerical Example 1. Therefore, our approach focuses on the single-sided optimality based on the lower and upper game values; (b) our approach transforms the team-against-team problem into a zero-sum game between two coordinators, which allows the deployment of dynamic programming; (c) the analysis in (Sanjari, Saldi, and Yüksel 2023) primarily offers asymptotic performance guarantees. In contrast, our results, which incorporate reachability-analysis and additional Lipschitz assumptions, provide the convergence rate of the finite-population team performance to its infinite-population limit.

Notations

We use $[n]$ to denote $\{1, 2, \dots, n\}$. The indicator function is denoted as $\mathbb{1}_{\cdot}(\cdot)$, such that $\mathbb{1}_a(b) = 1$ if $a = b$ and 0 otherwise. We use uppercase letters to denote random variables (e.g., X and $\mathcal{M}$) and lowercase letters to denote their realizations (e.g., x and μ). For a finite set E, we denote the space of all probability measures over E as $\mathcal{P}(E)$.

Problem Formulation

Finite-Population Team Games

Consider a discrete-time system with two large teams of agents that operate over a finite horizon T. The Blue team consists of N_1 *homogeneous* agents, and the Red team consists of N_2 homogeneous agents. The total system size is denoted as $N = N_1 + N_2$, and $\rho = N_1/N$ reflects the size ratio between the two teams. Let $X_{i,t}^{N_1} \in \mathcal{X}$ and $U_{i,t}^{N_1} \in \mathcal{U}$ denote the random variables representing the state and action taken by Blue agent $i \in [N_1]$ at time t. Here, $\mathcal{X}$ and $\mathcal{U}$ are the *finite* individual state and action spaces for each Blue agent, independent of i and t. Similarly, we use $Y_{j,t}^{N_2} \in \mathcal{Y}$ and $V_{j,t}^{N_2} \in \mathcal{V}$ to denote the individual state and action of Red agent $j \in [N_2]$. The joint state and action of the Blue team and the Red team 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.

Definition 1. The *empirical distribution* (ED) for the Blue and Red teams are defined as

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

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

Note that $\mathcal{M}_t^{N_1} \in \mathcal{P}(\mathcal{X})$ and $\mathcal{N}_t^{N_2} \in \mathcal{P}(\mathcal{Y})$, and $\mathcal{M}_t^{N_1}(x)$ gives the fraction of Blue agents at state x. We use the following two operators to denote the computations in (1):

$$\mathcal{M}_t^{N_1} = \mathrm{Emp}_{\mu}(\mathbf{X}_t^{N_1}), \qquad \mathcal{N}_t^{N_2} = \mathrm{Emp}_{\nu}(\mathbf{Y}_t^{N_2}).$$

Note that the Emp operators remove agent index information and thus one cannot tell the state of a specific Blue agent i given only the Blue ED.

We use total variation to measure the distance between distributions. Formally, for a finite set E, the total variation between two probability measures $\mu, \mu' \in \mathcal{P}(E)$ is given by

$$\mathrm{d_{TV}}(\mu, \mu') = \frac{1}{2} \sum_{e \in E} |\mu(e) - \mu'(e)| = \frac{1}{2} \|\mu - \mu'\|_1 .$$

Dynamics We consider weakly-coupled dynamics, where the dynamics of each individual agent is coupled with other agents through the EDs. For Blue agent i, its stochastic transition is governed by the transition kernel f_t and satisfies

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

$$= f_t(x^{N_1}_{i,t+1} | x^{N_1}_{i,t}, u^{N_1}_{i,t}, \mu^{N_1}_t, \nu^{N_2}_t),$$

where $\mu^{N_1}_t = \mathrm{Emp}_\mu(\mathbf{x}^{N_1}_t)$ and $\nu^{N_2}_t = \mathrm{Emp}_\nu(\mathbf{y}^{N_2}_t)$. Similarly, the dynamics of Red agent j is governed by the transition kernel $g_t(y^{N_2}_{j,t+1} | y^{N_2}_{j,t}, v^{N_2}_{j,t}, \mu^{N_1}_t, \nu^{N_2}_t)$.

Assumption 1 (Lipschitz Dynamics)**.** For all $x \in \mathbf{x}$, $\mu, \mu' \in \mathcal{P}(\mathcal{X})$, $\nu, \nu' \in \mathcal{P}(\mathcal{Y})$ and $t \in \{0, \ldots, T-1\}$, there exist a positive constant L_f such that

$$\sum_{x' \in \mathcal{X}} |f_t(x'|x,u,\mu,\nu) - f_t(x'|x,u,\mu',\nu')| \leq L_f \Big(\mathrm{d_{TV}}(\mu,\mu') + \mathrm{d_{TV}}(\nu,\nu') \Big).$$

We assume that g_t is L_g-Lipschitz as well.

Reward Structure Under the team-game framework, agents in the same team share the same reward. Similar to the dynamics, we consider a weakly-coupled team reward

$$r_t : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y}) \to [-R_{\max}, R_{\max}].$$

Assumption 2 (Lipschitz Rewards)**.** For all $\mu, \mu' \in \mathcal{P}(\mathcal{X})$, $\nu, \nu' \in \mathcal{P}(\mathcal{Y})$ and $t \in \{0, \ldots, T\}$, there exists $L_r > 0$ such that

$$|r_t(\mu,\nu) - r_t(\mu',\nu')| \leq L_r \big(\mathrm{d_{TV}}(\mu,\mu') + \mathrm{d_{TV}}(\nu,\nu') \big).$$

Under the zero-sum structure, we let the Blue team maximize the reward while the Red team minimizes it.

Information Structure We assume a mean-field sharing information structure (Arabneydi and Mahajan 2015). Specifically, at each time step t, Blue agent i observes its own state $X^{N_1}_{i,t}$ and the EDs $\mathcal{M}^{N_1}_t$ and $\mathcal{N}^{N_2}_t$. Similarly, Red agent j observes $Y^{N_2}_{j,t}$, $\mathcal{M}^{N_1}_t$ and $\mathcal{N}^{N_2}_t$. We consider the following mixed Markov policies:

$$\phi_{i,t} : \mathcal{U} \times \mathcal{X} \times \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y}) \to [0,1],$$
$$\psi_{j,t} : \mathcal{V} \times \mathcal{Y} \times \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y}) \to [0,1],$$

where $\phi_{i,t}(u | X^{N_1}_{i,t}, \mathcal{M}^{N_1}_t, \mathcal{N}^{N_2}_t)$ is the probability that Blue agent i selects action u given its state $X^{N_1}_{i,t}$ and the team EDs $\mathcal{M}^{N_1}_t$ and $\mathcal{N}^{N_2}_t$. An individual strategy is defined as a time sequence $\phi_i = \{\phi_{i,t}\}_{t=0}^T$. A Blue team strategy $\phi^{N_1} = \{\phi_i\}_{i=1}^{N_1}$ is the collection of individual strategies used by each Blue agent. We use Φ_t and Φ to denote, respectively, the set of individual policies and strategies available to each Blue agent. The set of Blue team strategies is then defined as the Cartesian product $\Phi^{N_1} = \times_{i=1}^{N_1} \Phi$. The notations extend naturally to the Red team.

