Distributed Team Orchestration via Supervisor Networks: Convergence, Optimality, and Resilience

Juntian Zhu, Guanpu Chen, Tongtian Zhu, Miguel de Carvalho, Zhouwang Yang, and Fengxiang He

***Abstract*—In this paper, we study zero-sum potential team games with a supervisor network, where agents rely on supervisor-provided belief information rather than accurate common beliefs. The main challenge is that such belief information can be inaccurate because of supervisors' belief-estimation errors and the misreporting of joint actions by Byzantine teams. We propose the distributed team-orchestrating algorithm (DTOA), which combines team fictitious play with supervisor-based distributed belief learning. We prove the convergence of supervisors' belief estimates and establish that the induced learning dynamics converge to a near team-Nash equilibrium (TNE) in terms of the team-Nash gap (TNG). In the Byzantine setting, we consider a misreporting attack model and develop a Byzantine-resilient DTOA. We further provide probabilistic guarantees for Byzantine-team identification and establish an asymptotic bound on the honest TNG. Numerical experiments illustrate the theoretical findings, compare DTOA with baseline learning methods, and evaluate its performance in a Markov decision process setting.**

***Index Terms*—Team game, distributed learning, algorithmic convergence, equilibrium, Byzantine resilience**

I. Introduction

With the rapid development of artificial intelligence (AI), decision-making problems in multi-agent systems have received increasing attention. A large body of work has investigated cooperative and competitive interactions in multi-agent systems, which can be broadly classified into three categories: purely cooperative single-team problems, single-team problems involving external adversaries, and multi-team games. Single-team problems, with or without external adversaries, consider settings in which all agents cooperate to achieve a common objective [1–4]. However, many complex decision-making environments involve multiple teams whose members cooperate internally while competing strategically with other teams, and such settings are naturally modeled by multi-team games [5–7].

Existing studies develop theoretical and algorithmic tools for multi-team interactions from several perspectives. Feng et al. [8] study a two-team mean-field game that incorporates both within-team cooperation and inter-team competition. Other studies consider coalition games in which each fixed coalition is treated as a team and analyze the corresponding equilibrium properties [9–12]. Approaches based on mean-field games and coalition games typically take intra-team cooperation as a premise induced by a team-level objective. In contrast, team fictitious play (team-FP) focuses on team-level learning dynamics under complete observation of all agents. Dönmez et al. [6] study a multi-team setting in which self-interested agents learn team-level cooperative behavior and propose team-FP, which is shown to converge to a team-Nash equilibrium (TNE).

However, team-FP [6] does not address multi-team learning in which self-interested agents learn to exhibit team-level cooperative behavior under incomplete and possibly unreliable information about opponents. This setting is relevant to applications such as market competition [13] and security problems [14], where self-interested agents may not naturally coordinate toward a team-level objective and decisions are often made under incomplete information about other teams. In addition, the information available for learning can be unreliable or deliberately manipulated, for example, through Byzantine attacks, which may affect agents' decisions and long-term strategic behavior. The challenge is to establish how self-interested agents can learn team-level cooperative behavior when incomplete information and Byzantine attacks affect their decisions and, consequently, their long-term behavior.

To address this challenge, it is important to introduce supervisors as high-level information intermediaries and construct a supervisor network for distributed information learning. The design of supervisors is motivated by information-intermediary roles studied in organizational management and networked systems [15, 16]. Each supervisor is associated with a subset of teams, receives the joint actions reported by these teams, and provides estimates of team-related information to the agents in these teams. Supervisors then exchange information over the supervisor network and update their estimates of team-related information for all teams. The supervisor network provides a basis for distributed learning under incomplete information and Byzantine attacks in multi-team games.

The above motivates the development of algorithms for team orchestration based on distributed information learning over the supervisor network. For algorithm design, supervisors estimate and exchange information about all teams, and correspondingly, agents choose their actions using the information provided by supervisors. In addition, algorithmic resilience against Byzantine attacks should be taken into account since such attacks can corrupt supervisors' information learning and affect agents' strategy optimization. On this basis, we will explore the convergence of the learning dynamics and investigate whether self-interested agents can learn team-level cooperative behavior, as well as the resilience against Byzantine attacks.

J. Zhu and Z. Yang are with the School of Artificial Intelligence and Data Science, University of Science and Technology of China, Hefei 230026, China. Email: zjt1229@mail.ustc.edu.cn, yangzw@ustc.edu.cn.

G. Chen is with the School of Automation, Southeast University, Nanjing 210096, China. Email: guanpu_chen@seu.edu.cn.

T. Zhu is with the College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China. Email: raiden@zju.edu.cn.

M. de Carvalho is with the School of Mathematics, University of Edinburgh, Edinburgh EH9 3FD, Scotland. Email: Miguel.deCarvalho@ed.ac.uk.

F. He is with the School of Informatics, University of Edinburgh, Edinburgh EH8 9AB, Scotland. Email: fhe@ed.ac.uk.

The primary contributions of this paper are as follows:

- We design a supervisor-network mechanism for distributed learning of team-related information. Compared with existing studies on multi-team games [6, 8, 12], the proposed mechanism does not require direct observation of agents in other teams or a priori intra-team cooperation. This enables agents to optimize their strategies using supervisor-mediated belief estimates, thereby supporting team orchestration.
- We develop the distributed team-orchestrating algorithm (DTOA) for team-FP learning over the supervisor network. We establish the convergence of the belief-estimation errors over the uniformly connected supervisor network using a gossip-matrix contraction argument (Theorem 1). We further derive an upper bound on the team-Nash gap (TNG) by comparing the actual learning dynamics with ideal and reference dynamics (Theorem 2).
- We further investigate a Byzantine attack setting in which some teams can misreport their actions to supervisors. We propose the Byzantine-resilient DTOA (BR-DTOA) and show the convergence of the belief-estimation errors for honest teams (Theorem 3). We also reveal an upper bound on the honest TNG, showing that near-TNE convergence for honest teams is resiliently preserved against Byzantine attacks with high-probability identification guarantees (Theorem 4).

The rest of this paper is organized as follows. Section II reviews related work. Section III formulates zero-sum potential team games (ZSPTGs) with a supervisor network and their Byzantine extension. Section IV presents DTOA and establishes convergence and near-TNE guarantees. Section V develops BR-DTOA and analyzes its convergence and resilience. Section VI presents the experimental results. Section VII concludes this paper. The code is available at https://github.com/zjt-1229/team_game_with_supervisor_network.

II. Related Work

This section provides a literature review.

Mean-field games (MFG). Studies on MFGs primarily characterize optimal responses of individuals or representative agents to the mean field in large-scale weakly coupled systems, together with a consistency condition between the induced aggregate statistics and the hypothesized mean field. In standard MFGs, participants are typically modeled as anonymously coupled individuals through aggregate distributions [17]. Subsequent studies further investigate population-level consistency induced by mean-field responses [18], as well as associated learning and computational methods [19]. In parallel, the mean-field team literature focuses on team-optimal decision-making under mean-field coupling [20] and extends this line to more general uncertainty settings [21]. More recently, mixed cooperative–competitive mean-field models have incorporated both cooperation and competition into the mean-field framework [8, 22]. Related studies have also used mixed-coalition formulations with intra-group cooperation and inter-group competition [23] and large-scale competitive team learning [24] to characterize richer collective interactions.

Despite these advances, mean-field-based approaches typically characterize agents' responses to aggregate population statistics rather than team-level equilibrium learning among strategically interacting teams. In contrast, our work addresses the complementary problem of equilibrium formation in repeated multi-team games under limited information, where agents' strategy optimization depends on estimated information acquired through the learning process.

Coalition games. Coalitional-game studies take coalitions as the basic units of analysis, focusing either on coalition formation, coalition values, and the stability of payoff allocations, or, under a fixed coalition partition, on equilibria induced by coalition-level objectives. The early coalition-game literature mainly developed around hedonic preferences [25] and broader modeling and stability analysis of coalition formation [26]. More recent studies further examine the endogenous grouping of self-interested agents and the structural stability of the resulting partitions from a group-formation perspective [27], and provide a systematic account of major classes of coalitional games and their distributed-network applications [28]. In parallel, the literature on multi-coalition games with fixed coalition partitions treats each coalition as a composite agent and studies the existence and distributed computation of generalized Nash equilibria (GNEs) and variational generalized Nash equilibria (vGNEs), together with extensions to nonsmooth, constrained, and dynamic settings [9–12].

However, studies on multi-coalition games with fixed coalition partitions typically model each coalition as a unified agent with a prescribed coalition-level objective and therefore do not capture how team-level behavior emerges from self-interested agent-level interactions. In contrast, our work studies equilibrium formation among teams induced by agent-level learning, which is relevant to multi-team systems without centralized team control.

Byzantine-resilient mechanisms. Uncertainty is an unavoidable issue in many multi-agent game-theoretic settings, and seeking robust or resilient equilibria has become a common approach for preserving desirable performance in uncertain environments [29–31]. Among different sources of uncertainty, Byzantine attacks represent a particularly challenging form, in which malicious agents can distort the information available to others by sending falsified or inconsistent reports. Research on Byzantine-resilient multi-agent systems mainly focuses on how normal agents maintain coordination, identify malicious information sources, and suppress the influence of such sources when adversarial agents disrupt the system by sending falsified messages [32]. Early work focuses primarily on resilient consensus and establishes convergence mechanisms in the presence of adversarial or Byzantine nodes [33]. Building on this line, subsequent studies introduce mechanisms such as distributed detection [34]. These ideas have also been extended beyond consensus to more general control and learning settings, including Byzantine-resilient output regulation [35] and cooperative multi-agent reinforcement learning [36].

These works mainly develop Byzantine-resilient mechanisms for consensus, control, and cooperative learning, with an emphasis on filtering malicious information and preserving coordination among normal agents. They provide useful insights for the Byzantine-resilient algorithm developed in this paper for multi-team games with self-interested agents.

III. Problem Formulation

In this section, we first revisit ZSPTGs and then formulate ZSPTGs with a supervisor network. We also specify the Byzantine attack setting considered in this paper.

A. Revisiting Zero-sum Potential Team Game

In multi-team games, teams compete against one another, while agents in the same team cooperate. A multi-team game is characterized by the tuple $\mathcal{G} = (\mathcal{I}, \mathcal{T}, \{\mathcal{A}^i\}_{i\in\mathcal{I}}, \{u^i\}_{i\in\mathcal{I}})$, where $\mathcal{I}$ and $\mathcal{T}$ denote the index sets of agents and teams, respectively. Let $\mathcal{I}^m$ denote the index set of agents in team $m \in \mathcal{T}$. These sets form a partition of $\mathcal{I}$, i.e., $\mathcal{I}^m \cap \mathcal{I}^{m'} = \varnothing$ for $m \neq m'$ and $\bigcup_{m\in\mathcal{T}} \mathcal{I}^m = \mathcal{I}$. Here, $\mathcal{A}^i$ denotes the finite action set of agent i, $\mathcal{A} \triangleq \prod_{i\in\mathcal{I}} \mathcal{A}^i$ denotes the finite joint action set of all agents, and $u^i : \mathcal{A} \to \mathbb{R}$ denotes the utility function of agent i. Let $\underline{\mathcal{A}}^m \triangleq \prod_{i\in\mathcal{I}^m} \mathcal{A}^i$ denote the joint action set of team $m \in \mathcal{T}$. The mixed-strategy spaces of agent i, team m, and all agents are $\Delta(\mathcal{A}^i)$, $\Delta(\underline{\mathcal{A}}^m)$, and $\Delta(\mathcal{A})$, respectively. We next recall the definition of ZSPTGs [6].

Definition 1 (Zero-sum Potential Team Game)**.** *A multi-team game is called a ZSPTG if, for every team $m \in \mathcal{T}$, there exists a potential function $\phi^m : \mathcal{A} \to \mathbb{R}$ such that*

$$\phi^m(\hat{a}^i, a^{-i}, \underline{a}^{-m}) - \phi^m(a) = u^i(\hat{a}^i, a^{-i}, \underline{a}^{-m}) - u^i(a), \quad (1)$$

for all $(\hat{a}^i, a) \in \mathcal{A}^i \times \mathcal{A}$ and all $i \in \mathcal{I}^m$, where $a^{-i} \triangleq \{a^j\}_{j\in\mathcal{I}^m\setminus\{i\}}$ and $\underline{a}^{-m} \triangleq \{\underline{a}^l\}_{l\in\mathcal{T}\setminus\{m\}}$. Moreover, the potential functions satisfy the zero-sum condition

$$\sum_{m\in\mathcal{T}} \phi^m(a) = 0, \quad \forall a \in \mathcal{A}. \quad (2)$$

The cross-team interactions are network-separable, so that the potential and utility functions can be decomposed as

$$\phi^m = \sum_{l\neq m} \phi^{ml} \text{ and } u^i = \sum_{l\neq m} u^{il}, \ \forall i \in \mathcal{I}^m, \quad (3)$$

for some $\phi^{ml} : \underline{\mathcal{A}}^m \times \underline{\mathcal{A}}^l \to \mathbb{R}$ and $u^{il} : \underline{\mathcal{A}}^m \times \underline{\mathcal{A}}^l \to \mathbb{R}$.

We next recall the notions of the TNG and TNE [6].

Definition 2 (Team-Nash Gap)**.** *Given a team strategy profile $\pi = \{\pi^m \in \Delta(\underline{\mathcal{A}}^m)\}_{m\in\mathcal{T}}$, the TNG for team m is defined as*

$$TNG^m(\pi) \triangleq \max_{\pi'\in\Delta(\underline{\mathcal{A}}^m)} \{\phi^m(\pi', \pi^{-m})\} - \phi^m(\pi).$$

The TNG is then defined as $TNG(\pi) \triangleq \sum_{m\in\mathcal{T}} TNG^m(\pi)$, where $\pi^{-m} = \{\pi^l\}_{l\in\mathcal{T}\setminus m}$. Correspondingly, π is called a TNE if $TNG(\pi) = 0$. Furthermore, for any $\varepsilon > 0$, π is called an ε-TNE if $TNG(\pi) < \varepsilon$.

Intuitively, a small TNG means that the induced team strategy profile is nearly stable against unilateral deviations of any single team from its current strategy. When the team strategy profile has a small TNG, the resulting behavior reflects an approximate pattern of within-team cooperation and inter-team competition. Thus, we use the TNG to evaluate the extent to which self-interested agents in each team behave cooperatively while competing with other teams.

To study how self-interested agents choose their actions to maximize their own utilities, we next revisit the smoothed best response used in the subsequent learning algorithms.

