Distributed Adaptive Flocking Control for Large-Scale Multiagent Systems

Shawon Dey and Hao Xu, *Member, IEEE*

***Abstract*— This article presents a novel distributed flocking control method for large-scale multiagent systems (LS-MASs) operating in uncertain environments. When dealing with a massive number of flocking agents in uncertain environments, existing flocking methods encounter the problem of communication complexity and "Curse of dimensionality" caused by the exponential growth of agent interactions while solving PDE-based optimal flocking control for large-scale systems. The mean field game (MFG) method addresses this issue by transforming interactions among all agents into the interaction of each individual agent with average effects represented by a probability density function (pdf) of other agents. However, relying solely on a pdf term to consider other agents' states can result in inefficient flocking performance due to the absence of a proficient coordination mechanism encompassing all agents involved in flocking. To overcome these difficulties and achieve the desired flocking performance for LS-MASs, the agents are decomposed into a finite number of subgroups. Each subgroup comprises a leader and followers, and a hybrid game theory is developed to manage both inter- and intragroup interactions. The method incorporates a cooperative game that links leaders from different groups to formulate distributed flocking control, a Stackelberg game that teams up leaders and followers within the same group to extend collective flocking behavior, and an MFG for followers to address the challenges of LS-MASs. Furthermore, to achieve distributed adaptive flocking using the hybrid game structure, we propose a hierarchical actor–critic-mass-based reinforcement learning technique. This approach incorporates a multiactor–critic method for leaders and an actor–critic-mass algorithm for followers, enabling adaptive flocking control in a distributed manner for large-scale agents. Finally, numerical simulation including comparison study and Lyapunov analysis demonstrates the effectiveness of the developed method.**

***Index Terms*— Flocking, mean field game (MFG), multiagent, reinforcement learning.**

I. Introduction

THE idea of flocking has been adopted from living beings such as flocks of birds [1], [2], [3] and schools of fishes [4], where a common objective is executed using cooperative behavior among the individuals in the group. Flocking control has been widely studied in research communities since 1987 when Reynolds proposed three fundamental rules, that is, 1) flock centering; 2) velocity matching; and 3) collision avoidance to achieve the flocking behavior [5]. Over the last few decades, a significant number of researches have been conducted in multiagent flocking control ([6], [7], [8], [9], [10]). Olfati-Saber [11] presented a theoretical framework for distributed flocking control. The framework addresses the flocking behavior of a finite group of agents relying on two major assumptions: first, all agents possess information about the virtual leader's state, and second, the virtual leader maintains a constant velocity. Additionally, the framework incorporates interaction among neighboring agents to achieve cohesive flocking behavior. The Olfati-Saber algorithm has been modified in [12] by removing these two assumptions. The revised approach involved a partial information dissemination strategy, where only a limited number of followers were provided information about the virtual leader. Meanwhile, the uninformed agents were influenced intermittently by the informed agents. However, the authors of these two major studies along with other existing studies in this domain did not address the potential issues arising from the exponential growth in dimensionality as the agents' number increases significantly, nor did they discuss the challenges associated with real-time communication among a large number of agents while interacting with each other to achieve the desired flocking behavior. To tackle these challenges, we integrate the mean field game (MFG) [13], [14], [15] approach, wherein each agent receives information about the other agents through a probability density function (pdf) rather than direct interactions. This allows us to meet the flocking requirements without requiring explicit interactions among the agents. In a more realistic scenario, a real leader is employed instead of a virtual leader. The leader and their followers utilize the pdf to align their velocities and achieve the desired flocking behavior. This approach is more plausible in real-world situations, as it allows for effective coordination without the need for direct interactions between individual agents. However, the efficiency of achieving flocking for large-scale agents can be impeded by relying on an agent to directly solve the Fokker–Planck–Kolmogorov (FPK) equation—a partial differential equation (PDE) to obtain the pdf information of the entire group. Without direct communication, the utilization of a single pdf function for other agents' information may result in inferior performance in effectively achieving flocking behavior. Additionally, in real-world scenarios, the decomposition of agents into multiple groups [16], [17], [18], [19] is a natural occurrence, where agents may have distinct roles and importance that influence their neighbors differently. For example, in a

Manuscript received 21 January 2023; revised 21 August 2023; accepted 1 December 2023. This work is supported by the National Science Foundation under grant 2144646 and Air Force Research Lab under grant FA8750-23-C-0509. *(Corresponding author: Hao Xu.)*

The authors are with the Department of Electrical and Biomedical Engineering, University of Nevada, Reno, Reno, NV 89557 USA (e-mail: shawondey@nevada.unr.edu; haoxu@unr.edu).