In summary, an instance of a finite-population zero-sum mean-field team game is defined as the tuple $\texttt{ZS-MFTG} = \langle \mathcal{X}, \mathcal{Y}, \mathcal{U}, \mathcal{V}, f_t, g_t, r_t, N_1, N_2, T \rangle$.

Optimization Problem The performance of the team strategy pair (ϕ^{N_1}, ψ^{N_2}) is given by the expected cumulative reward

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

where $\mathcal{M}^{N_1}_t = \mathrm{Emp}_\mu(\mathbf{X}^{N_1}_t)$ and $\mathcal{N}^{N_2}_t = \mathrm{Emp}_\nu(\mathbf{Y}^{N_2}_t)$, and the expectation is with respect to the distribution of all system variables induced by ϕ^{N_1} and ψ^{N_2}.

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

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

where $\underline{J}^{N*}$ is the lower game value for the finite-population game. Note that the game value may not always exist, i.e., max-min value may differ from the min-max value (Elliott and Kalton 1972). Consequently, we consider the following optimality condition for the Blue team strategy.

Definition 2. A Blue team strategy $\phi^{N_1 *}$ is ϵ-optimal if

$$\underline{J}^{N*} \geq \min_{\psi^{N_2} \in \Psi^{N_2}} J^{N,\phi^{N_1 *},\psi^{N_2}} \geq \underline{J}^{N*} - \epsilon.$$

The strategy $\phi^{N_1 *}$ is optimal if $\epsilon = 0$.

Similarly, the minimizing Red team considers a min-max optimization problem, which leads to the upper game value

$$\bar{J}^{N*} = \min_{\psi^{N_2} \in \Psi^{N_2}} \max_{\phi^{N_1} \in \Phi^{N_1}} J^{N,\phi^{N_1},\psi^{N_2}}.$$

The ϵ-optimality of Red team strategies is defined similarly.

A ZS-MFTG Example

Consider a simple team game on a two-node graph, where the Blue team aims to maximize its presence at node 2. The state spaces are given by $\mathcal{X} = \{x^1, x^2\}$ and $\mathcal{Y} = \{y^1, y^2\}$, and the action spaces are $\mathcal{U} = \{u^1, u^2\}$ and $\mathcal{V} = \{v^1, v^2\}$. The Blue action u^1 corresponds to staying on the current node and u^2 represents moving to the other node. The same connotations apply to Red actions v^1 and v^2. This scenario is visualized in the following figure.

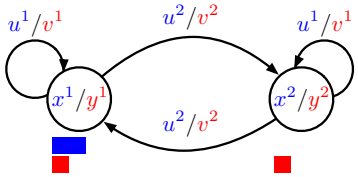


Figure 1: An example of ZS-MFTG over a two-node graph, where $N_1 = 2$, $N_2 = 2$ and $\rho = 0.5$.

The reward is $r_t(\mu,\nu) = \mu(x^2)$, which incentivizes Blue team agents to concentrate at node 2. An example of the Blue transition kernel at x^1 under u^2 can be

$$f_t(x^1|x^1,u^2,\mu,\nu) = 0.5\big(1 - \big(\rho\mu(x^1) - (1-\rho)\nu(y^1)\big)\big),$$
$$f_t(x^2|x^1,u^2,\mu,\nu) = 0.5\big(1 + \big(\rho\mu(x^1) - (1-\rho)\nu(y^1)\big)\big).$$

Under this transition kernel, the probability of a Blue agent transitioning from node 1 to node 2 depends on the Blue team's numerical advantage over the Red team at node 1.

The initial joint states depicted in Figure 1 are given by $\mathbf{x}^2_0 = [x^1, x^1]$ and $\mathbf{y}^2_0 = [y^1, y^2]$. The corresponding EDs are $\mu^2_0 = [1, 0]$, $\nu^2_0 = [0.5, 0.5]$, and the running reward is $r_0 = 0$. Suppose the Blue team applies a team strategy

such that $\phi_0^i(u^2|x^1,\mu_0^2,\nu_0^2)=1$ for both $i\in[2]$. The probability of an individual Blue agent transitioning to node 2 is 0.625. Thus, the next Blue ED is a random vector with three possible realizations: (i) $\mathcal{M}_1^2=[1,0]$ with probability 0.14 (both Blue agents remain on node 1); (ii) $\mathcal{M}_1^2=[0.5,0.5]$ with probability 0.47 (one moves and one remains); and (iii) $\mathcal{M}_1^2=[0,1]$ with probability 0.39 (both move). Suppose the game terminates at $T=1$, then the value under ϕ^2 is given by $J^{4,\phi^2,\psi^2}(\mathbf{x}_0^2,\mathbf{y}_0^2)=0+(0.14\cdot 0+0.47\cdot 0.5+0.39\cdot 1)=0.63$.

Infinite-Population Team Game

The preceding max-min optimization in (2) is intractable for large-population systems since the dimension of the joint policy spaces Φ^{N_1} and Ψ^{N_2} grows exponentially with the number of the agents. To address this scalability issue, we consider the infinite-population limit of the ZS-MFTGs, and further assume that agents in the same infinite-population team deploy the same strategy. As a result, we can model the behavior of an entire team as the distribution of a *typical agent*, i.e., the mean-field (Lasry and Lions 2007).

We first introduce the class of identical team strategies.

Definition 3 (Identical Blue Team Strategy)**.** The Blue team strategy $\phi^{N_1}=\{\phi_1,\dots,\phi_{N_1}\}$ is an identical team strategy, if $\phi_{i_1,t}=\phi_{i_2,t}$ for all $i_1,i_2\in[N_1]$ and $t\in\{0,1,\dots,T{-}1\}$.

When all Blue agents apply the same individual strategy ϕ, we slightly abuse the notation and use ϕ to denote the identical Blue team strategy. Consequently, we use Φ to denote both the set of Blue individual strategies and the set of identical Blue team strategies. The definitions and notations extend to the identical Red team strategies.

We define the mean-field (MF) as the state distribution of a typical agent in an infinite-population team game.