Definition 3 (Smoothed Best Response)**.** *Given a finite set X, for any distribution $D \in \Delta(X)$ and any function $f : X \to \mathbb{R}$, let $f(D) = \sum_{x\in X} D(x)f(x)$. For a temperature parameter $\tau > 0$, the smoothed best response to f is defined as*

$$br_\tau(f)(x) = \frac{e^{\frac{f(x)}{\tau}}}{\sum_{\tilde{x}\in X} e^{\frac{f(\tilde{x})}{\tau}}}, \ \forall x \in X.$$

Equivalently, $br_\tau(f) = \arg\max_{D\in\Delta(X)}\{f(D) + \tau\mathcal{H}(D)\}$, where $\mathcal{H}(D) \triangleq \sum_{x\in X} -D(x)\log D(x)$ is the entropy regularization term.

The smoothed best response provides a regularized decision rule that assigns positive probability to all feasible actions and balances objective optimization with exploration. The temperature parameter τ controls the degree of smoothing: smaller values make the response closer to a pure best response, whereas larger values induce more exploratory behavior. The maximizer is unique because $f(D)$ is linear in D and, for $\tau > 0$, the entropy regularization term makes the objective strictly concave on $\Delta(X)$.

Note that the utilities of all agents need not sum to zero. Although the definition of a ZSPTG requires the team potentials to sum to zero, the same analysis applies when their sum is a constant independent of the action profile $a \in \mathcal{A}$ and agents choose actions according to the smoothed best response. Indeed, the potentials can be normalized by subtracting this constant from any one team potential, which preserves all potential differences and therefore leaves the smoothed best-response structure unchanged.

B. ZSPTG with a supervisor network

In this paper, we focus on ZSPTGs with a supervisor network. A multi-team game with a supervisor network is defined by the tuple $\mathcal{G} = (\mathcal{I}, \mathcal{T}, \mathcal{S}, ST, \{SN_k\}_{k\geq 0}, \{\mathcal{A}^i\}_{i\in\mathcal{I}}, \{u^i\}_{i\in\mathcal{I}})$, where $\mathcal{I}$, $\mathcal{T}$, and $\mathcal{S}$ denote the index sets of agents, teams, and supervisors, respectively. We follow the notation in Section III-A: $\mathcal{I}^m$, $\mathcal{A}^i$, $\underline{\mathcal{A}}^m$, $\mathcal{A}$, and $u^i : \mathcal{A} \to \mathbb{R}$ denote the index set of agents in team m, the action set of agent i, the joint action set of team m, the joint action set of all agents, and the utility function of agent i, respectively. We assume that each agent observes only the actions of agents in the same team. The matrix $ST \in \{0,1\}^{|\mathcal{T}|\times|\mathcal{S}|}$ denotes the supervision relationships between teams and supervisors, and the matrix $SN_k \in \{0,1\}^{|\mathcal{S}|\times|\mathcal{S}|}$ denotes the communication structure among supervisors at round k.

Each supervisor supervises a subset of teams: it receives action reports from these teams and provides their agents with information about other teams. Specifically, $ST(m, s) = 1$ indicates that team m is supervised by supervisor s, and $ST(m, s) = 0$ otherwise. Similarly, $SN_k(s, s') = 1$ indicates that supervisors s and s' exchange information at round k, and $SN_k(s, s') = 0$ otherwise. Each agent aims to maximize its utility based on the information provided by the supervisors. For each supervisor $s \in \mathcal{S}$, let $\mathcal{T}^s \subseteq \mathcal{T}$ denote the set of teams supervised by s. The supervisor communication structure at

round k can be equivalently represented by an undirected graph $G_k = (\mathcal{S}, E_k)$, where $\mathcal{S}$ is the node set and $E_k = \left\{\{s, s'\} : SN_k(s, s') = 1\right\}$ is the edge set, for all $k \geq 0$.

In the team-FP learning dynamics, agents update their actions based on common beliefs about other teams' strategies [6], which are formed through direct observation of the actions of other agents. In ZSPTGs with a supervisor network, however, these beliefs are provided by supervisors that receive action reports from supervised teams and exchange information over the supervisor network. Consequently, belief-estimation errors can arise because each supervisor receives reports from only a subset of teams. It is therefore crucial to characterize how such errors affect the convergence of the learning dynamics and the induced TNG. This motivates the following problem.

Problem 1. *How to design a distributed team-orchestrating algorithm for ZSPTGs with a supervisor network while providing theoretical guarantees?*

This problem concerns whether the common-belief requirement in team-FP can be relaxed through distributed belief learning over the supervisor network. The key challenge is to control the long-term effect of supervisors' belief-estimation errors on the induced team behavior. We first characterize the belief-estimation process over the supervisor network and then quantify its effect on the induced learning dynamics and the TNG. We impose the following assumptions.

Assumption 1. *Every team is supervised by at least one supervisor, i.e.,* $\bigcup_{s\in\mathcal{S}} \mathcal{T}^s = \mathcal{T}$.

Assumption 2. *The supervisor communication graph* $G_k \triangleq (\mathcal{S}, E_k)$*, which represents the supervisor network, is time-varying and uniformly connected over time [37]. Specifically, let* B *be the smallest positive integer such that, for any* $n \in \mathbb{N}^+$*, the union graph* $G_{(n)} \triangleq (\mathcal{S}, \bigcup_{k=nB}^{(n+1)B-1} E_k)$ *is connected.*

Assumption 1 ensures that every team has at least one reporting link to the supervisor network and hence no team is completely isolated from it. Assumption 2 ensures that information can propagate among all supervisors within each finite communication window.

C. Byzantine Attack

We consider a Byzantine attack setting in which some teams can misreport their actions to the supervisors that supervise them at each round. Given a ZSPTG with a supervisor network $\mathcal{G} = (\mathcal{I}, \mathcal{T}, \mathcal{S}, ST, \{SN_k\}_{k\geq 0}, \{\mathcal{A}^i\}_{i\in\mathcal{I}}, \{u^i\}_{i\in\mathcal{I}})$, honest teams and Byzantine teams are defined as follows.

Definition 4 (Honest Team and Byzantine Team)**.** *A team* m *is called honest if it reports its joint action truthfully to the supervisors that supervise it, i.e.,* $\underline{a}^m_{k,r} = \underline{a}^m_k$ *for all* $k \geq 0$*, where* $\underline{a}^m_{k,r}$ *denotes the joint action reported by team* m *at round* k*. In contrast, a team* m *is called Byzantine if it can misreport its joint action to the supervisors that supervise it, i.e.,* $\underline{a}^m_{k,r} \neq \underline{a}^m_k$ *for some* $k \geq 0$*.*

Let $\mathcal{H} \subseteq \mathcal{T}$ and $\mathcal{B} \subseteq \mathcal{T}$ denote the sets of honest and Byzantine teams, respectively. We next define the honest TNG.

Definition 5 (Honest Team-Nash Gap)**.** *Given a team strategy profile* $\pi = \{\pi^m \in \Delta(\underline{\mathcal{A}}^m)\}_{m\in\mathcal{T}}$*, the honest TNG is defined as* $TNG_{\mathcal{H}}(\pi) \triangleq \sum_{m\in\mathcal{H}} TNG^m(\pi)$*. Correspondingly, the team strategy profile* π *is called an* ε*-honest TNE if* $TNG_{\mathcal{H}}(\pi) < \varepsilon$*.*

In the definition of the honest TNG, the strategies of Byzantine teams are treated as fixed exogenous factors, and the gap is evaluated only over honest teams. Since our objective is to optimize the performance of honest teams under Byzantine attacks, including the gaps for Byzantine teams in the honest TNG does not reflect the intended performance measure.

In this Byzantine attack setting, the learning dynamics are affected by Byzantine teams, whose misreports can distort supervisors' belief learning processes and influence the long-term behavior of agents in honest teams. This leads to the following Byzantine-resilient learning problem.

Problem 2. *How to design a distributed resilient algorithm for ZSPTGs with a supervisor network against Byzantine attacks?*

The key challenge is to identify Byzantine teams using limited information while preventing their misreports from disrupting the long-term learning behavior of honest teams. We address this challenge by developing a supervisor-based mechanism and quantifying the effect of Byzantine teams on the learning dynamics of honest teams. To this end, each supervisor can check whether the action reports from its supervised teams are consistent with their true actions. The checking outcome is not necessarily accurate: an honest report can be incorrectly regarded as a misreport, and a misreport can fail to be detected.

IV. Distributed Team-Orchestrating Algorithm

This section addresses Problem 1 by developing the DTOA. We then analyze the convergence of supervisors' belief-estimation errors and derive an upper bound on the TNG.

A. Algorithm Design

We first specify the main components of DTOA for ZSPTGs with a supervisor network. At round k, let a^i_k denote the action of agent i, and let $\underline{a}^m_k = \left(a^i_k\right)_{i\in\mathcal{I}^m}$ denote the joint action of team m. To compensate for the lack of complete opponent-action information, DTOA uses supervisors to provide belief estimates. Let $\pi^{m,s}_k \in \Delta\left(\underline{\mathcal{A}}^m\right)$ denote the belief of supervisor s regarding the strategy of team m. To induce team-level strategic behavior under incomplete opponent-action information, agent i in team m updates its action according to the smoothed best response in Definition 3, using the previous actions of the other agents in the same team and the beliefs provided by a supervisor s that supervises team m:

$$a^i_k \sim br_\tau\left(u^i(\cdot, a^{-i}_{k-1}, \pi^{-m,s}_k)\right), \tag{4}$$

where $a^i_k \sim p$ means that a^i_k is sampled from $p \in \Delta\left(\mathcal{A}^i\right)$.

The rule above specifies the action update of a selected agent. In DTOA, one agent per team is randomly selected at each round to update its action, while the other agents in the same team keep their actions unchanged. When such coordination is unavailable, we consider an independent variant, called independent DTOA (iDTOA), in which each agent updates its action independently with probability $\delta \in (0, 1)$.

Algorithm 1 Distributed Team-Orchestrating Algorithm (DTOA)

Initialize: $\{\pi^m_{s,0}\}_{m\in\mathcal{T}}$ and $\{a^i_{-1}\}_{i\in\mathcal{I}}$ arbitrarily
while round $k=0,1,\dots$ **do**
  for $m\in\mathcal{T}$ **do**
    select agent $i\in\mathcal{I}^m$ in team m randomly
    select $s\in\{s: i\in\mathcal{T}^s\}$ randomly to provide belief estimates about other teams
    agent i updates its action based on a^{-i}_{k-1} and $\pi^{-m,s}_k$: $a^i_k\sim br_\tau\left(u^i(\cdot,a^{-i}_{k-1},\pi^{m,s}_k)\right)$
    for agent $j\in\mathcal{I}^m\backslash\{i\}$ **do**
      repeat the last action: $a^j_k=a^j_{k-1}$
    end for
  end for
  for $s\in\mathcal{S}$ and $m\in\mathcal{T}$ **do**
    if $m\in\mathcal{T}^s$ **then**
      belief update: $\pi^{m,s}_{k+1}=\pi^{m,s}_k+\alpha_k\left(\underline{a}^m_k-\pi^{m,s}_k\right)$
    else
      belief update:
$$\pi^{m,s}_{k+1}=\begin{cases}\frac{1}{|\mathcal{N}^s_k|}\sum_{s'\in\mathcal{N}^s_k}\pi^{m,s'}_k, & \text{for } |\mathcal{N}^s_k|>0\\ \pi^{m,s}_k, & \text{for } |\mathcal{N}^s_k|=0\end{cases}$$
    end if
  end for
end while

To maintain belief estimates of teams' strategies, we design two belief-update rules: one for teams directly supervised by a given supervisor and the other for teams not directly supervised by that supervisor. For a supervisor $s\in\mathcal{S}$ and any team $m\in\mathcal{T}^s$, supervisor s directly updates its belief using the joint action reported by team m:

$$\pi^{m,s}_{k+1}=\pi^{m,s}_k+\alpha_k\left(\underline{a}^m_k-\pi^{m,s}_k\right),$$

where $\{\alpha_k\}_{k\geq 0}$ is the step-size sequence used by all supervisors, and $\underline{a}^m_k$ denotes its one-hot representation in $\Delta(\underline{\mathcal{A}}^m)$ for notational simplicity. Let $\mathcal{N}^s_k=\{s'\in\mathcal{S}: SN_k(s,s')=1\}$ denote the set of neighbors of supervisor s at round k. For any team $m\notin\mathcal{T}^s$, if $|\mathcal{N}^s_k|>0$, supervisor s updates its belief using information provided by its neighbors:

$$\pi^{m,s}_{k+1}=\frac{1}{|\mathcal{N}^s_k|}\sum_{s'\in\mathcal{N}^s_k}\pi^{m,s'}_k.$$

If $|\mathcal{N}^s_k|=0$, supervisor s keeps its beliefs unchanged, i.e., $\pi^{m,s}_{k+1}=\pi^{m,s}_k$ for all $m\notin\mathcal{T}^s$. The implementation of DTOA is summarized in Algorithm 1.

B. Convergence Analysis

We now turn to the convergence analysis of DTOA. Following the step-size conditions considered in [6], we impose the following standard assumptions on the step sizes.

Assumption 3. *The step-size sequence $\{\alpha_k\}_{k\geq 0}$ satisfies the following conditions:*

- $\alpha_k\in[0,1]$, *and* $\alpha_k\to 0$ *as* $k\to\infty$;
- $\sum_{k=0}^\infty\alpha_k=\infty$ *and* $\sum_{k=0}^\infty\alpha_k^2<\infty$;
- $\lim_{k\to\infty}\alpha_k/\alpha_{k+1}=1$ *and* $\alpha_k-\alpha_{k+1}\geq\alpha_k\alpha_{k+1}$.

The first two conditions in Assumption 3 ensure that the belief updates continue to incorporate new action information while the effect of sampling fluctuations is asymptotically averaged out. The last condition further ensures that action information from adjacent periods has comparable influence on the resulting beliefs. A standard choice satisfying these conditions is $\alpha_k=1/(k+1)$, which corresponds to empirical averaging over past actions.

To quantify supervisors' belief-estimation errors, we define true beliefs in the full-supervision case where each supervisor supervises all teams. The true beliefs evolve according to

$$\pi^m_{k+1}=\pi^m_k+\alpha_k\left(\underline{a}^m_k-\pi^m_k\right),\ \forall m\in\mathcal{T}. \tag{5}$$

Under Assumption 3, the effect of the initial beliefs vanishes asymptotically. Hence, the initialization does not affect the asymptotic convergence results. For ease of analysis, we set $\pi^{m,s}_0=\pi^m_0$ for all $m\in\mathcal{T}$ and $s\in\mathcal{S}$.