Digital Object Identifier 10.1109/TNNLS.2023.3343666

predator–prey system, each predator's role and importance can vary based on their position relative to the prey.

To address this knowledge gap, the large number of mean field agent games with a single leader is adapted into a hybrid game. In this modified approach, the large-scale multiagent system (LS-MAS) is decomposed into multiple groups, with each group being assigned a leader. This decomposition enables the agents to work with multiple pdfs instead of a single one, allowing for more efficient fulfillment of the flocking requirements. The hybrid game that has been developed comprises three key components: 1) a cooperative game [20], [21] involving leaders from multiple groups facilitating information exchange to coordinate flock centering, collision avoidance, and velocity matching for all agents; 2) a Stackelberg game [22] is employed to establish bonds among the leader and his followers within the group, ensuring that the leader informs its followers about flock centering as well as obtaining pdf information from followers for cohesive flocking; and 3) the MFG [13], [14], [15] is applied for intragroup followers to influence the group followers by locally generating pdfs, substituting all agents' states, and achieving collective flocking of large-scale followers, thereby mitigating dimensional explosion. The developed framework is illustrated in Fig. 1. It is worth noting that the cooperation between leaders has a minimal impact on computation and real-time communication compared to the communication required among large-scale agents, which approaches infinity.

To achieve globally distributed optimal flocking control for multiple groups, each leader must solve the Hamiltonian–Jacobi–Bellman (HJB) equation to guide their respective followers and contribute to the overall achievement of flocking behaviors for all agents. Additionally, followers from each group need to solve two coupled PDEs known as FPK equations for pdf estimates and the HJB equation for optimal flocking control. However, the complexity of solving the two coupled PDEs in real time presents challenges. To tackle this issue, the application of reinforcement learning [23], [24], [25], and adaptive dynamic programming [26] techniques has been adopted, enabling the efficient real-time solution of the coupled PDEs. The developed approach involves a hierarchical actor–critic-based reinforcement learning method that achieves adaptive flocking in a distributed manner. This algorithm combines actor–critic-mass-based learning for a large number of followers in each group and multiple actor–critic-based learning for leaders from different groups. In the proposed method, neural networks (NNs) for mass learning capture the behaviors of a large population in each group by estimating the solution of FPK equations. Critic NNs, on the other hand, learn the optimal cost function by solving the HJB equations, while actor NNs are responsible for solving the optimal flocking control for both leaders and followers within the same group. The major contributions are summarized as follows.

1) To resolve the "Curse of dimensionality" challenge in solving the PDE-based optimal control for multiagent systems while the number of agents is increasing significantly, multiagent flocking is extended to large-scale multiagent flocking, and agents are decomposed into multiple groups. A leader–follower system is introduced for all groups, employing a novel hybrid game theory to ensure efficient overall flocking behaviors and enable the decomposition of mean-field game pdfs along with the groups to achieve effective influence on all agents.
2) To solve the PDE-based optimal multigroup flocking with the developed hybrid game structure in practice, a hierarchical actor–critic-based learning control algorithm is proposed. This algorithm can achieve real-time multigroup flocking behavior even in the presence of environmental uncertainties.