Definition 4. Given identical team strategies $\phi\in\Phi$ and $\psi\in\Psi$, the MFs propagate according to the following *deterministic* dynamics with (μ_0^ρ,ν_0^ρ) as initial conditions:

$$\mu_{t+1}^\rho(x')=\sum_{x\in\mathcal{X}}\Big[\sum_{u\in\mathcal{U}}f_t(x'|x,u,\mu_t^\rho,\nu_t^\rho)\phi_t(u|x,\mu_t^\rho,\nu_t^\rho)\Big]\mu_t^\rho(x),$$
$$\nu_{t+1}^\rho(y')=\sum_{y\in\mathcal{Y}}\Big[\sum_{v\in\mathcal{V}}g_t(y'|y,v,\mu_t^\rho,\nu_t^\rho)\psi_t(v|y,\mu_t^\rho,\nu_t^\rho)\Big]\nu_t^\rho(y).$$

For simplicity, we express the above MF dynamics in a compact matrix form as

$$\begin{aligned}\mu_{t+1}^\rho&=\mu_t^\rho F_t(\mu_t^\rho,\nu_t^\rho,\phi_t),\\ \nu_{t+1}^\rho&=\nu_t^\rho G_t(\mu_t^\rho,\nu_t^\rho,\psi_t),\end{aligned}\tag{3}$$

where $F_t\in\mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|}$ is the transition matrix for a typical Blue agent under ϕ_t, which can be computed based on the transition kernel f_t. The matrix G_t is defined similarly.

Consider the infinite-population limit of the example in Figure 1 with $\mu_0^{0.5}=[1,0]$, $\nu_0^{0.5}=[0.5,0.5]$ and $\rho=0.5$. If the Blue team applies the identical team strategy $\phi_0(u^2|x^1,\mu_0^{0.5},\nu_0^{0.5})=1$, then the next Blue MF is deterministically given by $\mu_1^{0.5}=[0.375,0.625]$.

Later, in Theorem 2 and Lemma 1 we will show that the *deterministic* MF above is an approximation of the (stochastic) finite-population ED, and the approximation error goes to zero when $N_1,N_2\to\infty$. Thus, we can regard the mean-field as the empirical distribution of an infinite-population team. On the other hand, Theorem 2 justifies the identical team strategy assumption we made when constructing the infinite-population game.

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

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

where the propagation of μ_t^ρ and ν_t^ρ are subject to (3).

The worst-case performance of the maximizing Blue team is then given by the lower game value

$$\underline{J}^{\rho*}(\mu_0^\rho,\nu_0^\rho)=\max_{\phi\in\Phi}\min_{\psi\in\Psi}J^{\rho,\phi,\psi}(\mu_0^\rho,\nu_0^\rho).\tag{4}$$

Remark 1. As shown in Numerical Example 1 below, the max-min and min-max coordinator game value can differ, since the team best-response correspondence may not be convex-valued (Owen 2013). In the extended version of this work (Guan, Afshari, and Tsiotras 2023), we show that a Nash equilibrium exists when the agents' dynamics are completely decoupled from each other.

Remark 2. Different from the infinite-population game value in (4), the finite-population value in (2) takes joint states as arguments rather than the EDs. The difference comes from the non-identical strategies considered in the finite-population game, which require agents' index information to sample actions and predict the game's evolution.

Reachable Sets

Due to the deterministic dynamics in (3), designing the policies ϕ_t and ψ_t at time t is equivalent to selecting the next desirable MFs. Consequently, we examine the set of MFs that can be reached at the next time step. We use $\pi_t:\mathcal{U}\times\mathcal{X}\to[0,1]$ to denote a local Blue policy, which is *open-loop* with respect to the MFs. Specifically, $\pi_t(u|x)$ is the probability that the typical Blue agent selects action u at state x regardless of the current MFs. The set of open-loop Blue local policies is denoted as Π_t. Similarly, $\sigma_t:\mathcal{V}\times\mathcal{Y}\to[0,1]$ and Σ_t denote a Red local policy and its admissible set. Under the local policy π_t, the Blue MF propagates as

$$\mu_{t+1}^\rho(x')=\sum_{x\in\mathcal{X}}\Big[\sum_{u\in\mathcal{U}}f_t(x'|x,u,\mu_t^\rho,\nu_t^\rho)\pi_t(u|x)\Big]\mu_t^\rho(x),$$

and the Red team MF dynamics under Red local policies is defined similarly.

The reachable sets are then defined as the collection of all MFs that can be reached using a local policy at the next step.

Definition 5. The Blue team reachable set, starting from μ_t^ρ and ν_t^ρ, is defined as

$$\mathcal{R}_{\mu,t}(\mu_t^\rho,\nu_t^\rho)\triangleq\{\mu_{t+1}^\rho|\exists\pi_t\in\Pi_t\text{ s.t.}\,\mu_{t+1}^\rho=\mu_t^\rho F_t(\mu_t^\rho,\nu_t^\rho,\pi_t)\}.$$

Similarly, the Red team reachable set is defined as

$$\mathcal{R}_{\nu,t}(\mu_t^\rho,\nu_t^\rho)\triangleq\{\nu_{t+1}^\rho|\exists\sigma_t\in\Sigma_t\text{ s.t.}\,\nu_{t+1}^\rho=\nu_t^\rho G_t(\mu_t^\rho,\nu_t^\rho,\sigma_t)\}.$$

Later on, we regard the reachable sets as correspondences, i.e., set-valued functions (Freeman and Kokotovic 2008).

Zero-Sum Game Between Coordinators

To obtain a dynamic program that effectively solves (4), we construct a fictitious centralized coordinated system (Mahajan and Nayyar 2015) for the *infinite-population* game with a Blue and a Red team coordinator. At time t, the Blue coordinator observes the MFs of both teams and prescribes a local policy $\pi_t \in \Pi_t$ to all agents within its team. The local policy is selected according to:

$$\pi_t = \alpha_t\big(\mu_t^\rho, \nu_t^\rho\big),$$

where $\alpha_t : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y}) \to \Pi_t$ is a deterministic Blue *coordination policy*, and $\pi_t(u_t|x_t) \triangleq \alpha_t(\mu_t^\rho, \nu_t^\rho)(u_t|x_t)$ gives the probability that a Blue agent selects action u_t at state x_t. Similarly, the Red coordinator observes the MFs and selects a local policy $\sigma_t \in \Sigma_t$ according to $\sigma_t = \beta_t\big(\mu_t^\rho, \nu_t^\rho\big)$.

We refer to the time sequence $\alpha = \big(\alpha_1, \dots, \alpha_T\big)$ as the *coordination strategy* for the Blue team and $\beta = \big(\beta_1, \dots, \beta_T\big)$ as the Red team coordination strategy. The sets of admissible coordination strategies are denoted as $\mathcal{A}$ and $\mathcal{B}$.