Remark 1. *In the full-supervision case, for any supervisor $s\in\mathcal{S}$ and any team $m\in\mathcal{T}$, we have $\pi^{m,s}_k=\pi^m_k$ for all $k\geq 0$. By contrast, under a general supervision structure, a supervisor $s\in\mathcal{S}$ may not directly supervise a team $m\in\mathcal{T}$, i.e., $m\notin\mathcal{T}^s$, in which case the equality $\pi^{m,s}_k=\pi^m_k$ may no longer hold.*

For supervisor $s\in\mathcal{S}$ and team $m\in\mathcal{T}$, we define the belief-estimation error with respect to the true belief as $\|\pi^{m,s}_k-\pi^m_k\|_\infty$. The following theorem establishes the asymptotic convergence of all supervisors' belief-estimation errors under DTOA. The proof is provided in Appendix A.

Theorem 1. *For a ZSPTG with a supervisor network*

$$\mathcal{G}=(\mathcal{I},\mathcal{T},\mathcal{S},ST,\{SN_k\}_{k\geq 0},\{\mathcal{A}^i\}_{i\in\mathcal{I}},\{u^i\}_{i\in\mathcal{I}}),$$

under Assumptions 1 and 2, the supervisors' belief-estimation errors in DTOA converge to zero; specifically, $\forall s\in\mathcal{S}$, $\forall m\in\mathcal{T}$,

$$\|\pi^{m,s}_k-\pi^m_k\|_\infty\leq O(\rho^{[\frac{k}{2}]})+O(\alpha_{[\frac{k}{2}]}), \tag{6}$$

where $\rho=\left(1-\left(\frac{1}{|\mathcal{S}|}\right)^{(|\mathcal{S}|+1)B}\right)^{\frac{1}{(|\mathcal{S}|+1)B}}<1$. If Assumption 3 also holds, (6) simplifies to $\|\pi^{m,s}_k-\pi^m_k\|_\infty\leq O(\alpha_{[\frac{k}{2}]})$.

Theorem 1 shows that the convergence rate of the supervisors' beliefs depends on the supervisor network $\{SN_k\}_{k\geq 0}$, the number of supervisors $|\mathcal{S}|$, and the step-size sequence $\{\alpha_k\}_{k\geq 0}$. Here, B characterizes the length of the communication window over which information is propagated among supervisors. The first term in (6) indicates that, for fixed B and $\{\alpha_k\}_{k\geq 0}$, a larger number of supervisors $|\mathcal{S}|$ leads to slower convergence. We next fix $|\mathcal{S}|$ and $\{\alpha_k\}_{k\geq 0}$ and examine how B affects the convergence rate. Since the logarithm is strictly increasing, the monotonicity of ρ with respect to B is equivalent to the monotonicity of $\log\rho$ with respect to B. For all $|\mathcal{S}|>1$ and all $B>0$, we have

$$\frac{\partial(\log\rho)}{\partial B}=\frac{(|\mathcal{S}|+1)B\cdot\frac{\log|\mathcal{S}|}{|\mathcal{S}|^{(|\mathcal{S}|+1)B}-1}}{(|\mathcal{S}|+1)^2B^2}>0.$$

This implies that ρ increases with B; hence, a larger B leads to a slower convergence rate. The second term in (6) further shows that the step-size sequence $\{\alpha_k\}_{k\geq 0}$ also affects the convergence rate of the supervisors' beliefs.

C. TNG Analysis

Having established the convergence of supervisors' belief estimates, we next analyze TNG convergence under DTOA. The main challenge in proving TNG convergence for DTOA is that supervisors' belief-estimation errors affect agents' actions at each round, thereby influencing subsequent belief updates and action decisions. Although supervisors' belief-estimation errors converge to zero as $k \to \infty$, it remains nontrivial to show that their cumulative influence on agents' long-term behavior also vanishes asymptotically.

To address this difficulty, we introduce an ideal scenario for the analysis of DTOA and refer to the original setting with belief-estimation errors as the actual scenario. The repeated play is divided into epochs, each consisting of T rounds. In the ideal scenario, at the beginning of each epoch, all supervisors are initialized with the history of all agents' actions from the actual scenario and supervise all teams throughout the epoch. Consequently, they share common beliefs at each round, which are referred to as ideal beliefs. Since the actual and ideal scenarios use different supervision structures during epoch n, they generally induce different distributions over the joint actions of all teams at round k of epoch n.

Remark 2. *The ideal beliefs generally differ from the true beliefs. Specifically, the true beliefs are determined by the weighted empirical average of the action history generated in the actual scenario, whereas the ideal beliefs depend on both the actions generated during the current epoch in the ideal scenario and the action history generated before epoch n in the actual scenario. Nevertheless, at the beginning of each epoch, the ideal beliefs coincide with the true beliefs because both are constructed from the action history generated before epoch n in the actual scenario.*

The key idea is to use an epoch-wise comparison between the ideal and actual scenarios to quantify how supervisors' belief-estimation errors propagate to the induced action distributions. Let $\nu^m_{k,(n)}$ and $a\text{-}\nu^m_{k,(n)}$ denote the joint-action distributions of team m at round k of epoch n in the ideal and actual scenarios, respectively. The following lemma quantifies how the difference between $\nu^m_{k,(n)}$ and $a\text{-}\nu^m_{k,(n)}$ is bounded in terms of supervisors' belief-estimation errors.

Lemma 1. *For a ZSPTG with a supervisor network*

$$\mathcal{G} = (\mathcal{I}, \mathcal{T}, \mathcal{S}, ST, \{SN_k\}_{k\geq 0}, \{\mathcal{A}^i\}_{i\in\mathcal{I}}, \{u^i\}_{i\in\mathcal{I}}),$$

under Assumptions 1 and 2, the difference between the induced action distributions for DTOA in the actual and ideal scenarios can be bounded as

$$\|\nu^m_{k,(n)} - a\text{-}\nu^m_{k,(n)}\|_1 \leq C\delta_{(n)},$$

where $\delta_{(n)} = \max_{k\in[nT-1,(n+1)T-1]} \sum_{m\in\mathcal{T}} \max_{s\in\mathcal{S}} \|\pi^{m,s}_k - \pi^m_k\|_1$ denotes the maximum aggregate supervisor belief-estimation error in epoch n. Since Theorem 1 implies $\delta_{(n)} \to 0$ as $n \to \infty$, it follows that $\|\nu^m_{k,(n)} - a\text{-}\nu^m_{k,(n)}\|_1$ also converges to zero.

Lemma 1 establishes that the difference between the action distributions induced by the actual and ideal beliefs is bounded by a term proportional to the maximum aggregate supervisor belief-estimation error in epoch n. This bound is crucial for analyzing agents' long-term behavior in the actual scenario. The proof is provided in Appendix B.

Remark 3. *At round $k = nT$, Remark 2 and the properties of the smoothed best response imply that $\|\nu^m_{nT,(n)} - a\text{-}\nu^m_{nT,(n)}\|_1$ can be controlled by $\|\pi^{m,s}_{nT} - \pi^m_{nT}\|_1$. However, this argument does not directly extend to $nT < k \leq (n+1)T-1$, because the action distributions in an epoch are affected by the accumulated discrepancy between the actual and ideal scenarios.*

We now present the TNG convergence result. Building on Theorem 1 and Lemma 1, the following theorem establishes an almost-sure upper bound on the TNG under DTOA.

Theorem 2. *For a ZSPTG with a supervisor network*

$$\mathcal{G} = (\mathcal{I}, \mathcal{T}, \mathcal{S}, ST, \{SN_k\}_{k\geq 0}, \{\mathcal{A}^i\}_{i\in\mathcal{I}}, \{u^i\}_{i\in\mathcal{I}}),$$

under Assumptions 1, 2, and 3, DTOA satisfies, for any $s \in \mathcal{S}$,

$$\limsup_{k\to\infty} TNG(\pi^s_k) \leq \begin{cases} \tau \log|\mathcal{A}|, & \text{for DTOA}, \\ \tau \log|\mathcal{A}| + |\mathcal{T}|^2 \overline{\phi} \Lambda(\delta, \varepsilon_\phi), & \text{for iDTOA}, \end{cases}$$

where $\overline{\phi} = \max_{(m,l,a)} |\phi^{ml}(a)|$, $\pi^s_k = \{\pi^{m,s}_k\}_{m\in\mathcal{T}}$, and $\Lambda(\delta, \varepsilon_\phi)$ is a function decaying to zero as $\delta \to 0^+$ for any $0 < \varepsilon_\phi \leq \min_{a\in\mathcal{A}} br_\tau\left(a, a^{-i}\right)$.

To proceed with the TNG convergence analysis, we introduce a reference scenario following [6]. In this reference scenario, at any round k in epoch n, agents update their actions using the true beliefs at the beginning of epoch n. Specifically, for $nT \leq k \leq (n+1)T-1$, the selected agent i follows the reference update rule $a^i_k \sim br_\tau\left(u^i(\cdot, a^{-i}_{k-1}, \pi^{-m}_{nT})\right)$. The comparison between the actual and reference scenarios can then be decomposed into two parts: the comparison between the actual and ideal scenarios and the comparison between the ideal and reference scenarios. The proof is in Appendix B.

Theorem 2 shows that the long-term behavior of all agents induces a near TNE. This implies that agents in the same team learn to cooperate with one another while competing against other teams, even though each agent only knows its own utility. The resulting TNG bound matches that in Theorem 4.2 of [6]. In contrast, our analysis explicitly accounts for belief-estimation errors induced by supervisor-network learning. Accordingly, DTOA replaces the common-belief requirement in team-FP with supervisor-based distributed belief learning. DTOA enables agents to optimize their strategies for team orchestration while ensuring the convergence of belief-estimation errors and providing a small-TNG guarantee. These results provide an affirmative answer to Problem 1.

V. Byzantine Resilience

This section addresses Problem 2 by developing a supervisor-based Byzantine-identification mechanism and adapting DTOA to the Byzantine attack setting.

We consider a Byzantine attack setting for ZSPTGs with a supervisor network, adapted from Byzantine-resilient multi-agent learning [38] and imperfect verification [39]. In this setting, each team is either honest or Byzantine; that is, $\mathcal{H} \cap \mathcal{B} =$

$\emptyset$ and $\mathcal{H} \cup \mathcal{B} = \mathcal{T}$. Agents in honest teams update their actions according to (4) and report their joint actions truthfully to their supervisors, whereas agents in Byzantine teams update their actions arbitrarily and can misreport their joint actions. Supervisors aim to identify Byzantine teams and mitigate their influence on honest teams. The supervisors' checking policy and the Byzantine teams' attack policy are specified below.

Supervisors' checking policy. For a supervisor $s \in \mathcal{S}$ and a team $m \in \mathcal{T}^s$, supervisor s checks, with a prescribed probability, whether the joint action reported by team m is consistent with the joint action actually taken by team m. We use a binary variable $v_k^{m,s} \in \{0,1\}$ to indicate this decision, where $v_k^{m,s} = 1$ means that supervisor s checks team m at round k. After checking team m at round k, supervisor s receives a verification signal $z_k^{m,s} \in \{\text{fail}, \text{pass}\}$. The verification signals $z_k^{m,s}$ are not necessarily accurate. Specifically, we assume that there exist constants $\eta_{FP} \in (0,1]$ and $\eta_{FN} \in [0,1)$ such that, for any $s \in \mathcal{S}$, $m \in \mathcal{T}^s$, and $k \geq 0$,

$$\mathbb{P}\left(z_k^{m,s} = \text{fail} \mid v_k^{m,s} = 1,\ \underline{a}_{k,r}^m = \underline{a}_k^m\right) \leq \eta_{FP},$$
$$\mathbb{P}\left(z_k^{m,s} = \text{pass} \mid v_k^{m,s} = 1,\ \underline{a}_{k,r}^m \neq \underline{a}_k^m\right) \leq \eta_{FN}.$$

Here, η_{FP} and η_{FN} denote upper bounds on the false-positive and false-negative probabilities, respectively. Smaller values of η_{FP} and η_{FN} correspond to more accurate verification.

Byzantine attack policy. A Byzantine team can be identified more readily if it misreports its joint action at every round. We therefore assume that there exists $p_{\text{lie}} \in [0,1]$ such that each Byzantine team misreports its joint action to its supervisors with probability p_{lie} at each round. If a Byzantine team does not misreport at a given round, then it reports its true joint action at that round. In this case, the Byzantine team is indistinguishable from an honest team from the supervisors' perspective, because the verification signal depends only on whether the reported joint action is consistent with the joint action actually taken. Moreover, supervisors have no access to any agent's utility function or any team's potential function.

In practice, a verification signal is meaningful only if it is informative about whether a team has misreported its joint action. We impose the following consistency condition on the verification signals. Specifically, η_{FP}, η_{FN}, and p_{lie} satisfy

$$\eta_{FP} < (1 - p_{\text{lie}})\eta_{FP} + p_{\text{lie}}(1 - \eta_{FN}). \tag{7}$$

Condition (7) means that a checked Byzantine team has a higher probability of generating a fail signal than a checked honest team. Equivalently, supervisors are more likely to receive a pass signal after checking an honest team than after checking a Byzantine team.

The Byzantine attack setting described above is referred to as a Byzantine ZSPTG with a supervisor network and is defined by the tuple

$$\mathcal{G}_B = (\mathcal{I}, \mathcal{H}, \mathcal{B}, \mathcal{S}, ST, \{SN_k\}_{k\geq 0}, \{\mathcal{A}^i\}_{i\in\mathcal{I}}, \{u^i\}_{i\in\mathcal{I}}, p),$$

where $\mathcal{H}$ and $\mathcal{B}$ denote the sets of honest and Byzantine teams, respectively, and $p = (\eta_{FP}, \eta_{FN}, p_{\text{lie}})$ denotes the vector of verification parameters satisfying (7). This formulation serves as the basis for the subsequent analysis of Problem 2.