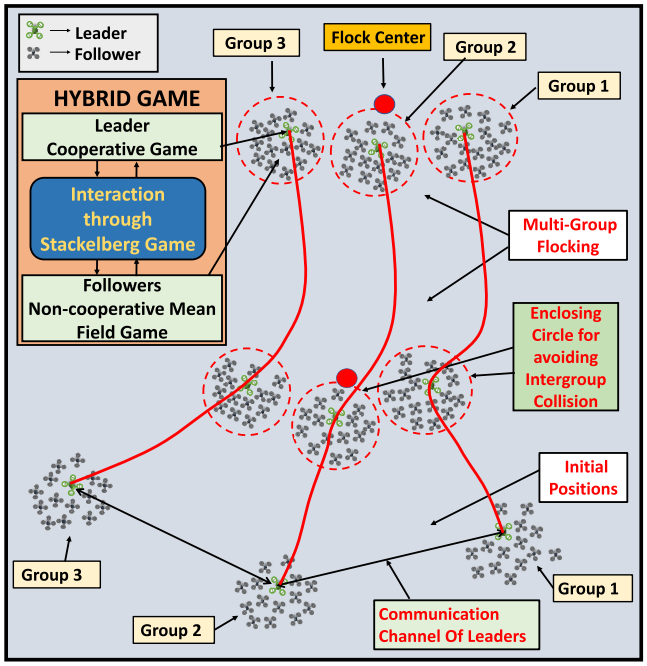


Fig. 1. Hybrid game theory-based flocking system.

II. Problem Formulation

Consider a flock of multigroup LS-MASs. Each group of the flock has a leader and significant followers. Let n denote the total number of groups in the system. Thus, there are n number of leaders in the system. Each leader is identified by the index i. Furthermore, the total number of followers in the ith group is N_i. Throughout this study, it is important to highlight that i denotes the leader of the respective ith group, considering the single leader in each group. Next, the state of the leader from the ith group is defined as $x_i = \left[p_i^T \ v_i^T\right]^T \in \mathbb{R}^{2m}$. The state includes position $p_i \in \mathbb{R}^m$ and velocity $v_i \in \mathbb{R}^m$ of the leader. Also, j represents the follower index. Then, jth follower state in the ith group can be represented as $x_{i,j} = \left[p_{i,j} \ v_{i,j}\right]^T \in \mathbb{R}^{2m}$ with $p_{i,j} \in \mathbb{R}^m$,$v_{i,j} \in \mathbb{R}^m$ being the jth follower's position and velocity in the ith group, respectively.

A. Leader–Follower Dynamic System for Flocking

Without sacrificing generality, the ith group agents' system dynamics are derived next. First, the leaders' dynamic i is represented as

$$\begin{aligned} dp_i &= v_i dt \\ dv_i &= [f_a(p_i, v_i) + g_a(p_i, v_i)u_i]dt + \sigma_{i,v} d\omega_{i,v} \end{aligned} \tag{1}$$

where $\omega_{i,v} \in \mathbb{R}^m$ represent wiener processes, and the Wiener process coefficient matrix is represented by $\sigma_{i,v}$. Also, $u_i \in \mathbb{R}$ is the control input of the leader. The nonlinear dynamics of the leaders are described by the functions $f_a(\cdot) \in \mathbb{R}^m$ and $g_a(\cdot) \in \mathbb{R}^m$. Then, (1) can be reformulated as

$$dx_i = [f_a'(x_i) + g_a'(x_i)u_i]dt + \sigma_i d\omega_i \tag{2}$$

where $f_a'(x_i) = \left[v_i^T \;\; f_a^T(p_i, v_i)\right]^T \in \mathbb{R}^{2m}$ and $g_a'(x_i) = \left[0 \;\; g_a^T(p_i, v_i)\right]^T \in \mathbb{R}^{2m}$ are the augmented nonlinear dynamic of the leader. Also, $\sigma_i = \left[0 \;\; \sigma_{i,v}^T\right]^T \in \mathbb{R}^{2m\times 2m}$ and $\omega_i = \left[0 \;\; \omega_{i,v}^T\right]^T \in \mathbb{R}^{2m}$ are the augmented coefficient matrix and Wiener process, respectively. Now, the jth follower dynamics from the group i can be derived as

$$\begin{aligned} dp_{i,j} &= v_{i,j}dt \\ dv_{i,j} &= [f_s(p_{i,j}, v_{i,j}) + g_s(p_{i,j}, v_{i,j})u_{i,j}]dt + \sigma_{i,v}d\omega_{i,j,v} \end{aligned} \tag{3}$$