Remark 3. There is a one-to-one correspondence between the Blue (Red) coordination strategies and the identical Blue (Red) team strategies such that

$$\phi_t(u|x,\mu,\nu) = \underbrace{\alpha_t(\mu,\nu)}_{\pi_t}(u|x).$$

The equivalent centralized system can be viewed as a zero-sum game played between the two coordinators, where the game state is the joint MF (μ_t^ρ, ν_t^ρ) that follows the dynamics in (3), and the actions are the local policies π_t and σ_t selected by the coordinators. Formally, the zero-sum coordinator game is defined as the tuple $\texttt{ZS-CG} = \langle \mathcal{P}(\mathcal{X}), \mathcal{P}(Y), \Pi_t, \Sigma_t, F_t, G_t, r_t, \rho, T\rangle$, where both the state and action spaces are continuous.

Coordinator Game Values

Similar to the standard two-player zero-sum games, we use a dynamic programming backward recursion scheme to find the lower value of the coordinator game. The lower value at the terminal time T is given by $\underline{J}^{\rho*}_{\mathrm{cor},T}(\mu_T^\rho, \nu_T^\rho) = r_T(\mu_T^\rho, \nu_T^\rho)$. For all previous time steps $t = 0, \dots, T-1$, the two coordinators optimize their cumulative reward functions by choosing their actions (i.e., local policies) π_t and σ_t. Consequently, we have

$$\underline{J}^{\rho*}_{\mathrm{cor},t}(\mu_t^\rho, \nu_t^\rho) = r_t(\mu_t^\rho, \nu_t^\rho) \tag{5}$$
$$+ \max_{\pi_t \in \Pi_t} \min_{\sigma_t \in \Sigma_t} \underline{J}^{\rho*}_{\mathrm{cor},t+1}\big(\mu_t^\rho F_t(\mu_t^\rho, \nu_t^\rho, \pi_t), \nu_t^\rho G_t(\mu_t^\rho, \nu_t^\rho, \sigma_t)\big).$$

With the optimal value function, the optimal Blue team coordination policy can then be easily constructed via

$$\alpha_t^*(\mu_t^\rho, \nu_t^\rho) \in \tag{6}$$
$$\operatorname*{argmax}_{\pi_t \in \Pi_t} \min_{\sigma_t \in \Sigma_t} \underline{J}^{\rho*}_{\mathrm{cor},t+1}\big(\mu_t^\rho F_t(\mu_t^\rho, \nu_t^\rho, \pi_t), \nu_t^\rho G_t(\mu_t^\rho, \nu_t^\rho, \sigma_t)\big).$$

Exploiting the deterministic mean-field dynamics, we can change the optimization domains in (5) from the policy spaces to the corresponding reachable sets as follows

$$\underline{J}^{\rho*}_{\mathrm{cor},t}(\mu_t^\rho, \nu_t^\rho) = r_t(\mu_t^\rho, \nu_t^\rho) \tag{7}$$
$$+ \max_{\mu_{t+1}^\rho \in \mathcal{R}_{\mu,t}(\mu_t^\rho, \nu_t^\rho)} \min_{\nu_{t+1}^\rho \in \mathcal{R}_{\nu,t}(\mu_t^\rho, \nu_t^\rho)} \underline{J}^{\rho*}_{\mathrm{cor},t+1}(\mu_{t+1}^\rho, \nu_{t+1}^\rho).$$

We can then employ a dynamic programming scheme to solve the previous equation backward in time, starting from $\underline{J}^{\rho*}_{\mathrm{cor},T}(\mu_T^\rho, \nu_T^\rho) = r_T(\mu_T^\rho, \nu_T^\rho)$. Later on, we primarily work with the reachability-based optimization in (7). There are two advantages to such an approach: First, the reachable sets generally have a lower dimension than the coordinator action spaces[1], which is desirable for numerical algorithms, and; Second, the reachability-based optimization allows us to apply the forthcoming Theorem 2 and study the performance loss due to the identical-strategy assumption introduced by the mean-field approximation.

Lipschitz Continuity of the Value Functions

We examine the continuity of the coordinator game values, which is essential for the performance guarantees. We start with the continuity of the reachability correspondences under the Hausdorff distance $\mathrm{dist_H}$.[2]

Lemma 1. *For all $\mu_t, \mu_t' \in \mathcal{P}(\mathcal{X})$ and $\nu_t, \nu_t' \in \mathcal{P}(\mathcal{Y})$, the reachability correspondence $\mathcal{R}_{\mu,t}$ satisfies*

$$\mathrm{dist_H}(\mathcal{R}_{\mu,t}(\mu_t,\nu_t), \mathcal{R}_{\mu,t}(\mu_t',\nu_t')) \tag{8}$$
$$\leq L_{R_\mu}\big(\mathrm{d_{TV}}\big(\mu_t,\mu_t'\big) + \mathrm{d_{TV}}\big(\nu_t,\nu_t'\big)\big),$$

where the Lipschitz constant is given by $L_{R_\mu} = 1 + \frac{1}{2}L_f$. The Red reachability correspondence satisfies a similar inequality with a Lipschitz constant $L_{R_\nu} = 1 + \frac{1}{2}L_g$.

Leveraging the continuity of the reachability correspondences, the following theorem establishes the Lipschitz continuity of the optimal coordinator game value.

Theorem 1. *For all $\mu_t^\rho, \mu_t^{\rho\prime} \in \mathcal{P}(\mathcal{X})$ and $\nu_t^\rho, \nu_t^{\rho\prime} \in \mathcal{P}(\mathcal{Y})$, the lower coordinator game value satisfies*

$$\big|\underline{J}^{\rho*}_{\mathrm{cor},t}(\mu_t^\rho, \nu_t^\rho) - \underline{J}^{\rho*}_{\mathrm{cor},t}(\mu_t^{\rho\prime}, \nu_t^{\rho\prime})\big| \tag{9}$$
$$\leq L_{J,t}\big(\mathrm{d_{TV}}\big(\mu_t^\rho,\mu_t^{\rho\prime}\big) + \mathrm{d_{TV}}\big(\mu_t^\rho,\nu_t^{\rho\prime}\big)\big),$$

where the Lipschitz constant is given by $L_{J,t} = L_r\big(1 + L_{\mathcal{R}}(1 - L_{\mathcal{R}}^{T-t})/(1 - L_{\mathcal{R}})\big)$ and $L_{\mathcal{R}} = L_{R_\mu} + L_{R_\nu}$.