Algorithm 2 Byzantine-resilient DTOA (BR-DTOA)

Initialise: $\{\pi_{s,0}^m\}_{m\in\mathcal{T}}$ and $\{a_{-1}^i\}_{i\in\mathcal{I}}$ arbitrarily, $F_0^{m,s} = 0$
while round $k = 0, 1, \ldots$ **do**
  for $m \in \mathcal{T}$ **do**
    select agent $i \in \mathcal{I}^m$ in team m randomly
    if m is an honest team **then**
      select $s \in \{s : i \in \mathcal{T}^m\}$ randomly to provide belief estimates about other teams
      agent i updates its action based on a_{k-1}^{-i} and $\pi_k^{m,s}$: $a_k^i \sim br_\tau\left(u^i(\cdot, a_{k-1}^{-i}, \pi_k^{m,s})\right)$
    else
      agent i updates its action randomly
    end if
    for $j \in \mathcal{I}^m \backslash \{i\}$ **do**
      repeat the last action: $a_k^j = a_{k-1}^j$
    end for
    report the joint action $\underline{a}_{k,r}^m$ to supervisors $\mathcal{T}^m$
  end for
  for $s \in \mathcal{S}$ and $m \in \mathcal{T}$ **do**
    if $m \in \mathcal{T}^s$ **then**
      check whether team m lies with probability q and get a verification signal $z_k^{m,s} \in \{\text{pass}, \text{fail}\}$
      update $F_k^{m,s} = F_{k-1}^{m,s} + 1\,(z_k^{m,s} = fail)$
      update $f_k^{m,s} = F_k^{m,s}/k$
      if $k < K$ or $f_k^{m,s} < f$ **then**
        belief update: $\pi_{k+1}^{m,s} = \pi_k^{m,s} + \alpha_k\left(\underline{a}_{k,r}^m - \pi_k^{m,s}\right)$
      end if
    else
      belief update:
      $$\pi_{k+1}^{m,s} = \begin{cases} \frac{1}{|\mathcal{N}_k^s|}\sum_{s'\in\mathcal{N}_k^s} \pi_k^{m,s'}, & \text{for } |\mathcal{N}_k^s| > 0 \\ \pi_k^{m,s}, & \text{for } |\mathcal{N}_k^s| = 0 \end{cases}$$
    end if
  end for
end while

A. Algorithm Design

To address the Byzantine attack setting described above, we extend DTOA with a Byzantine-resilient mechanism that operates at the supervisor-team level. The resulting algorithm, referred to as BR-DTOA, is summarized in Algorithm 2.

Byzantine-resilient mechanism. Since each checking outcome can be inaccurate, supervisors do not rely on a single fail verification signal to identify Byzantine teams. Instead, they accumulate checking evidence over time. For a supervisor $s \in \mathcal{S}$ and a team $m \in \mathcal{T}^s$, let $F_k^{m,s}$ denote the number of times that supervisor s has received a fail signal from team m up to round k. Supervisor s checks team m at round k with probability q and updates $F_k^{m,s}$ as follows:

$$F_k^{m,s} = F_{k-1}^{m,s} + \mathbf{1}\{v_k^{m,s} = 1,\ z_k^{m,s} = \text{fail}\}.$$

The empirical frequency of receiving a fail signal from team m is then defined as $f_k^{m,s} = F_k^{m,s}/k$. To reduce the effect of early-stage fluctuations, supervisors start labeling teams only after a burn-in period of K rounds. Given a threshold

f satisfying $q\eta_{FP} < f < q\left[(1-p_{\text{lie}})\eta_{FP} + p_{\text{lie}}(1-\eta_{FN})\right]$, if $k > K$ and $f_k^{m,s} > f$, then supervisor s labels team m as Byzantine. Such an f exists by Condition (7).

Once a team m is labeled as Byzantine by a supervisor s, supervisor s keeps this label in all subsequent rounds and stops providing belief estimates to team m. This label is also used to prevent identified Byzantine teams from further affecting supervisors' belief estimates. At round k, supervisor s updates its belief regarding team m according to

$$\pi_{k+1}^{m,s}=\begin{cases}\pi_k^{m,s}+\omega_k^{m,s}\alpha_k\Big(\underline{a}_{k,r}^m-\pi_k^{m,s}\Big), & \text{for } m\in\mathcal{T}^s,\\ \frac{1}{|\mathcal{N}_k^s|}\sum_{s'\in\mathcal{N}_k^s}\pi_k^{m,s'}, & \text{for } m\notin\mathcal{T}^s \text{and } |\mathcal{N}_k^s|>0,\\ \pi_k^{m,s}, & \text{for } m\notin\mathcal{T}^s \text{and } |\mathcal{N}_k^s|=0,\end{cases}$$

where $\omega_k^{m,s}=1$ if team m has not been labeled as Byzantine by supervisor s, and $\omega_k^{m,s}=0$ otherwise.

B. Resilience Analysis

We analyze BR-DTOA in two steps. We first examine the Byzantine-resilient mechanism and then study the convergence of BR-DTOA and the associated honest TNG. The following lemma provides exponential bounds on Byzantine-team identification errors.

Lemma 2. *Given a Byzantine ZSPTG with a supervisor network $\mathcal{G}_B=(\mathcal{I},\mathcal{H},\mathcal{B},\mathcal{S},ST,\{SN_k\}_{k\geq 0},\{\mathcal{A}^i\}_{i\in\mathcal{I}},\{u^i\}_{i\in\mathcal{I}},p)$, for any supervisor s and any team $m\in\mathcal{T}^s$, the following statements hold under BR-DTOA:*

- *if $m\in\mathcal{H}$, then*

$$\mathbb{P}\left(f_k^{m,s}>f\right)\leq\exp\left(-kD(f\|q\eta_{FP})\right);$$

- *if $m\in\mathcal{B}$, then*

$$\mathbb{P}\left(f_k^{m,s}<f\right)\leq\exp\Big(-kD\big(f\|q\left[(1-p_{\text{lie}})\eta_{FP}+p_{\text{lie}}(1-\eta_{FN})\right]\big)\Big);$$

where $D(x\|y)\triangleq x\log\left(\frac{x}{y}\right)+(1-x)\log\left(\frac{1-x}{1-y}\right)$ denotes the binary relative entropy.

Lemma 2 shows that the probabilities of falsely labeling an honest team as Byzantine and failing to identify a Byzantine team by round k decay exponentially as $k\to\infty$. A larger checking probability q and smaller error rates η_{FP} and η_{FN} can improve the identification performance by reducing the misidentification probabilities. The threshold f controls the trade-off between missed identifications of Byzantine teams and false alarms for honest teams. Increasing f reduces the probability of falsely labeling an honest team as Byzantine but increases the probability of failing to identify a Byzantine team. Conversely, decreasing f facilitates Byzantine-team identification but raises the risk of misclassifying honest teams.

We now derive convergence guarantees for BR-DTOA from the identification result in Lemma 2 and the convergence analysis of DTOA in the previous section. The first result concerns supervisors' belief-estimation errors for honest teams.

Theorem 3. *For a Byzantine ZSPTG with a supervisor network $\mathcal{G}_B=(\mathcal{I},\mathcal{H},\mathcal{B},\mathcal{S},ST,\{SN_k\}_{k\geq 0},\{\mathcal{A}^i\}_{i\in\mathcal{I}},\{u^i\}_{i\in\mathcal{I}},p)$, under the conditions of Theorem 1, for any supervisor $s\in\mathcal{S}$, the following statement holds for BR-DTOA with probability at least $1-\delta_B$:*

$$\|\pi_k^{m,s}-\pi_k^m\|_\infty\leq O(\rho^{[\frac{k}{2}]})+O(\alpha_{[\frac{k}{2}]}),\quad\forall m\in\mathcal{H},\quad(8)$$

where $\delta_B=\delta_B^{\mathcal{H}}+\delta_B^{\mathcal{B}}$ with $\delta_B^{\mathcal{H}}=|\mathcal{H}|\cdot\frac{\exp(-K\cdot D(f\|q\eta_{FP}))}{1-\exp(-D(f\|q\eta_{FP}))}$ and $\delta_B^{\mathcal{B}}=|\mathcal{B}|\cdot\exp\Big(-K\cdot D\big(f\|q\left[(1-p_{\text{lie}})\eta_{FP}+p_{\text{lie}}(1-\eta_{FN})\right]\big)\Big)$, and ρ is defined as in Theorem 1. If Assumption 3 also holds, (8) simplifies to $\|\pi_k^{m,s}-\pi_k^m\|_\infty\leq O(\alpha_{[\frac{k}{2}]})$.

Lemma 2 shows that, after the burn-in period, BR-DTOA avoids falsely labeling honest teams and identifies Byzantine teams with high probability. On this event, Theorem 1 can be applied to bound the belief-estimation errors for honest teams, yielding Theorem 3. Thus, the convergence rate remains the same as in the non-Byzantine case, while the guarantee holds with probability at least $1-\delta_B$ because of possible identification errors. Moreover, $\delta_B\to 0$ as $K\to\infty$, so a longer burn-in period improves the reliability of Byzantine-team identification.

The next result establishes the corresponding honest-TNG convergence guarantee under BR-DTOA.

Theorem 4. *For a Byzantine ZSPTG with a supervisor network $\mathcal{G}_B=(\mathcal{I},\mathcal{H},\mathcal{B},\mathcal{S},ST,\{SN_k\}_{k\geq 0},\{\mathcal{A}^i\}_{i\in\mathcal{I}},\{u^i\}_{i\in\mathcal{I}},p)$, under the conditions of Theorem 2, for any supervisor $s\in\mathcal{S}$, the following inequality holds for BR-DTOA with probability at least $1-\delta_B$*

$$\limsup_{k\to\infty}TNG_{\mathcal{H}}(\pi_k^s)\leq\begin{cases}\tau\log|\mathcal{A}_{\mathcal{H}}|, & \text{for BR-DTOA},\\ \tau\log|\mathcal{A}_{\mathcal{H}}|+|\mathcal{H}|^2\overline{\phi}\Lambda(\delta,\epsilon), & \text{for BR-iDTOA},\end{cases}$$

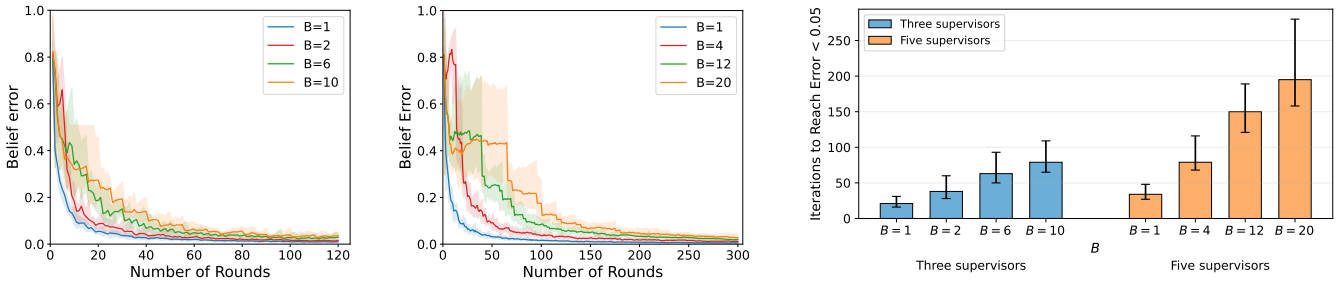
where δ_B is defined as in Theorem 3, and $\mathcal{A}_{\mathcal{H}}=\prod_{m\in\mathcal{H}}\underline{\mathcal{A}}^m$ is the joint action set of honest teams.

Theorem 4 extends the TNG convergence result in Theorem 2 to the Byzantine attack setting. Conditional on the high-probability event characterized by Lemma 2, the honest-team learning dynamics satisfy an honest-TNG bound of the same form as in the non-Byzantine case, but with the gap evaluated only over honest teams. This yields a sharper guarantee than treating all teams uniformly.

Together, the above results show that BR-DTOA identifies Byzantine teams with vanishing error probability. The Byzantine resilience of BR-DTOA is further established by showing that, in the presence of Byzantine teams, it preserves the convergence of belief-estimation errors for honest teams and provides a guarantee on the honest TNG. These results provide an affirmative answer to Problem 2.

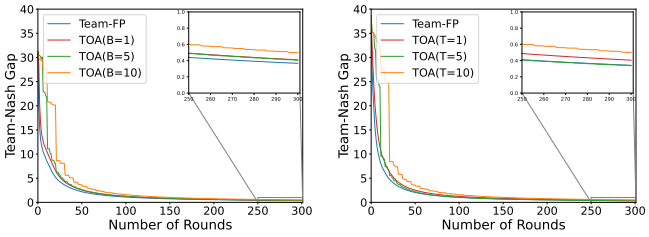
VI. Experimental Validation

In this section, we present numerical experiments to illustrate the theoretical results in Sections IV and V, including the convergence, optimality, and Byzantine resilience of the proposed team-orchestrating algorithms. We also examine the scalability of DTOA. Furthermore, we compare DTOA with two baseline methods and numerically extend DTOA to the Markov decision process (MDP) setting.



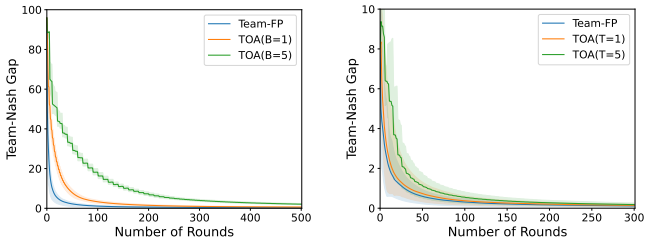
(a) Three supervisors. (b) Five supervisors. (c) Rounds needed for small errors under different B.

Fig. 1: Convergence of belief-estimation errors.



(a) DTOA. (b) Independent DTOA.

Fig. 2: Convergence of TNG.



(a) More supervisors and teams. (b) More agents per team.

Fig. 3: Scalability tests.

Unless otherwise specified, the experiments are conducted on a repeated ZSPTG with a supervisor network consisting of five teams, three supervisors, and two agents in each team. Each agent has the binary action set $\{0, 1\}$. The utility and potential functions are chosen separately for different experiments to reflect the corresponding scenarios while preserving the ZSPTG structure. The supervision relationship between teams and supervisors is given by $\mathcal{T}^1$=$\{1, 2\}$, $\mathcal{T}^2$=$\{3, 4\}$, and $\mathcal{T}^3$=$\{5\}$. We set the step size to $\alpha_k = 1/(k+1)$ and the temperature parameter to $\tau = 0.1$. The shaded regions in the figures indicate the variability across ten independent runs. The code is available at https://github.com/zjt-1229/team_game_with_supervisor_network.