where $u_{i,j}(t) \in \mathbb{R}$ denotes the follower control input. In addition, $\omega_{i,j,v} \in \mathbb{R}^{2m}$ represents independent Wiener processes and $\sigma_{i,v} \in \mathbb{R}^{m\times m}$ denotes the coefficient matrix. The functions $f_s(\cdot) \in \mathbb{R}^m$ and $g_s(\cdot) \in \mathbb{R}^m$ are the nonlinear homogeneous dynamics. Now, (3) is rewritten as

$$dx_{i,j} = [f_s'(x_{i,j}) + g_s'(x_{i,j})u_{i,j}]dt + \sigma_i d\omega_{i,j} \tag{4}$$

where $f_s'(x_{i,j}) = \left[v_{i,j} \;\; f_s^T(p_{i,j}, v_{i,j})\right]^T \in \mathbb{R}^{2m}$ and $g_s'(x_{i,j}) = \left[0 \;\; g_s^T(p_{i,j}, v_{i,j})\right]^T \in \mathbb{R}^{2m}$ are the augmented nonlinear intrinsic dynamics of the follower. Besides that, $\sigma_i = \left[0 \;\; {\sigma_{i,v}}^T\right]^T \in \mathbb{R}^{2m\times 2m}$ and $\omega_{i,j} = \left[0 \;\; \omega_{i,j,v}\right]^T \in \mathbb{R}^{2m}$.

B. Hybrid Game-Based Flocking Control Cost Formulation

This section develops and applies hybrid game theory to formulate the multileader–follower system optimal flocking control with multiple groups.

1) Group Leader Flocking: Each leader in a group must adhere to three fundamental rules governing collective flocking behavior, that is, 1) flock centering [5]; 2) velocity matching; and 3) collision avoidance. To facilitate cohesive flocking behavior, the leader within each group engages in interactions with neighborhood leaders, adhering to these three flocking rules. Now, consider $x_r(t) \in \mathbb{R}^{2m}$ is the state of the virtual flocking center, and the flock centering error for the leader as $e_i(t) = x_i(t) - x_r(t), \forall i = 1, 2, \ldots, n$. Then, using (2), the flock-centering error dynamic is defined as follows:

$$\begin{aligned} de_i(t) &= dx_i(t) - dx_r(t) \\ &= [f_{ar}'(e_i) + g_{ar}'(e_i)u_i]dt + \sigma_i d\omega_i \end{aligned} \tag{5}$$

with $f_{ar}'(e_i) = f_a'(e_i + x_r) - (dx_r/dt)$ and $g_{ar}'(e_i) = g_a'(e_i + x_r)$. Subsequently, the leader cost function associated with tracking the flock center is defined as follows:

$$\Phi_{\mathrm{FC,L}}(x_i) = \|x_i - x_r\|^2_{Q_{\mathrm{FC,L}}} \tag{6}$$

with $Q_{\mathrm{FC,L}}$ being the positive definitive matrix (PDM). Next, the leader's cooperation is characterized by how it interacts with neighboring leaders to enable the entire group to flock together effectively. Let $h > 0$ represent the sensing distance that enables the leader to detect its neighborhood leaders. The set $\mathcal{M}_i$ represents the neighborhood of the group i's leader

$$\mathcal{M}_i(t) = \{i_{\mathrm{nei}} \in \mathcal{V} : \|x_i - x_{i_{\mathrm{nei}}}\|^2_{Q_{\mathrm{nei}}} < h, i_{\mathrm{nei}} \neq i\} \tag{7}$$