Proof. Observe that the lower value in (7) takes the form: $f(x,y) = c(x,y) + \max_{p\in\Gamma(x,y)} \min_{q\in\Theta(x,y)} g(p,q)$, which is an extension of the maximization marginal function (Freeman and Kokotovic 2008) to the max-min case. We present a continuity result for this type of marginal function in the extended version of this paper (Guan, Afshari, and Tsiotras 2023), based on which we can prove the above theorem through an inductive argument. □

Main Results

Recall that the optimal Blue team coordination strategy α^* is constructed for the infinite-population game assuming identical team strategies. This section establishes the performance guarantees for α^* in the finite-population games where both teams are allowed to deploy non-identical strategies.

[1] The Blue reachable set is a subset of $\mathcal{P}(\mathcal{X})$, while the Blue coordinator action space is given by $(\mathcal{P}(\mathcal{U}))^{|\mathcal{X}|}$.

[2] The Hausdorff distance between sets $A, B \subseteq \mathcal{X}$ is defined as $\mathrm{dist_H}(A,B) = \max\big\{\sup_{a\in A}\inf_{b\in B}\|a-b\|, \sup_{b\in B}\inf_{a\in A}\|a-b\|\big\}$.

Approximation Error

As α^* is solved at the infinite-population limit, it is essential to understand how well the infinite-population game approximates the original finite-population problem. The following theorem states that the reachable set constructed using identical strategies is rich enough to approximate the empirical distributions induced by *non-identical* team strategies in finite-population games.

Theorem 2. *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 team ED induced by a (potentially non-identical) Blue team policy $\phi_t^{N_1} \in \Phi_t^{N_1}$ as $\mathcal{M}_{t+1}^{N_1}$. Then, there exists $\mu_{t+1} \in \mathcal{R}_{\mu,t}(\mathcal{M}_t^{N_1}, \mathcal{N}_t^{N_2})$ such that*

$$\mathbb{E}_{\phi_t^{N_1}}\Big[\mathrm{d_{TV}}\big(\mathcal{M}_{t+1}^{N_1}, \mu_{t+1}\big)\big|\mathbf{X}_t^{N_1}, \mathbf{Y}_t^{N_2}\Big] \leq \frac{|\mathcal{X}|}{2}\sqrt{\frac{1}{N_1}}. \quad (10)$$

Proof. The key step is to construct an identical local policy $\pi_{\mathrm{apprx},t}$ that has its action distribution matching the average of the policies used by the Blue agents at each state. One can then leverage $\pi_{\mathrm{apprx},t}$ to mimic the population behavior and use a modified law of large numbers to show that the MF induced by $\pi_{\mathrm{apprx},t}$ satisfies the error bound in (10). This idea is visualized in Figure 2. A detailed proof is presented in the extended version (Guan, Afshari, and Tsiotras 2023). □

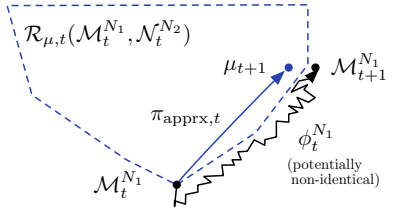


Figure 2: An illustration of the key idea behind Theorem 2.

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:*

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

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

Performance Guarantees

The following main theorem compares the worst-case performance of the identical Blue team strategy induced by α^* (Remark 3) to the original max-min optimization in (2), where non-identical strategies are allowed.

Theorem 3. *The optimal Blue coordination strategy α^* in* (6) *induces an ϵ-optimal Blue team strategy. Formally, for all $\mathbf{x}^{N_1} \in \mathcal{X}^{N_1}$ and $\mathbf{y}^{N_2} \in \mathcal{Y}^{N_2}$,*

$$\underline{J}^{N*}(\mathbf{x}^{N_1}, \mathbf{y}^{N_2}) \geq \min_{\psi^{N_2} \in \Psi^{N_2}} J^{N,\alpha^*,\psi^{N_2}}(\mathbf{x}^{N_1}, \mathbf{y}^{N_2}) \quad (11)$$
$$\geq \underline{J}^{N*}(\mathbf{x}^{N_1}, \mathbf{y}^{N_2}) - \mathcal{O}\Big(\frac{1}{\sqrt{\underline{N}}}\Big),$$

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

Proof. The first inequality is straightforward. We break the second inequality into two steps: (i) $\min_{\psi^{N_2}} J^{N,\alpha^*,\psi^{N_2}} \geq \underline{J}^{\rho*}_{\mathrm{cor}} - \mathcal{O}(1/\sqrt{\underline{N}})$; and (ii) $\underline{J}^{\rho*}_{\mathrm{cor}} \geq J^{N*} - \mathcal{O}(1/\sqrt{\underline{N}})$. The proofs for both steps are constructed based on induction, and we only present the proof for step (i) in the appendix of this paper. A detailed proof of Theorem 3 is presented in the extended version (Guan, Afshari, and Tsiotras 2023). □

Remark 4. Step (i) suggests that the Red team's performance improvement, when employing non-identical team strategies against α^*, is limited to $\mathcal{O}(1/\sqrt{\underline{N}})$. Step (ii) suggests a similar result for the Blue team.

Remark 5. The continuity of the coordinator game value (Theorem 1) is essential in the proof of Theorem 3, as bridges the mean-field approximation error (Theorem 2) and the performance loss. Hence, Assumptions 1 and 2 are necessary to translate the infinite-population performance back to the finite-population game. See the appendix of (Guan, Afshari, and Tsiotras 2023) for a discontinuous example where the infinite-population game value is different from that of the finite-population counterpart.

Numerical Examples

In this section, we provide two numerical examples. For both examples, the state spaces are $\mathcal{X} = \{x^1, x^2\}$ and $\mathcal{Y} = \{y^1, y^2\}$, and the action spaces are $\mathcal{U} = \{u^1, u^2\}$ and $\mathcal{V} = \{v^1, v^2\}$. The two-state state spaces allow the *joint* MFs to be characterized solely by $\mu_t(x^1)$ and $\nu_t(y^1)$.