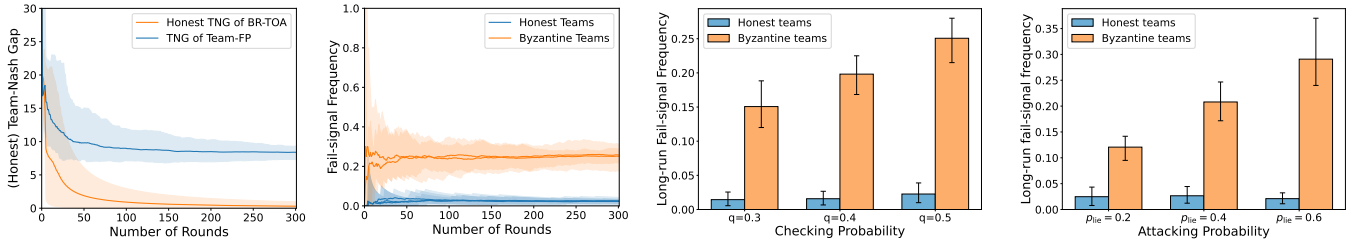
A. Numerical Results for DTOA

We first present numerical results for DTOA. These experiments aim to illustrate the convergence behavior and TNG performance characterized by the theoretical analysis and to examine the scalability of DTOA.

Hypothesis I: convergence of belief-estimation errors. Theorem 1 shows that supervisors' belief-estimation errors converge to zero, with slower convergence when the number of supervisors $|\mathcal{S}|$ or the communication window length B increases. We examine the convergence of supervisors' belief-estimation errors in two instances: one with three supervisors and six teams, and the other with five supervisors and ten teams. In both instances, each supervisor supervises two teams, each team contains two agents, and each agent has the binary action set $\{0, 1\}$. Each agent randomly selects an action at every round, so the experiment isolates supervisor-network belief learning under random action sequences. Fig. 1 reports the results. The errors in Fig. 1b converge more slowly than those in Fig. 1a, illustrating that increasing $|\mathcal{S}|$ slows the convergence rate. Moreover, in both instances, a larger B leads to slower convergence, as reflected in Fig. 1c by the increased number of rounds required to reach a small error level. These observations are consistent with Theorem 1.

Hypothesis II: upper bound on TNG. Theorem 2 implies that DTOA and independent DTOA converge to a near TNE in terms of the TNG, with an asymptotic bound comparable to that of team-FP [6], while sparser supervisor communication can slow convergence. We examine TNG convergence under DTOA and independent DTOA. The independent update probability in independent DTOA is set to $\delta = 0.5$. Fig. 2 shows that both DTOA and independent DTOA eventually attain TNGs comparable to those of the corresponding team-FP benchmarks, which is consistent with Theorem 2. The numerical results also show that the TNG curves under DTOA and independent DTOA converge more slowly than those under team-FP. The slower convergence can be attributed to belief-estimation errors induced by distributed belief learning over the supervisor network. A larger communication window length B further slows convergence in both instances, suggesting that sparser supervisor communication delays the reduction of the TNG in the considered setting.

Scalability tests. To examine the scalability, we consider two larger instances: one with increased numbers of supervisors and teams, and the other with an increased number of agents per team. In the first instance, we consider eight supervisors and fifteen teams with two agents per team; in the second



(a) Honest TNG. (b) Fail-signal frequencies. (c) Effect of q. (d) Effect of p_{lie}.

Fig. 4: Performance of the BR-DTOA.

instance, we consider three supervisors and five teams with five agents per team. Fig. 3 shows that DTOA still ensures a small TNG in both larger instances, consistent with the small-scale experiments.

B. Numerical Results for BR-DTOA

We next present numerical results for BR-DTOA. These experiments illustrate the implications of the theoretical analysis for Byzantine-team identification and honest-TNG convergence.

Hypothesis III: convergence of honest TNG. The theoretical results indicate that BR-DTOA identifies Byzantine teams through accumulated fail-signal frequencies and provides an honest-TNG guarantee under Byzantine attacks. We consider an instance with two Byzantine teams. The parameters are set as $\eta_{FP} = \eta_{FN} = 0.05$, $K = 50$, $f = 0.1$, $q = 0.5$, and $p_{\text{lie}} = 0.5$. To examine parameter sensitivity, we also vary p_{lie} and q separately while keeping all other parameters unchanged. Fig. 4a shows that the honest TNG decreases to a small value under Byzantine attacks, which is consistent with Theorem 4. By contrast, team-FP does not include a Byzantine-resilient mechanism and cannot distinguish Byzantine teams from honest teams. Consequently, the overall TNG under team-FP remains high in the Byzantine attack setting, indicating that team-FP fails to provide an effective performance guarantee for honest teams. Fig. 4b shows that the fail-signal frequencies of honest and Byzantine teams become separated after the burn-in period $K = 50$. Figs. 4c and 4d show that the fail-signal frequencies of Byzantine teams increase as p_{lie} and q increase, respectively. These observations are consistent with Lemma 2.

C. Comparison With Baseline Methods

We compare DTOA with multiplicative weights update (MWU) and smoothed fictitious play (SFP) on a two-team ZSPTG instance with one supervisor assigned to each team. For MWU and SFP, we relax the information constraint by allowing access to full opponent information, whereas DTOA relies on supervisor-based distributed belief learning. Fig. 5 shows the TNG results. In this instance, DTOA attains a lower TNG after sufficiently many rounds, while MWU exhibits persistent oscillations and SFP stabilizes at a higher TNG level. This comparison suggests that DTOA can maintain a smaller TNG even under a more restrictive information structure.

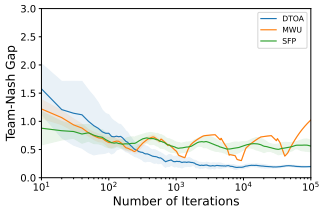


Fig. 5: Comparison.

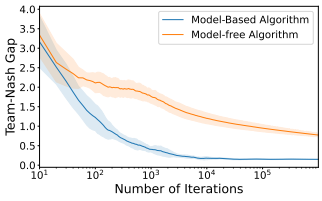


Fig. 6: MDP setting.

D. Extension to an MDP Setting

Although the theoretical analysis in this paper focuses on repeatedly played ZSPTGs, we also examine the proposed supervisor-network learning framework in an MDP setting. The experiment considers three teams and three supervisors, with each team supervised by one supervisor. Fig. 6 shows the results. Both the model-based and model-free variants reduce the TNG through learning, suggesting that the proposed framework can be applied numerically in an MDP setting. Compared with the corresponding results in [6], the reduction in the TNG is slower, which may be attributed to the additional belief-estimation errors induced by the supervisor network.

VII. Conclusions

In this paper, we studied team-orchestrating learning in repeatedly played ZSPTGs with a supervisor network under distributed belief information. We proposed the DTOA, which combines team-FP with distributed belief learning over a supervisor network, established the convergence of supervisors' belief-estimation errors, and showed that the induced learning dynamics converged to a near TNE with a small TNG. We further considered a Byzantine attack setting, where Byzantine teams could misreport their joint actions and developed the BR-DTOA by integrating DTOA with a supervisor-based identification mechanism. For BR-DTOA, we established the convergence of supervisors' belief-estimation errors for honest teams and derived an honest TNG guarantee. Numerical simulations illustrated the convergence of the proposed algorithms and the effectiveness of the Byzantine-resilient mechanism.

References

[1] A. Bagchi and T. Basar, "Team decision theory for linear continuous-time systems," *IEEE Transactions on Automatic Control*, vol. 25, no. 6, pp. 1154–1161, 1980.

[2] A. A. Malikopoulos, "On team decision problems with nonclassical information structures," *IEEE Transactions on Automatic Control*, vol. 68, no. 7, pp. 3915–3930, 2023.

[3] R. Cao and Y. Zhao, "Distributed average tracking over directed communication networks: A nonsmooth surplus approach," *IEEE Transactions on Automatic Control*, vol. 71, no. 4, pp. 2450–2465, 2026.

[4] J. Wang, D. W. C. Ho, X. Jin, F. Li, and Y. Tang, "Multi-agent target-attacker-defender differential games with anomalous defenders under limited perception," *IEEE Transactions on Automatic Control*, pp. 1–16, 2025.

[5] J. Kim, T. R. Palfrey, and J. R. Zeidel, "Games played by teams of players," *American Economic Journal: Microeconomics*, vol. 14, no. 4, pp. 122–157, 2022.

[6] A. Dönmez, Y. Arslantaş, and M. Sayin, "Team-fictitious play for reaching team-Nash equilibrium in multi-team games," *Advances in Neural Information Processing Systems*, vol. 37, pp. 111 515–111 543, 2024.

[7] J. Xiao, C. Qiu, Y. Xu, J. Zhang, S. Qi, and X. Wang, "Solving equilibrium for adversarial team games utilizing fictitious team play with refined team plans," *Expert Systems with Applications*, p. 129496, 2025.

[8] X. Feng, J. Huang, and Z. Qiu, "Mixed social optima and Nash equilibrium in linear-quadratic-gaussian mean-field system," *IEEE Transactions on Automatic Control*, vol. 67, no. 12, pp. 6858–6865, 2021.

[9] X. Zeng, J. Chen, S. Liang, and Y. Hong, "Generalized Nash equilibrium seeking strategy for distributed non-smooth multi-cluster game," *Automatica*, vol. 103, pp. 20–26, 2019.

[10] T. Ma, Z. Deng, and C. Hu, "A fully distributed Nash equilibrium seeking algorithm for N-coalition games of euler–lagrange players," *IEEE Transactions on Control of Network Systems*, vol. 10, no. 1, pp. 205–213, 2022.

[11] H. Zhang, G. Chen, and Y. Hong, "Distributed algorithm for continuous-type Bayesian Nash equilibrium in subnetwork zero-sum games," *IEEE Transactions on Control of Network Systems*, vol. 11, no. 2, pp. 915–927, 2023.

[12] Z. Deng and J. Luo, "Distributed algorithm for nonsmooth multi-coalition games and its application in electricity markets," *Automatica*, vol. 161, p. 111494, 2024.

[13] Y. Shi and B. Zhang, "Multi-agent reinforcement learning in Cournot games," in *2020 59th IEEE Conference on Decision and Control*. IEEE, 2020, pp. 3561–3566.

[14] G. Alcantara-Jiménez and J. B. Clempner, "Repeated Stackelberg security games: Learning with incomplete state information," *Reliability Engineering & System Safety*, vol. 195, p. 106695, 2020.

[15] J. W. Neal, Z. P. Neal, and B. Brutzman, "Defining brokers, intermediaries, and boundary spanners: a systematic review," *Evidence & policy*, vol. 18, no. 1, pp. 7–24, 2022.

[16] F. Boutcher, W. Berta, R. Urquhart, and A. R. Gagliardi, "The roles, activities and impacts of middle managers who function as knowledge brokers to improve care delivery and outcomes in healthcare organizations: a critical interpretive synthesis," *BMC Health Services Research*, vol. 22, no. 1, p. 11, 2022.

[17] M. Huang, R. P. Malhamé, and P. E. Caines, "Large population stochastic dynamic games: closed-loop Mckean-Vlasov systems and the Nash certainty equivalence principle," *Communication Information Systems*, vol. 6, no. 1, pp. 221–252, 2006.

[18] B. Moll and L. Ryzhik, "Mean field games without rational expectations," *Communications in Contemporary Mathematics*, p. 2640007, 2026.

[19] Z. Wu, M. Laurière, S. J. C. Chua, M. Geist, O. Pietquin, and A. Mehta, "Population-aware online mirror descent for mean-field games by deep reinforcement learning," in *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, 2024, pp. 2561–2563.

[20] S. Sanjari and S. Yüksel, "Optimal solutions to infinite-player stochastic teams and mean-field teams," *IEEE Transactions on Automatic Control*, vol. 66, no. 3, pp. 1071–1086, 2020.

[21] X. Feng, J. Huang, Y. Jia, and M. Yu, "Mean-field team in backward linear-quadratic control problems with model uncertainty," *Science China Information Sciences*, vol. 68, no. 11, p. 210203, 2025.

[22] S. Aggarwal, M. A. u. Zaman, M. Bastopcu, and T. Başar, "Semantic communication in multiteam dynamic games: A mean field perspective," *IEEE Transactions on Automatic Control*, vol. 71, no. 1, pp. 49–64, 2026.

[23] J. Huang, Z. Qiu, S. Wang, and Z. Wu, "Linear quadratic mean-field game-team analysis: A mixed coalition approach," *Automatica*, vol. 159, p. 111358, 2024.

[24] B. Jeloka, Y. Guan, and P. Tsiotras, "Learning large-scale competitive team behaviors with mean-field interactions," in *The Seventeenth Workshop on Adaptive and Learning Agents*, 2025.

[25] J. H. Dreze and J. Greenberg, "Hedonic coalitions: Optimality and stability," *Econometrica: Journal of the Econometric Society*, pp. 987–1003, 1980.

[26] J. Hajduková, "Coalition formation games: A survey," *International Game Theory Review*, vol. 8, no. 04, pp. 613–641, 2006.

[27] C. Wang, M. Moharrami, K. Jin, D. Kempe, P. J. Brantingham, and M. Liu, "Structural stability of a family of group formation games," in *2021 60th IEEE Conference on Decision and Control*, 2021, pp. 3080–3085.

[28] W. Saad, Z. Han, T. Basar, M. Debbah, and A. Hjorungnes, "Hedonic coalition formation for distributed task allocation among wireless agents," *IEEE Transactions on Mobile Computing*, vol. 10, no. 9, pp. 1327–1344, 2010.

[29] Q. Zhu and T. Basar, "Game-theoretic methods for robustness, security, and resilience of cyberphysical control systems: games-in-games principle for optimal cross-layer resilient control systems," *IEEE Control Systems Magazine*, vol. 35, no. 1, pp. 46–65, 2015.

[30] G. Chen, Y. Ming, Y. Hong, and P. Yi, "Distributed algorithm for ε-generalized Nash equilibria with uncertain coupled constraints," *Automatica*, vol. 123, p. 109313, 2021.

[31] G. Chen, G. Xu, F. He, D. Tao, T. Parisini, and K. H.

Johansson, "Inverse learning of black-box aggregator for robust Nash equilibrium," *IEEE Transactions on Automatic Control*, 2025.
[32] J. Wang, X. Deng, J. Guo, and Z. Zeng, "Resilient consensus control for multi-agent systems: A comparative survey," *Sensors*, vol. 23, no. 6, p. 2904, 2023.
[33] H. J. LeBlanc, H. Zhang, X. Koutsoukos, and S. Sundaram, "Resilient asymptotic consensus in robust networks," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 4, pp. 766–781, 2013.
[34] L. Yuan and H. Ishii, "Secure consensus with distributed detection via two-hop communication," *Automatica*, vol. 131, p. 109775, 2021.
[35] J. Yan, C. Deng, and C. Wen, "Resilient output regulation in heterogeneous networked systems under Byzantine agents," *Automatica*, vol. 133, p. 109872, 2021.
[36] S. Li, J. Guo, J. Xiu, R. Xu, X. Yu, J. Wang, A. Liu, Y. Yang, and X. Liu, "Byzantine robust cooperative multi-agent reinforcement learning as a Bayesian game," in *The 12th International Conference on Learning Representations*, 2024.
[37] L. Moreau, "Stability of multiagent systems with time-dependent communication links," *IEEE Transactions on automatic control*, vol. 50, no. 2, pp. 169–182, 2005.
[38] J. Li, W. Abbas, M. Shabbir, and X. Koutsoukos, "Byzantine resilient distributed learning in multirobot systems," *IEEE Transactions on Robotics*, vol. 38, no. 6, pp. 3550–3563, 2022.
[39] A. Haeberlen, P. Kouznetsov, and P. Druschel, "The case for Byzantine fault detection," in *Proceedings of the 2nd Conference on Hot Topics in System Dependability*, 2006, pp. 5–5.
[40] L. E. Blume, "The statistical mechanics of strategic interaction," *Games and economic behavior*, vol. 5, no. 3, pp. 387–424, 1993.
[41] J. R. Marden and J. S. Shamma, "Revisiting log-linear learning: Asynchrony, completeness and payoff-based implementation," *Games and Economic Behavior*, vol. 75, no. 2, pp. 788–808, 2012.
[42] M. J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge university press, 2019, vol. 48.

APPENDIX

A. Proof of Theorem 1

We prove Theorem 1 by first establishing a technical result on the supervisor networks $\{SN_k\}_{k\geq 0}$ and then proving the convergence of the belief-estimation errors.

Lemma 3. *Let $W_k \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ denote the gossip matrix associated with the supervisor network at round k, where*

$$W_k(s,s') = \begin{cases} \frac{SN_k(s,s')}{\|SN_k(s,\cdot)\|_1}, & \text{if } \|SN_k(s,\cdot)\|_1 > 0, \\ 1, & \text{if } \|SN_k(s,\cdot)\|_1 = 0 \text{ and } s' = s, \\ 0, & \text{if } \|SN_k(s,\cdot)\|_1 = 0 \text{ and } s' \neq s. \end{cases}$$

If Assumption 2 holds, there exists some $L \in \mathbb{N}_+$ such that, for any supervisors s, s' and any $k \geq 0$, the following statement holds: $\big[W_{k+L}\cdots W_k\big](s,s') \geq \left[\frac{1}{|\mathcal{S}|}\right]^L$.

Proof. We consider a setting in which $W_k(s,s')$ represents the probability of s reaching s' at round k; therefore, W_k is a transition matrix. A path on the dynamic supervisor network $\{SN_k\}_{k\geq 0}$ is defined as $\mathcal{P} = (X_k \in \mathcal{S})_{k\geq 0}$ such that $W_k(X_k, X_{k+1}) > 0$. Given $k \geq 0$ and $s = X_k$, there exists $n \in \mathbb{N}$ such that $nB \leq k < (n+1)B$, and from the definition of W_k, we have

$$\big[W_{(n+1)B}\cdots W_k\big](s,s) \geq \left[\frac{1}{|\mathcal{S}|}\right]^{(n+1)B-k} \geq \left[\frac{1}{|\mathcal{S}|}\right]^{B}.$$

Let $\mathcal{N}^s(l)$ denote the set of supervisors that can be reached with positive probability at time step $(n+1)B+l$. It follows that $\mathcal{N}^s(l) \subseteq \mathcal{N}^s(l+1)$ since $W_l(s',s') > 0$ for all $s' \in \mathcal{N}^s(l)$. If $\mathcal{N}^s(n'B) \neq \mathcal{S}$, there must exist $n'B \leq l' < (n'+1)B$, $s \in \mathcal{N}^s(n'B)$ and $s' \notin \mathcal{N}^s(n'B)$ such that $\{s,s'\} \in E_l$ because the graph $G_{(n')}$ is connected (Assumption 2). Hence, we obtain $\mathcal{N}^s(n'B) \subsetneq \mathcal{N}^s(l') \subseteq \mathcal{N}^s\big((n'+1)B\big)$. Then we can conclude that $\mathcal{N}^s(|\mathcal{S}|B) = \mathcal{S}$. Let $L = (|\mathcal{S}|+1)B$. Then we have $\mathbb{P}\big(X_{k+L} = s' \big| X_k = s\big) > 0$. By the definition of $W_{l'}$, $W_l(s,s') > 0$ implies $W_{l'}(s,s') \geq \frac{1}{|\mathcal{S}|}$. It follows that that there exists at least one path satisfying

$$\mathbb{P}\big(X_{k+L}=s' \big| X_k=s\big) > 0,$$

namely, $\mathcal{P}_{s,s',L}=(X_k=s,\ldots,X_{l'}=s, X_{l'+1}=s',\ldots,X_{k+L}=s')$. Thus, we have

$$\mathbb{P}\big(X_{k+L} = s' \big| X_k = s\big) \geq \mathbb{P}\left(\mathcal{P}_{s,s',L}\right) \geq \left[\frac{1}{|\mathcal{S}|}\right]^L.$$

This completes the proof. □

Now we can prove Theorem 1. We first define the error of $\pi_k^{m,s}$ as the difference between $\pi_k^{m,s}$ and π_k^m:

$$e_k^{m,s} = \pi_k^{m,s} - \pi_k^m, \quad \forall m \in \mathcal{T},\ \forall s \in \mathcal{S}, \tag{9}$$

and then we analyze the decay rate of $e_k^{m,s}$. Given a team m, we introduce the following notation: $D^m = \operatorname{diag}\big(ST(m,1),\ldots,ST(m,|\mathcal{S}|)\big)$. Then we obtain

$$\begin{pmatrix} \pi_{k+1}^{m,1} \\ \vdots \\ \pi_{k+1}^{m,|\mathcal{S}|} \end{pmatrix} = D^m \begin{pmatrix} \pi_{k+1}^{m} \\ \vdots \\ \pi_{k+1}^{m} \end{pmatrix} + (I - D^m) W_k \begin{pmatrix} \pi_{k}^{m,1} \\ \vdots \\ \pi_{k}^{m,|\mathcal{S}|} \end{pmatrix}. \tag{10}$$

Since $\|W_k(s,\cdot)\|_1 > 0$, combining (9) and (10), we obtain

$$\begin{aligned} \begin{pmatrix} e_{k+1}^{m,1} \\ \vdots \\ e_{k+1}^{m,|\mathcal{S}|} \end{pmatrix} &= \begin{pmatrix} \pi_{k+1}^{m,1} \\ \vdots \\ \pi_{k+1}^{m,|\mathcal{S}|} \end{pmatrix} - \begin{pmatrix} \pi_{k+1}^{m} \\ \vdots \\ \pi_{k+1}^{m} \end{pmatrix} \\ &= (I - D^m) W_k \left[\begin{pmatrix} e_{k}^{m,1} \\ \vdots \\ e_{k}^{m,|\mathcal{S}|} \end{pmatrix} + \begin{pmatrix} \pi_{k}^{m} \\ \vdots \\ \pi_{k}^{m} \end{pmatrix} \right] + (D^m - I) \begin{pmatrix} \pi_{k+1}^{m} \\ \vdots \\ \pi_{k+1}^{m} \end{pmatrix} \\ &= (I - D^m) W_k \begin{pmatrix} e_{k}^{m,1} \\ \vdots \\ e_{k}^{m,|\mathcal{S}|} \end{pmatrix} + (I - D^m) \begin{pmatrix} \pi_{k}^{m} - \pi_{k+1}^{m} \\ \vdots \\ \pi_{k}^{m} - \pi_{k+1}^{m} \end{pmatrix}. \end{aligned} \tag{11}$$

Let $e_k^m \triangleq \begin{pmatrix} e_k^{m,1} \\ \vdots \\ e_k^{m,|\mathcal{S}|} \end{pmatrix}$, and $b_{k+1}^m \triangleq (I - D^m)\begin{pmatrix} \pi_k^m - \pi_{k+1}^m \\ \vdots \\ \pi_k^m - \pi_{k+1}^m \end{pmatrix}$ and $B_k^m \triangleq (I - D^m)W_k$. Then we rewrite (11) as $e_{k+1}^m = B_k^m e_k^m + b_{k+1}^m$. Using the above recursion, we obtain the expression for e_k^m:

$$e_k^m = \mathcal{B}^m(k,0)e_0^m + \sum_{l=0}^{k-1} \mathcal{B}^m(k,l+1)b_l^m,$$

where $\mathcal{B}^m(k,l) = B_{k-1}^m \cdots B_l^m$. By reordering the indices, W_k^m and D^m can be written as

$$W_k^m = \begin{pmatrix} W_{k,U^mU^m} & W_{k,U^mO^m} \\ W_{k,O^mU^m} & W_{k,O^mO^m} \end{pmatrix}, \ \tilde{D}^m = \begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix}.$$

Then we have

$$B_k^m = \begin{pmatrix} W_{k,U^mU^m} & W_{k,U^mO^m} \\ 0 & 0 \end{pmatrix},$$
$$\mathcal{B}^m(k,l) = \begin{pmatrix} B_{UU}^m(k,l) & B_{UO}^m(k,l) \\ 0 & 0 \end{pmatrix},$$

where we denote $B_{UU}^m(k,l) = W_{k-1,U^mU^m} \cdots W_{l,U^mU^m}$ and $B_{UO}^m(k,l) = W_{k-1,U^mO^m} \cdots W_{l,U^mO^m}$. Let $W^m(k,l) = W_{k-1,U^mU^m} \cdots W_{l,U^mU^m}$ and $\|W^m(k,l)\|_\infty = R^m(k,l) = \max_{s\in U^m} \sum_{s'\in U^m} [W^m(k,l)](s,s')$. By Lemma 3, we have $R^m(l+L,l) \leq 1-\alpha$ with $\alpha = \left[\frac{1}{|\mathcal{S}|}\right]^L$. Therefore, we can conclude $R^m(k,l) \leq \alpha^{[\frac{k-l}{L}]} \leq C'\rho^{k-l}$, where $C' = (1-\alpha)^{-1} > 1$ and $\rho = (1-\alpha)^{\frac{1}{L}} < 1$. For any $x = (x_{U^m}, x_{O^m})^T \in \mathbb{R}^m$, we get $B_l^m x = (y_{U^m}, 0)^T$. Then we obtain $\mathcal{B}^m(k,l)x = (W^m(k,l+1)y_{U^m}, 0)^T$. Thus, we have

$$\begin{aligned} \|\mathcal{B}^m(k,l)x\|_\infty &\leq \|W^m(k,l+1)\|_\infty \|y_{U^m}\|_\infty \\ &\leq C'\rho^{k-l-1}\|y_{U^m}\|_\infty. \end{aligned} \tag{12}$$

From $y_{U^m} = W_{l,U^mU^m}x_{U^m} + W_{l,U^mO^m}x_{O^m} \triangleq M(l)x$ and $\|M(l)\|_\infty \leq 1$, we can say that

$$\|y_{U^m}\|_\infty \leq \|x\|_\infty \tag{13}$$

Combine (12) and (13) we can conclude that $\|\mathcal{B}^m(k,l)\|_\infty \leq C''\rho^{k-l}$. Hence, we obtain

$$\begin{aligned} \|e_k^m\|_\infty &= \|\mathcal{B}^m(k,0)e_0^m + \sum_{l=0}^{k-1} \mathcal{B}^m(k,l+1)b_l^m\|_\infty \\ &\leq C''\rho^k\|e_0^m\|_\infty + \sum_{l=1}^{k-1} C''\rho^{k-l-1}\|b_l^m\|_\infty \\ &= C''\rho^k\|e_0^m\|_\infty + \sum_{l=1}^{[\frac{k}{2}]-1} C''\rho^{k-l-1}\|b_l^m\|_\infty + \sum_{l=[\frac{k}{2}]}^{k-1} C''\rho^{k-l-1}\|b_l^m\|_\infty \\ &\leq C''\rho^k\|e_0^m\|_\infty + C''\sum_{l=k-[\frac{k}{2}]}^{\infty} \rho^l \sup_{l\geq 0}\|b_l^m\|_\infty + C''\sum_{l=0}^{\infty}\rho^l \sup_{l\geq[\frac{k}{2}]}\|b_l^m\|_\infty. \end{aligned}$$

Then, we get

$$\begin{aligned} &\|e_k^m\|_\infty \\ &\leq C''\rho^k\|e_0^m\|_\infty + C''\sup_{l\geq 0}\|b_l^m\|_\infty \frac{\rho^{k-[\frac{k}{2}]}}{1-\rho} + C''\sup_{l\geq[\frac{k}{2}]}\|b_l^m\|_\infty \frac{1}{1-\rho} \\ &\leq C''\rho^k\|e_0^m\|_\infty + C''\sup_{l\geq 0}\|b_l^m\|_\infty \frac{\rho^{\frac{k}{2}}}{1-\rho} + C''\sup_{l\geq[\frac{k}{2}]}\|b_l^m\|_\infty \frac{1}{1-\rho} \\ &= C''\rho^k\|e_0^m\|_\infty + C''\sup_{l\geq 0}\alpha_{l-1}\|(I-D^m)(\underline{a}_{l-1}^m - \pi_{l-1}^m)\|_\infty \frac{\rho^{\frac{k}{2}}}{1-\rho} \\ &\quad + \frac{C''}{1-\rho}\sup_{l\geq[\frac{k}{2}]}\alpha_{l-1}\|(I-D^m)(\underline{a}_{l-1}^m - \pi_{l-1}^m)\|_\infty \\ &\leq C''\rho^k\|e_0^m\|_\infty + 2C''\alpha_0 \cdot \frac{\rho^{\frac{k}{2}}}{1-\rho} + \frac{2C''}{1-\rho}\cdot\alpha_{[\frac{k}{2}]-1} \\ &= O(\rho^{[\frac{k}{2}]}) + O(\alpha_{[\frac{k}{2}]}). \end{aligned}$$

Finally, we can conclude $\|\pi_k^{m,s} - \pi_k^m\|_\infty \leq O(\rho^{[\frac{k}{2}]}) + O(\alpha_{[\frac{k}{2}]})$ and complete the proof of Theorem 1.

B. Proof of Theorem 2

The TNG of Algorithm 1 is closely related to the true belief π_k^m. The proof of Theorem 2 proceeds in two steps: 1) constructing a reference scenario where true beliefs are frozen and showing that members within a team can learn to team up in spite of the errors caused by the supervisor networks; 2) bounding the TNG by leveraging stochastic differential inclusion approximations.