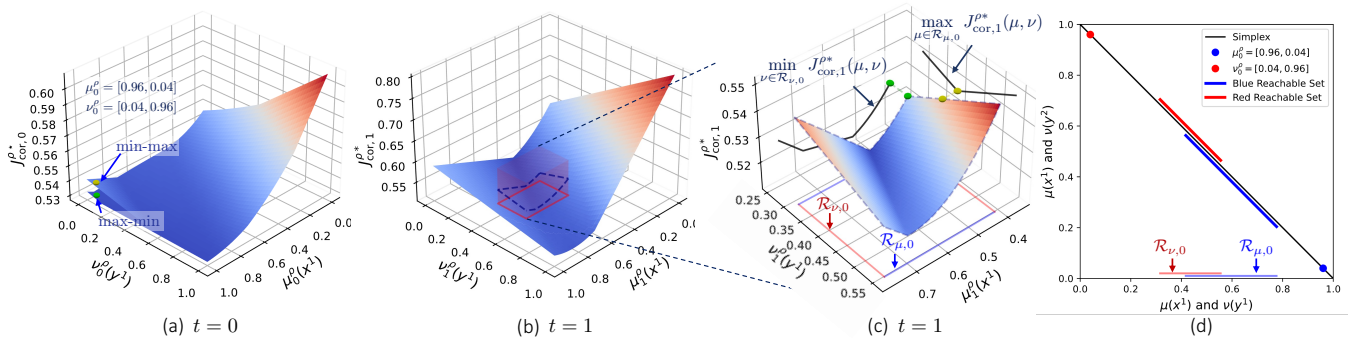
Numerical Example 1

This example serves to illustrate the reachability-based optimization in equation (7) and to demonstrate that the coordinator game value may not exist, contrary to the continuous setting as discussed in (Sanjari, Saldi, and Yüksel 2023). For a comprehensive description of the dynamics and reward setup, see (Guan, Afshari, and Tsiotras 2023).

The coordinator game values in Figure 3 are computed through discretization, where the two-dimensional simplexes $\mathcal{P}(\mathcal{X})$ and $\mathcal{P}(\mathcal{Y})$ are meshed into 1,000 bins.[3]. Note that the value function $J^{\rho*}_{\mathrm{cor},1}$ in subplot (b) is not convex-concave, which implies that the upper (max-min) and lower (min-max) game values at the previous step $t=0$ may differ, as observed in subplot (a). Specifically, at $\mu_0^\rho = [0.96, 0.04]$ and $\nu_0^\rho = [0.04, 0.96]$, we have the lower value $\underline{J}^{\rho*}_{\mathrm{cor},0} = 0.5298$ and the upper value $\bar{J}^{\rho*}_{\mathrm{cor},0} = 0.5384$, which are visualized as the green and yellow points. This discrepancy in the game values implies the absence of a Nash equilibrium in this coordinator game.

Based on (7), the optimization domains of $J^{\rho*}_{\mathrm{cor},0}$ are $\mathcal{R}_{\mu,0}(\mu_0^\rho, \nu_0^\rho)$ for the maximization and $\mathcal{R}_{\nu,0}(\mu_0^\rho, \nu_0^\rho)$ for the minimization, both of which are plotted in (d). Since $\mathcal{R}_{\mu,0}$ lives in a two-dimensional simplex, it is fully characterized by the range of $\mu(x^1)$, which enables us to visualize the optimization domain as the box in (c). Subplot (c) presents a

[3] An error bound on the difference between the discretized value and the true optimal value can be readily derived using the Lipschitz constants of the coordinator game values.



(a) $t=0$ (b) $t=1$ (c) $t=1$ (d)

Figure 3: Subplots (a)-(c) present the game values computed via discretization. The x- and y-axes correspond to $\mu_t^\rho(x^1)$ and $\nu_t^\rho(y^1)$, respectively. Subplot (d) illustrates the reachable sets starting from $\mu_0 = [0.96, 0.04]$ and $\nu_0 = [0.04, 0.96]$.

zoom-in for the optimization $\max_{\mathcal{R}_{\mu,0}} \min_{\mathcal{R}_{\nu,0}} J^{\rho*}_{\text{cor},1}$ and its min-max counterpart. The marginal functions are also plotted, from which the max-min and min-max values at $t=0$ can be directly obtained.

Numerical Example 2

It is generally challenging to verify the suboptimality bound in Theorem 3, since computing the true optimal performance of a finite-population team game is intractable. However, for the following designed example, we have the optimal team strategies for large finite-population teams.

Consider a ZS-MFTG with $T=2$. The game setup is similar to the one in Figure 1 but with different dynamics and rewards. The (minimizing) Red team's objective is to maximize its presence at state y^1 at $t=2$, which translates to the following reward function.

$$r_0(\mu,\nu) = r_1(\mu,\nu) = 0, \qquad r_2(\mu,\nu) = -\nu(y^1).$$

The Blue transition is time-invariant, deterministic, and independent of the MFs. Formally, for all μ, ν and $t \in \{0,1\}$,

$$\begin{aligned} &f_t(x^1|x^1,u^1,\mu,\nu)=1, \quad f_t(x^2|x^1,u^2,\mu,\nu)=1,\\ &f_t(x^2|x^2,u^1,\mu,\nu)=1, \quad f_t(x^1|x^2,u^2,\mu,\nu)=1. \end{aligned} \tag{12}$$

At $t=0$, all Red agents are frozen at both states, i.e., no action can change a Red agent's state. At $t=1$, Red agents at y^1 are frozen, and Red agents at y^2 can use v^2 to transition to y^1 and the transition probability is given by

$$\begin{aligned} &g_1^\rho(y^1|y^2,v^2,\mu_1,\nu_1) \qquad\qquad (13)\\ &= \min\left\{5\Big((\mu_1(x^1)-\frac{1}{\sqrt{2}})^2+(\mu_1(x^2)-(1-\frac{1}{\sqrt{2}}))^2\Big),1\right\}. \end{aligned}$$

Note that, under the above dynamics, if the Blue team achieves the target distribution $\mu_1 = [1/\sqrt{2}, 1-1/\sqrt{2}]$, no Red agent can transition from y^2 to y^1.

Infinite-population case. For the Red team, only the decisions of Red agents at y^2 at time $t=1$ have an impact on the game outcome. As a result, the above setup leads to a dominant optimal Red team strategy: all Red team agents at y^2 use v^2 at $t=1$ to transit to state y^1. On the other hand, the Blue team should try to match the distribution $\mu_1 = [1/\sqrt{2}, 1-1/\sqrt{2}]$ to minimize the probability of Red team agents transitioning from y^2 to y^1 at $t=1$. The dynamics in (12) ensures that the Blue team reachable set covers the entire simplex $\mathcal{P}(\mathcal{X})$ regardless of the initial distributions. Hence, the target distribution can always be achieved at $t=1$ with an *infinite* population.

Under the optimal strategies discussed above, the Blue team completely blocks the Red team agents' migration from y^2 to y^1, and thus only the Red agents spawn on y^1 will count towards the terminal rewards. Consequently, the infinite-population game value is given by $J^{\rho*}(\mu_0,\nu_0) = -\nu_0(y^1)$.