Note that, given a team strategy profile $\pi = \{\pi^m \in \Delta(\underline{\mathcal{A}}^m)\}_{m\in\mathcal{T}}$, an agent $i \in \mathcal{I}^m$ and any $a^{-i} \in \prod_{j\in\mathcal{I}^m, j\neq i}\mathcal{A}^j$, the ZSPTG property in (1) implies

$$br_\tau\left(u^i(\cdot, a^{-i}, \pi^{-m})\right) = br_\tau\left(\phi^m(\cdot, a^{-i}, \pi^{-m})\right),$$

where $\pi^{-m} \triangleq \{\pi^l\}_{l\neq m}$. We divide the learning process into epochs of length T. Then, by accumulating the true belief update (5) from nT to $(n+1)T-1$, we obtain

$$\begin{aligned} &\pi_{(n+1)T}^m \\ &= \left[\prod_{k=nT}^{(n+1)T-1}(1-\alpha_k)\right]\pi_{nT}^m + \sum_{k=nT}^{(n+1)T-1}\alpha_k\left[\prod_{l=k+1}^{(n+1)T-1}(1-\alpha_l)\right]\underline{a}_k^m, \end{aligned} \tag{14}$$

where $n = 0, 1, \ldots$ denotes the epoch index. Let $\pi_{(n)}^m \triangleq \pi_{nT}^m$ and $\pi_{(n)}^{m,s} \triangleq \pi_{nT}^{m,s}$ denote the true belief and the actual belief about team m after learning n epochs. Furthermore, we define

$$\beta_k \triangleq \alpha_k \prod_{l=k+1}^{(n+1)T-1}(1-\alpha_l) \ \text{ and } \ \beta_{(n)} \triangleq \sum_{k=nT}^{(n+1)T-1}\beta_k. \tag{15}$$

Then, we rewrite (14) as

$$\pi_{(n+1)T}^m = \left(1-\beta_{(n)}\right)\cdot\pi_{(n)}^m + \beta_{(n)}\left[\sum_{k=nT}^{(n+1)T-1}\frac{\beta_k}{\beta_{(n)}}\cdot\underline{a}_k^m\right]. \tag{16}$$

By Assumption 3 and Lemma 5.4 of [6], we obtain

$$\beta_{(n)} \to 0 \text{ as } n\to\infty, \ \sum_{n=0}^{\infty}\beta_{(n)} = \infty \text{ and } \sum_{n=0}^{\infty}\beta_{(n)}^2 < \infty. \tag{17}$$

Let $\mathcal{F}_{(n)}$ denote the filtration generated by the σ-algebra $\sigma(\mathcal{S}, ST, \{SN_k\}_{k\geq 0}, A_0, \ldots, A_{nT-1})$, where A_t denotes the joint action profile of all teams at time step t, i.e., $A_t = (\underline{a}_t^1, \ldots, \underline{a}_t^{|\mathcal{T}|})$. Note that $\pi_{(n)}^m$ and $\pi_{(n)}^{m,s}$ are $\mathcal{F}_{(n)}$-measurable. The joint action distributions of team m based on the true beliefs, i.e., in the ideal scenario, are defined for DTOA and independent DTOA at time k in epoch n as

$$\text{team-}\nu_{(n),k}^m \triangleq \mathbb{E}\left[\underline{a}_k^m \middle| \mathcal{F}_{(n)}\right], \ \text{indp-}\nu_{(n),k}^m \triangleq \mathbb{E}\left[\underline{a}_k^m \middle| \mathcal{F}_{(n)}\right]$$

for $k = nT, \ldots, (n+1)T-1$. We also define the distributions of the joint action of team m based on the actual beliefs at time step k in epoch n for DTOA and independent DTOA as

$$\text{a-team-}\nu_{(n),k}^m \triangleq \mathbb{E}\left[\underline{a}_k^m \middle| \mathcal{F}_{(n)}\right], \text{a-indp-}\nu_{(n),k}^m \triangleq \mathbb{E}\left[\underline{a}_k^m \middle| \mathcal{F}_{(n)}\right]$$

for $k = nT, \ldots, (n+1)T-1$. For notational simplicity, we use $\nu_{(n),k}^m$ to represent team-$\nu_{(n),k}^m$ for DTOA and indp-$\nu_{(n),k}^m$ for independent DTOA, and use a-$\nu_{(n),k}^m$ to represent a-team-$\nu_{(n),k}^m$ for DTOA and a-indp-$\nu_{(n),k}^m$ for independent DTOA. Then, we rewrite (16) in the form of stochastic approximation:

$$\begin{aligned}&\pi_{(n+1)}^m\\ =&\left(1-\beta_{(n)}\right)\pi_{(n)}^m + \beta_{(n)}\left[\sum_{k=nT}^{(n+1)T-1} \frac{\beta_k}{\beta_{(n)}} \cdot \nu_{(n),k}^m + \text{a-}\omega_{(n+1)}^m\right], \quad (18)\end{aligned}$$

where a-$\omega_{(n+1)}^m$ is defined as

$$\begin{aligned}\text{a-}\omega_{(n+1)}^m \triangleq& \sum_{k=nT}^{(n+1)T-1} \frac{\beta_k}{\beta_{(n)}}\left[\underline{a}_k^m - \nu_{(n),k}^m\right]\\ =&\sum_{k=nT}^{(n+1)T-1} \frac{\beta_k}{\beta_{(n)}}\left[\underline{a}_k^m - \text{a-}\nu_{(n),k}^m\right] + \sum_{k=nT}^{(n+1)T-1} \frac{\beta_k}{\beta_{(n)}}\left[\text{a-}\nu_{(n),k}^m - \nu_{(n),k}^m\right]\\ =&\omega_{(n+1)}^m + \text{a-}e_{(n+1)}^m, \quad (19)\end{aligned}$$

where $\omega_{(n+1)}^m \triangleq \sum_{k=nT}^{(n+1)T-1} \frac{\beta_k}{\beta_{(n)}}\left[\underline{a}_k^m - \text{a-}\nu_{(n),k}^m\right]$ is a martingale difference sequence and we define the actual error as $\text{a-}e_{(n+1)}^m \triangleq \sum_{k=nT}^{(n+1)T-1} \frac{\beta_k}{\beta_{(n)}}\left[\text{a-}\nu_{(n),k}^m - \nu_{(n),k}^m\right]$. It follows that $\mathbb{E}\left[\omega_{(n+1)}^m \middle| \mathcal{F}_{(n)}\right] = 0$. Then we are ready to prove Lemma 1. Recall the result in Lemma 1: the difference between a-$\nu_{(n),k}^m$ and $\nu_{(n),k}^m$ can be bounded as $\|\text{a-}\nu_{(n),k}^m - \nu_{(n),k}^m\|_1 \leq C\delta_{(n)}$, where $\delta_{(n)} \to 0$ as $n \to \infty$.

Proof. We first define the joint-action process from the start of epoch n. We denote the joint-action profiles of all teams by $\{\omega_k\}_{k\geq nT} \triangleq \left\{(\underline{a}_k^m)_{m\in\mathcal{T}} \middle| \mathcal{F}_{(n)}\right\}$, and view each joint-action profile ω_k as a state. The transitions between states depend only on the beliefs about all teams, ST, and $\{SN_k\}_{k\geq 0}$. Let $P_{(n),k}$ denote the transition probabilities between states and let a-$\nu_{(n),k}$ denote the state distribution at time step k in the actual scenario. Let $P_{(n)}^*$ denote the transition probabilities between states and let $\nu_{(n),k}$ denote the state distribution at time step k in the ideal scenario. Then, the joint-action distributions of team m at time step k within epoch n in the actual and ideal scenarios are defined as

$$a\text{-}\nu_{(n),k}^m(a) = \sum_{\omega: a^m = a} a\text{-}\nu_{(n),k}(\omega), \quad (20)$$

$$\nu_{(n),k}^m(a) = \sum_{\omega: a^m = a} \nu_{(n),k}(\omega). \quad (21)$$

Since ϕ^m is linear in π, the smoothed best response is Lipschitz in π. Thus, there exists $L_\pi > 0$ such that

$$\begin{aligned}&\left\|P_{(n),k}\left(\cdot|\omega_k,\ldots,\omega_{nT};\pi_k^{\mathcal{S}}\right) - P_{(n),k}^*\left(\cdot|\omega_k,\ldots,\omega_{nT};\pi_k\right)\right\|_{TV}\\ \leq& L_\pi \sum_{m\in\mathcal{T}} \max_{s\in\mathcal{S}} \|\pi_k^{m,s} - \pi_k^m\|_1,\end{aligned}$$

where $\pi_k^{\mathcal{S}}$ denotes the belief estimates provided by the supervisor network, based on which agents choose their actions. Consider the error $D_k \triangleq \|\text{a-}\nu_{(n),k} - \nu_{(n),k}\|_1$. Since $\text{a-}\nu_{(n),k+1} = \text{a-}\nu_{(n),k}P_{(n),k}$ and $\nu_{(n),k+1} = \nu_{(n),k}P_{(n),k}^*$, we have the decomposition

$$\begin{aligned}D_{k+1} =&\|\text{a-}\nu_{(n),k}P_{(n),k} - \nu_{(n),k}P_{(n),k}^*\|_1\\ =&\left\|\left(\text{a-}\nu_{(n),k} - \nu_{(n),k}\right)P_{(n),k}^* + \text{a-}\nu_{(n),k}\left(P_{(n),k} - P_{(n),k}^*\right)\right\|_1.\end{aligned}$$

Using the above decomposition and the fact that $\|xP\|_1 \leq \|x\|_1$, we obtain

$$\begin{aligned}D_{k+1} \leq& \left\|\left(\text{a-}\nu_{(n),k} - \nu_{(n),k}\right)P_{(n),k}^*\right\|_1 + \left\|\text{a-}\nu_{(n),k}\left(P_{(n),k} - P_{(n),k}^*\right)\right\|_1\\ \leq& D_k + \left\|\text{a-}\nu_{(n),k}\left(P_{(n),k} - P_{(n),k}^*\right)\right\|_1\\ \leq& D_k + 2\sup_\omega \|P_{(n),k}(\omega,\cdot) - P_{(n),k}^*(\omega,\cdot)\|_{TV}\\ \leq& D_k + 2L_\pi \sum_{m\in\mathcal{T}} \max_{s\in\mathcal{S}} \|\pi_k^{m,s} - \pi_k^m\|_1.\end{aligned}$$

By the definition of the maximum belief-estimation error in epoch n,

$$\delta_{(n)} = \max_{k\in[nT-1,(n+1)T-1]} \sum_{m\in\mathcal{T}} \max_{s\in\mathcal{S}} \|\pi_k^{m,s} - \pi_k^m\|_1,$$

we have $D_k \leq D_{nT} + 2TL_\pi\delta_{(n)}$. Moreover, $D_{nT} \leq \sum_{m\in\mathcal{T}} C'\|\pi_{nT-1}^{m,s} - \pi_{nT-1}^m\|_1$. Thus we have $D_k \leq C\delta_{(n)}$, where $C = C' + 2TL_\pi$. From (20) and (21), we have

$$\|\text{a-}\nu_{(n),k}^m - \nu_{(n),k}^m\|_1 \leq \|\text{a-}\nu_{(n),k} - \nu_{(n),k}\|_1 = D_k \leq C\delta_{(n)}. \quad (22)$$

This completes the proof of Lemma 1. □

Finally, we can get

$$\|\text{a-}e_{(n+1)}^m\|_1 \leq \sum_{k=nT}^{(n+1)T-1} \frac{\beta_k}{\beta_{(n)}} C\delta_{(n)} = C\delta_{(n)}.$$

Furthermore, $\delta_{(n)} \to 0$ as $n \to \infty$ by Theorem 1.

Now we introduce a reference scenario to facilitate the analysis. Let $\hat{\pi}_t^m$ denote the belief for team m at time step t in the reference scenario. In the reference scenario, $\hat{\pi}_t^m$ is only updated at the end of each epoch. In other words, for all $nT \leq t \leq (n+1)T-1$ and $m \in \mathcal{T}$, we have $\hat{\pi}_t^m = \pi_{(n)}^m$. Since the beliefs are fixed, team-FP dynamics reduce to log-linear learning in the reference scenario. Let $\hat{a}_{(n),k}^m$ denote the joint action of team m at time step k under the fixed beliefs in the reference scenario.

Due to the nature of log-linear learning, $\{\hat{a}_{(n),k}^m\}_{k=nT}^\infty$ forms a homogeneous Markov chain (MC). In contrast, the actual action profiles $\{\underline{a}_k^m\}_{k=nT}^\infty$ do not. We define the stationary distributions of the MC in the reference scenario as $\check{\nu}_{(n),*}^m$ and $\hat{\nu}_{(n),*}^m$ for team-FP and independent team-FP, respectively. By

[40, 41], it follows that $\check{\nu}^m_{(n),*} = br_\tau\left(\phi^m(\cdot,\pi^{-m}_{(n)})\right)$. Therefore, we write the true belief update (18) as

$$\pi^m_{(n+1)} = \left(1-\beta_{(n)}\right)\pi^m_{(n)} + \beta_{(n)}\Big[br_\tau\Big(\phi^m(\cdot,\pi^{-m}_{(n)})\Big) + \omega^m_{(n+1)} + e^m_{(n)} + a\text{-}e^m_{(n+1)}\Big]. \quad (23)$$

The error for Team-FP is

$$\begin{aligned}\text{team-}e^m_{(n)} &= \sum_{k=nT}^{(n+1)T-1}\frac{\beta_k}{\beta_{(n)}}\cdot\text{team-}\nu^m_{(n),k} - \check{\nu}^m_{(n),*}\\ &= \sum_{k=nT}^{(n+1)T-1}\frac{\beta_k}{\beta_{(n)}}\left(\text{team-}\nu^m_{(n),k} - \check{\nu}^m_{(n),*}\right). \quad (24)\end{aligned}$$

The error for independent team-FP can be decomposed as $\text{indp-}e^m_{(n)} \triangleq \hat{e}^m_{(n)} + \check{e}^m_{(n)}$, where

$$\begin{aligned}\hat{e}^m_{(n)} &\triangleq \sum_{k=nT}^{(n+1)T-1}\frac{\beta_k}{\beta_{(n)}}\cdot\text{indp-}\nu^m_{(n),k} - \hat{\nu}^m_{(n),*}\\ &= \sum_{k=nT}^{(n+1)T-1}\frac{\beta_k}{\beta_{(n)}}\left(\text{indp-}\nu^m_{(n),k} - \hat{\nu}^m_{(n),*}\right),\\ \check{e}^m_{(n)} &\triangleq \hat{\nu}^m_{(n),*} - \check{\nu}^m_{(n),*}.\end{aligned}$$

Lemma 4. *There exist constants $c, d, \rho \geq 0$ such that, for all $m \in \mathcal{T}$,*

$$\begin{aligned}\|\textit{team-}\nu^m_{(n),k} - \check{\nu}^m_{(n),*}\|_1 &\leq c\rho^{k-nT} + dT\alpha_{nT},\\ \|\textit{indp-}\nu^m_{(n),k} - \hat{\nu}^m_{(n),*}\|_1 &\leq c\rho^{k-nT} + dT\alpha_{nT}.\end{aligned}$$

The detailed proof of Lemma 4 can be found in Appendix B.1 of [6]. Let $\text{team-}e^m_{(n)}$ and $\text{indp-}e^m_{(n)}$ denote the errors for team-FP and independent team-FP, respectively. Lemma 4 yields that $\text{team-}e^m_{(n)}$ and $\hat{e}^m_{(n)}$ can be bounded by

$$\|\text{team-}e^m_{(n)}\|_1 \leq c\sum_{k=nT}^{(n+1)T-1}\frac{\beta_k}{\beta_{(n)}}\rho^{k-nT} + dR\alpha_{nT}, \quad (25)$$

$$\|\hat{e}^m_{(n)}\|_1 \leq c\sum_{k=nT}^{(n+1)T-1}\frac{\beta_k}{\beta_{(n)}}\rho^{k-nT} + dR\alpha_{nT}. \quad (26)$$

From Assumption 3 and (15), we have $\frac{\beta_{k+1}}{\beta_k} \leq 1$. Thus, we can bound $\frac{\beta_k}{\beta_{(n)}}$ by

$$\frac{\beta_k}{\beta_{(n)}} \leq \frac{\beta_{nT}}{T\beta_{(n+1)T-1}} \leq \frac{\alpha_{nT}}{T\alpha_{(n+1)T}}. \quad (27)$$

By Assumption 3, we obtain

$$\lim_{n\to\infty}\frac{\alpha_{nT}}{T\alpha_{(n+1)T}} = \frac{1}{T}\lim_{n\to\infty}\prod_{k=nT}^{(n+1)T-1}\frac{\alpha_k}{\alpha_{k+1}} = \frac{1}{T}. \quad (28)$$

Using (25), (26), (27), (28), and the decay property of α_k, we obtain, for all $m \in \mathcal{T}$,

$$\limsup_{n\to\infty}\left\|\text{team-}e^m_{(n)}\right\|_1 \leq \frac{c}{T}\frac{1}{1-\rho} \quad \text{and} \quad \limsup_{n\to\infty}\left\|\hat{e}^m_{(n)}\right\|_1 \leq \frac{c}{T}\frac{1}{1-\rho}.$$

Let $\hat{\nu}$ and $\check{\nu}$ denote the unique stationary distributions induced by the classical and independent log-linear learning, respectively. A small $\delta > 0$ implies close stationary distributions in the classical and independent settings. That is,

$$\|\hat{\nu} - \check{\nu}\|_1 \leq \Lambda(\delta, \varepsilon_\phi), \quad (29)$$

for some function Λ, and the difference $\|\hat{\nu} - \check{\nu}\|_1$ decays to zero as $\delta \to 0^+$ for any $\varepsilon_\phi > 0$. Based on (29), we can bound $\check{e}^m_{(n)}$ as $\check{e}^m_{(n)} \leq \Lambda(\delta, \varepsilon)$ for some $\varepsilon > 0$. Hence, given $\varepsilon > 0$, there exists $N_\varepsilon \in \mathbb{N}_+$ such that for any $n > N_\varepsilon$,

$$\|\text{team-}e^m_{(n)}\|_1 < C_{team}(\varepsilon, T), \quad (30)$$

$$\|\text{indp-}e^m_{(n)}\|_1 < C_{indp}(\varepsilon, T), \quad (31)$$

$$\|\text{a-}e^m_{(n+1)}\|_1 \leq \varepsilon, \quad (32)$$

where $C_{team}(\varepsilon, T) = \varepsilon + \frac{c}{T}\frac{1}{1-\rho}$ and $C_{indp}(\varepsilon, T) = \varepsilon + \frac{c}{T}\frac{1}{1-\rho} + \Lambda(\delta, \varepsilon)$. Note that $C_{team}(\varepsilon, T) \to 0$ as $\varepsilon \to 0,\ T \to \infty$, and $C_{indp}(\varepsilon, T) \to \Lambda(\delta, \varepsilon)$ as $\varepsilon \to 0,\ T \to \infty$.

We analyze the convergence of the TNG based on the above results. To this end, we define the set-valued mapping

$$F(\pi) \triangleq \left\{\left(br_\tau\left(\phi^m(\cdot,\pi^{-m})\right) - \pi^m + e^m\right)_{m\in\mathcal{T}}\,\middle|\, \forall m \in \mathcal{T}:\ \|e^m\|_1 \leq C(\varepsilon, T),\ br_\tau\left(\phi^m(\cdot,\pi^{-m})\right) + e^m \in \Delta\left(\underline{\mathcal{A}}^m\right)\right\}$$

for all $\pi \in \prod_{m\in\mathcal{T}}\Delta\left(\underline{\mathcal{A}}^m\right) \triangleq \Pi$, where

$$C(\varepsilon, T) = \begin{cases} C_{\text{team}}(\varepsilon, T) + \varepsilon & \text{for DTOA,}\\ C_{\text{indp}}(\varepsilon, T) + \varepsilon & \text{for iDTOA.}\end{cases}$$

Then, (23), (30), (31), and (32) imply that, for sufficiently large n, we obtain

$$\pi_{(n+1)} - \pi_{(n)} - \beta_{(n)}\cdot\omega_{(n+1)} \in \beta_{(n)}\cdot F\left(\pi_{(n)}\right).$$

The following argument is similar to the proof in Appendix A.2 of [6]. Finally, we get

$$\limsup_{k\to\infty} TNG(\pi^s_k) \leq \begin{cases}\tau\log|\mathcal{A}|, & \text{for DTOA,}\\ \tau\log|\mathcal{A}| + |\mathcal{T}|^2\overline{\phi}\Lambda(\delta, \varepsilon_\phi), & \text{for iDTOA,}\end{cases}$$

This completes the proof of Theorem 2.

C. Proof of Lemma 2 and Theorem 4

Proof of Lemma 2. Given a team $m \in \mathcal{T}$ and its supervisor s, define $X^{m,s}_k = \mathbf{1}\{v^{m,s}_k = 1,\ z^{m,s}_k = \text{fail}\}$. Then we have

$$\begin{aligned}\mathbb{E}\left(X^{m,s}_k\right) &= \begin{cases} q\cdot\eta_{FP}, & m \in \mathcal{H}\\ q\cdot\left[(1-p_{\text{lie}})\eta_{FP} + p_{\text{lie}}(1-\eta_{FN})\right], & m \in \mathcal{B}\end{cases}\\ &= \mathbb{P}\left(X^{m,s}_k = 1\right).\end{aligned}$$

For any $t > 0$, $F^{m,s}_t = \sum_{k=1}^t X^{m,s}_k$ is the number of fail signals up to time step t. Using the KL-form Chernoff bound (see, e.g., [42], Chapter 2) and (7), we obtain

$$\begin{aligned}\mathbb{P}\left(\frac{F^{m,s}_t}{t} > f\right) &\leq \exp\left(-tD(f\|q\eta_{FP})\right),\ m \in \mathcal{H},\\ \mathbb{P}\left(\frac{F^{m,s}_t}{t} < f\right) &\leq \exp\Big(-tD\big(f\|q[(1-p_{\text{lie}})\eta_{FP}\\ &\qquad + p_{\text{lie}}(1-\eta_{FN})]\big)\Big),\ m \in \mathcal{B},\end{aligned}$$

where $D(x\|y) \triangleq x\log\frac{x}{y} + (1-x)\log\frac{1-x}{1-y}$. This completes the proof of Lemma 2.

Proof of Theorem 4. For $m \in \mathcal{H}$,

$$\mathbb{P}\left(\exists t \geq t_0 : \frac{F_t^{m,s}}{t} > f\right) \leq \sum_{t=t_0}^{\infty} \exp\left(-tD(f\|q\eta_{FP})\right) = \frac{\exp\left(-t_0 D(f\|q\eta_{FP})\right)}{1-\exp\left(-D(f\|q\eta_{FP})\right)}.$$

We define the following events:

$$\mathcal{E}_0(t_0) = \bigcap_{m\in\mathcal{H}} \left\{\forall t \geq t_0 : \frac{F_t^{m,s}}{t} \leq f\right\},$$
$$\mathcal{E}_1(t_1) = \bigcap_{m\in\mathcal{B}} \left\{\exists t \leq t_1 : \frac{F_t^{m,s}}{t} > f\right\}.$$

We have

$$\mathbb{P}\left(\mathcal{E}_0(t_0)\right) \geq 1 - |\mathcal{H}| \cdot \frac{\exp\left(-t_0 D(f\|q\eta_{FP})\right)}{1-\exp\left(-D(f\|q\eta_{FP})\right)} = 1-\delta_{t_0},$$
$$\mathbb{P}\left(\mathcal{E}_1(t_1)\right) \geq 1 - |\mathcal{B}| \cdot \exp\Big(-t_1 D\big(f\|q\big[(1-p_{\text{lie}})\eta_{FP} + p_{\text{lie}}(1-\eta_{FN})\big]\big)\Big) = 1-\delta_{t_1}.$$

Define the good event as $\mathcal{E}(t_0,t_1) = \mathcal{E}_0(t_0) \cap \mathcal{E}_1(t_1)$. Then $\mathbb{P}(\mathcal{E}(t_0,t_1)) \geq 1-\delta_{t_0}-\delta_{t_1} = 1-\delta_B$. On the good event $\mathcal{E}$, for $k > t_1$, the beliefs of Byzantine teams are frozen, while those of honest teams are not. Following the proof of Theorem 2, we can conclude

$$\limsup_{k\to\infty} TNG_{\mathcal{H}}(\pi_k^s) \leq \begin{cases} \tau\log|\mathcal{A}_{\mathcal{H}}|, & \text{for BR-DTOA,} \\ \tau\log|\mathcal{A}_{\mathcal{H}}| + |\mathcal{H}|^2\overline{\phi}\Lambda(\delta,\varepsilon), & \text{for BR-iDTOA.} \end{cases}$$

This completes the proof of Theorem 4.

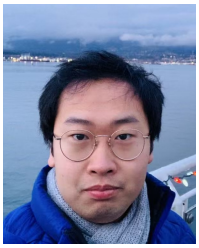
Juntian Zhu is a Ph.D. candidate with the School of Artificial Intelligence and Data Science, University of Science and Technology of China. She received the B.Sc. degree in mathematics from Sichuan University, Chengdu, China, in 2022. Her research interests are in (1) learning theory in uncertain AI environments; (2) multi-agent learning, decision-making and game-theoretic analysis under limited information; and (3) uncertainty-aware reasoning in large language models.

Guanpu Chen received his B.Sc. degree from University of Science and Technology of China, Hefei, China, in 2017, and Ph.D. degree from Academy of Mathematics and Systems Science, Chinese Academy of Sciences (CAS), Beijing, China, in 2022. He is currently a Professor with the School of Automation, Southeast University, Nanjing, China. He used to be a postdoctoral researcher with the School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden. His research interests include multi-agent systems, network games, as well as robustness, resilience, and security in cyber-physical systems. Prof. Chen was the recipient of the President Award of CAS, the Best Paper Award at IEEE International Conference on Control and Automation (ICCA), Guan Zhao-Zhi Award at Chinese Control Conference (CCC), and the Best Student Paper Honorable Mention at IEEE CSS TCSP.

Tongtian Zhu is a Ph.D. candidate with the College of Computer Science and Technology, Zhejiang University. His research interests are in (1) understanding modern deep learning systems from an optimization-theoretic perspective; and (2) the theoretical foundations and algorithms for scalable and autonomous decentralized, distributed, and multi-agent learning. He has made pioneering contributions to the study of implicit sharpness bias and data influence in decentralized learning, with related findings published at top-tier machine learning conferences, including ICML, NeurIPS, and ICLR, as oral and spotlight presentations.

Miguel de Carvalho is the Professor and Chair of Statistical Data Science, a Fellow of Generative AI Laboratory, and the Co-Director of the Edinburgh Centre for Financial Innovations at the University of Edinburgh. He is also an Honorary Professor at Universidade de Aveiro. He received a BSc in Mathematics from the NOVA School of Science and Technology in 2004, an MSc in Economics from the NOVA School of Business and Economics in 2009, and a PhD in Mathematics with emphasis on Statistics from the NOVA School of Science and Technology in 2009. His research is in Extreme Value Theory, Interfaces between Statistics and AI, and Bayesian Analysis. He is the Editor-in-Chief of the Springer book series on Courses in Advanced Statistics and Data Science, and an Associate Editor of American Statistician, the Annals of Applied Statistics, Bayesian Analysis, Computational Statistics & Data Analysis, Extremes, the Journal of the American Statistical Association, and Statistics and Public Policy.

Zhouwang Yang is a Professor in the School of Mathematical Sciences at University of Science and Technology of China (USTC). He received his Bachelor degree, Master degree and PhD degree in Mathematics from USTC in 1997, 2000 and 2005, respectively. He worked at Seoul National University as a postdoctoral researcher from December 2006 to November 2007. He was a visiting scholar in School of Industrial and Systems Engineering (ISyE) at Georgia Institute of Technology during August 2010 - August 2011. He has been working on geometric modeling and processing; optimization methods and their applications in sparse recovery. He also has a particular interest in the integration of optimization, machine learning and statistics for solving big data problems.



Fengxiang He is a Lecturer at the University of Edinburgh, and a Fellow of its Generative AI Laboratory. He received a BSc in statistics from the University of Science and Technology of China, an MPhil and a PhD in computer science from the University of Sydney in 2017, 2019, and 2021, respectively. His research interests are in (1) understanding AI from both learning-theoretical and game-theoretical views; (2) new models and algorithms leveraging symmetries in data, task, environment; (3) collaboration and interactions between AI agents in decentralized and multi-agent settings; and (4) applications in economics and finance. He is an Area Chair of ICML, NeurIPS, ICLR, UAI, and AISTATS, an Editorial Board member of Machine Learning, and an Associate Editor of Pattern Recognition.