Finite-population case. The Red team's optimal strategy remains the same as the infinite-population case. Note that this Red team strategy is optimal against all Blue team strategies. The Blue team, on the other hand, cannot achieve the *irrational* target distribution with a finite number of agents. While the Blue team can still match the target distribution *in expectation* using a stochastic identical team strategy, the following analysis shows that a non-identical deterministic team strategy achieves a better performance.

Consider a Blue team with three agents and $\mu_0^3 = [1,0]$. The optimal Blue coordination strategy prescribes that all Blue agents pick u^1 ("stay") with probability $1/\sqrt{2}$ and u^2 ("move to x^2") with probability $(1-1/\sqrt{2})$ to reach the target distribution in expectation. Such action selection leads to the following four possible outcomes of the next Blue team ED μ_1^3: $\mathbb{P}([1,0]) = 0.354$, $\mathbb{P}([2/3,1/3]) = 0.439$, $\mathbb{P}([1/3,2/3]) = 0.182$, and $\mathbb{P}([0,1]) = 0.025$. In expectation, these empirical distributions lead to a transition probability of 0.518 for a Red team agent moving from y^2 to y^1. Consequently, we have the worst-case performance of the optimal Blue coordinator strategy as $\min_{\psi^{N_2}} J^{3,\alpha^*,\psi^{N_2}}(\mu_0,\nu_0) = -\nu_0(y^1) - 0.518\nu_0(y^2)$.

Next, consider the non-identical deterministic Blue team strategy, such that Blue team agents 1 and 2 apply action u^1 and Blue team agent 3 applies u^2. This Blue team strategy deterministically leads to $\mu_1^3 = [2/3, 1/3]$ at $t=1$, and the resultant Red team transition probability from y^2 to y^1 is 0.016. Clearly, the non-identical Blue team strategy significantly outperforms the identical mixed team strategy in this *three-agent* case. Furthermore, this Blue team strategy is

optimal over the entire non-identical Blue team strategy set, resulting in a finite-population game value $J^{3*}(\mu_0, \nu_0) = -\nu_0(y^1) - 0.016\nu_0(y^2)$.

The red line in Figure 4 plots the suboptimality of the coordinator strategy, which verifies the $\mathcal{O}(1/\sqrt{N})$ decrease rate predicted by Theorem 3.

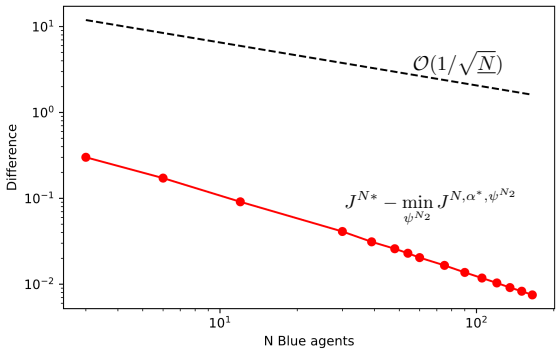


Figure 4: Performance loss of the optimal Blue coordination strategy with $\mu_0 = [1, 0]$ and $\nu_0 = [0.4, 0.6]$.

Conclusion

In this work, we introduced a discrete zero-sum mean-field team game to model the behaviors of competing large-population teams. We developed a dynamic programming approach that approximately solves this large-population game at its infinite-population limit where only identical team strategies are considered. Our analysis demonstrated that the identical strategies constructed are ϵ-optimal within the general class of non-identical team strategies when deployed in the original finite-population game. The derived performance guarantees are verified through numerical examples. Future work will investigate the LQG setup of this problem and explore machine-learning techniques to address more complex zero-sum mean-field team problems. Additionally, we aim to generalize our results to the infinite-horizon discounted problems, the general-sum case, and problems with heterogeneous teams.

Appendix

Proof of Theorem 3. We only show the proof for the first step here, i.e., $\min_{\psi^{N_2}} J^{N,\alpha^*,\psi^{N_2}} \geq \underline{J}^{\rho*}_{\text{cor}} - \mathcal{O}(1/\sqrt{N})$.

The proof for part (i) is constructed based on induction. Fix an arbitrary Red team strategy $\psi^{N_2} \in \Psi^{N_2}$.

Base case: At time T, we have for all $\mathbf{x}_T^{N_1}$ and $\mathbf{y}_T^{N_2}$ that

$$J_T^{N,\alpha^*,\psi^{N_2}}(\mathbf{x}_T^{N_1}, \mathbf{y}_T^{N_2}) = \underline{J}^{\rho*}_{\text{cor},T}(\mu_T^{N_1}, \nu_T^{N_1}) = r_T(\mu_T^{N_1}, \nu_T^{N_2}),$$

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

Inductive hypothesis: Assume that for all $\mathbf{x}_{t+1}^{N_1}$ and $\mathbf{y}_{t+1}^{N_2}$,

$$J_{t+1}^{N,\alpha^*,\psi^{N_2}}(\mathbf{x}_{t+1}^{N_1}, \mathbf{y}_{t+1}^{N_2}) \geq \underline{J}^{\rho*}_{\text{cor},t+1}(\mu_{t+1}^{N_1}, \nu_{t+1}^{N_2}) - \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big).$$

Induction: At timestep t, consider arbitrary joint states $(\mathbf{x}_t^{N_1}, \mathbf{y}_t^{N_2})$ and the corresponding EDs $(\mu_t^{N_1}, \nu_t^{N_2})$. Define

$$\mu_{t+1}^* = \mu_t^{N_1} F_t(\mu_t^{N_1}, \nu_t^{N_2}, \alpha_t^*).$$

Note that, from the optimality of α_t^* in (6), we have

$$\mu_{t+1}^* \in \underset{\mu_{t+1} \in \mathcal{R}_{\mu,t}(\mu_t^{N_1}, \nu_t^{N_2})}{\text{argmax}} \ \min_{\nu_{t+1} \in \mathcal{R}_{\nu,t}(\mu_t^{N_1}, \nu_t^{N_2})} \underline{J}^{\rho*}_{\text{cor},t+1}(\mu_{t+1}, \nu_{t+1}). \tag{14}$$

Furthermore, from Theorem 2, there exists a $\nu_{\text{apprx},t+1} \in \mathcal{R}_{\nu,t}(\mu_t^{N_1}, \nu_t^{N_2})$ for the fixed Red policy $\psi_t^{N_2}$ such that

$$\mathbb{E}_{\psi_t^{N_2}}\Big[\mathrm{d}_{\text{TV}}\big(\mathcal{N}_{t+1}^{N_2}, \nu_{\text{apprx},t+1}\big)\Big] \leq \frac{|\mathcal{Y}|}{2}\sqrt{\frac{1}{N_2}}. \tag{15}$$

Then, for all $\mathbf{x}_t^{N_1} \in \mathcal{X}^{N_1}$ and $\mathbf{y}_t^{N_2} \in \mathcal{Y}^{N_2}$, we have

$$\begin{aligned}
&J_t^{N,\alpha^*,\psi^{N_2}}(\mathbf{x}_t^{N_1}, \mathbf{y}_t^{N_2}) \\
&= r_t(\mu_t^{N_1}, \nu_t^{N_2}) + \mathbb{E}_{\alpha^*,\psi^{N_2}}\Big[J_{t+1}^{N,\alpha^*,\psi^{N_2}}(\mathbf{X}_{t+1}^{N_1}, \mathbf{Y}_{t+1}^{N_2})\Big] \\
&\overset{\text{(i)}}{\geq} r_t(\mu_t^{N_1}, \nu_t^{N_2}) + \mathbb{E}_{\alpha^*,\psi^{N_2}}\Big[\underline{J}^{\rho*}_{\text{cor},t+1}(\mathcal{M}_{t+1}^{N_1}, \mathcal{N}_{t+1}^{N_2})\Big] - \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big) \\
&= r_t(\mu_t^{N_1}, \nu_t^{N_2}) - \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big) + \mathbb{E}_{\alpha^*,\psi^{N_2}}\Big[\underline{J}^{\rho*}_{\text{cor},t+1}(\mathcal{M}_{t+1}^{N_1}, \mathcal{N}_{t+1}^{N_2}) \\
&\qquad - \underline{J}^{\rho*}_{\text{cor},t+1}(\mu_{t+1}^*, \nu_{\text{apprx},t+1}) + \underline{J}^{\rho*}_{\text{cor},t+1}(\mu_{t+1}^*, \nu_{\text{apprx},t+1})\Big] \\
&\overset{\text{(ii)}}{\geq} r_t(\mu_t^{N_1}, \nu_t^{N_2}) + \underline{J}^{\rho*}_{\text{cor},t+1}(\mu_{t+1}^*, \nu_{\text{apprx},t+1}) - \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big) \\
&\quad - L_{J,t+1}\Big(\underbrace{\mathbb{E}_{\alpha^*}[\mathrm{d}_{\text{TV}}(\mathcal{M}_{t+1}^{N_1}, \mu_{t+1}^*)]}_{\mathcal{O}(1/\sqrt{N_1}) \text{ due to Lemma 1}} + \underbrace{\mathbb{E}_{\psi^{N_2}}[\mathrm{d}_{\text{TV}}(\mathcal{N}_{t+1}^{N_2}, \nu_{\text{apprx},t+1})]}_{\mathcal{O}(1/\sqrt{N_2}) \text{ due to (15)}}\Big) \\
&\overset{\text{(iii)}}{\geq} r_t(\mu_t^{N_1}, \nu_t^{N_2}) + \underline{J}^{\rho*}_{\text{cor},t+1}(\mu_{t+1}^*, \nu_{\text{apprx},t+1}) - \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big) \\
&\overset{\text{(iv)}}{\geq} r_t(\mu_t^{N_1}, \nu_t^{N_2}) + \min_{\nu_{t+1} \in \mathcal{R}_{\nu,t}(\mu_t^{N_1}, \nu_t^{N_2})} \underline{J}^{\rho*}_{\text{cor},t+1}(\mu_{t+1}^*, \nu_{t+1}) - \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big) \\
&\overset{\text{(v)}}{=} \underline{J}^{\rho*}_{\text{cor},t}(\mu_t^{N_1}, \nu_t^{N_2}) - \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big).
\end{aligned}$$

Inequality (i) is due to the inductive hypothesis; inequality (ii) leverages the Lipschitz continuity of the coordinator value function (Theorem 1); inequality (iii) bounds the error terms using Theorem 2 and Lemma 1; inequality (iv) is due to the fact that $\nu_{\text{apprx},t+1}$ is in the reachable set; equality (v) comes from the optimality of μ_{t+1}^* in (14).

Finally, since $\psi^{N_2} \in \Psi^{N_2}$ is arbitrary, we have

$$\min_{\psi^{N_2} \in \psi^{N_2}} J_0^{N,\alpha^*,\psi^{N_2}}(\mathbf{x}^{N_1}, \mathbf{y}^{N_2}) \geq \underline{J}^{\rho*}_{\text{cor}}(\mu^{N_1}, \nu^{N_2}) - \mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big),$$

thus completing the proof. □

References

Arabneydi, J.; and Mahajan, A. 2014. Team optimal control of coupled subsystems with mean-field sharing. In *53rd IEEE Conference on Decision and Control*, 1669–1674. Los Angeles, CA.

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

Bernstein, D. S.; Givan, R.; Immerman, N.; and Zilberstein, S. 2002. The complexity of decentralized control of Markov decision processes. *Mathematics of Operations Research*, 27(4): 819–840.

Elliott, R. J.; and Kalton, N. J. 1972. *The Existence of Value in Differential Games*, volume 126. American Mathematical Soc.

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

Gibbons, R.; Roberts, J.; et al. 2013. *The Handbook of Organizational Economics*. Princeton University Press Princeton, NJ.

Guan, Y.; Afshari, M.; and Tsiotras, P. 2023. Zero-Sum Games between Mean-Field Teams: A Common Information and Reachability based Analysis. *arXiv preprint arXiv:2303.12243*.

Huang, M.; Caines, P. E.; and Malhamé, R. P. 2007. Large-population cost-coupled LQG problems with nonuniform agents: individual-mass behavior and decentralized ϵ-Nash equilibria. *IEEE Transactions on Automatic Control*, 52(9): 1560–1571.

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

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

Laurière, M.; Perrin, S.; Geist, M.; and Pietquin, O. 2022. Learning Mean Field Games: A Survey. *arXiv preprint arXiv:2205.12944*.

Li, J.; Tinka, A.; Kiesel, S.; Durham, J. W.; Kumar, T. S.; and Koenig, S. 2021. Lifelong multi-agent path finding in large-scale warehouses. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 11272–11281.

Mahajan, A.; and Nayyar, A. 2015. Sufficient Statistics for Linear Control Strategies in Decentralized Systems With Partial History Sharing. *IEEE Transactions on Automatic Control*, 60(8): 2046–2056.

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

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

Sanjari, S.; Saldi, N.; and Yüksel, S. 2023. Nash equilibria for exchangeable team against team games and their mean field limit. In *American Control Conference*, 1104–1109. San Diego, CA.

Tyler, J.; Arnold, R.; Abruzzo, B.; and Korpela, C. 2020. Autonomous Teammates for Squad Tactics. In *International Conference on Unmanned Aircraft Systems*, 1667–1672. Athens, Greece, Sept. 1–4, 2020.