where the set $\mathcal{V}$ comprises all the leaders' node for undirected graph $\mathcal{G}$. Moreover, $\mathcal{A} = [a_{i,i_{\mathrm{nei}}}] \in \mathbb{R}^{n\times n}$ denotes symmetric adjacency matrix, with $a_{i,i_{\mathrm{nei}}}$ is selected as $a_{i,i_{\mathrm{nei}}} = a_{i_{\mathrm{nei}},i} = 1$ if $(i, i_{\mathrm{nei}}) \in \mathcal{E}$. Here, notation $\mathcal{E}$ represents the edge set defined as $\mathcal{E} \subseteq \{(i, i_{\mathrm{nei}}) : i, i_{\mathrm{nei}} \in \mathcal{V}, i \neq i_{\mathrm{nei}}\}$. Now, assuming that the leader of the group i can communicate with the neighborhood leader i_{nei} if $i_{\mathrm{nei}} \in \mathcal{M}_i(t)$. Then, the collision avoidance cost of the leader to meet the flocking requirements is defined as

$$\Phi_{\mathrm{CO,L}}(x_i) = w_{\mathrm{CO,L}} \sum_{i_{\mathrm{nei}}\in\mathcal{M}_i} [e^{(\|x_i - x_{i_{\mathrm{nei}}}\|^2_{Q_{\mathrm{CO,L}}} - d_{i_{\mathrm{nei}}})} - 1]^{-1} \tag{8}$$

where $Q_{\mathrm{CO,L}} = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{2m\times 2m}$ with identity matrix $I \in \mathbb{R}^{m\times m}$, $w_{\mathrm{CO,L}}$ is positive weighting scalar, and $x_{-i} = \{x_{i_{\mathrm{nei}}}\}_{i_{\mathrm{nei}}\in\mathcal{M}_i}$, with $\mathcal{M}_i$ being neighbor leaders set. Also, $d_{i_{\mathrm{nei}}}$ represents the distance threshold between leader i and leader i_{nei}. This distance is carefully chosen to ensure the prevention of any potential collisions between the respective leaders and their accompanying followers. As the distance between any two leaders approaches threshold distance, the collision avoidance function tends toward infinity. The velocity requirement of the leader is achieved by the following function:

$$\Phi_{\mathrm{VM,L}}(x_i, x_{-i}) = w_{\mathrm{VM,L}} \sum_{i_{\mathrm{nei}}\in\mathcal{M}_i} a_{i,i_{\mathrm{nei}}} \left[\|x_i(t) - x_{i_{\mathrm{nei}}}(t)\|^2_{Q_{\mathrm{VM,L}}}\right] \tag{9}$$

where $w_{\mathrm{VM,L}}$ is the positive weighting scalar, $a_{i,i_{\mathrm{nei}}}$ is the element of the symmetric adjacency matrix of the leader i and neighborhood leader i_{nei}, and $Q_{\mathrm{VM,L}} = \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} \in \mathbb{R}^{2m\times 2m}$. This function ensures that the leaders from multiple groups match their corresponding velocities to achieve the collective flocking of a large number of agents. Also, the identity matrix $Q_{\mathrm{VM,L}}$ is used to evaluate the velocity errors of the leader.

2) Coupling Functions for Flocking: While the leader tracks the virtual center and meets the flocking requirements, it is also essential to ensure the successful tracking of respective leaders by corresponding followers to achieve overall flocking within the decomposed group of agents. The coupling cost functions are designed by the Stackelberg game [22] to guarantee that the followers of the same group can flock with their group leader. It includes a flocking center coupling cost function as

$$\Phi_{\mathrm{FC,C}}(x_{i,j}, x_i) = \left[e^{(r_i - \|x_{i,j} - x_i\|^2_{Q_{\mathrm{FC,C}}})} - 1\right]^{-1} \tag{10}$$

where $Q_{\mathrm{FC,C}} = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{2m\times 2m}$ and r_i signifies the predefined safety distance threshold for the followers within group i. In this scenario, r_i is considered as the radius of a circular region that serves as a zone devoid of collisions for the distinct groups. Note that the threshold distance $d_{i_{\mathrm{nei}}}$ of the leaders satisfy the condition $d_{i_{\mathrm{nei}}} \geq r_i + r_{i_{\mathrm{nei}}}$ to ensure the prevention of collisions between the followers of neighborhood groups.

When the follower approaches the boundary of this circle, then the power term of e approaches zero, and the cost function approaches infinity. As the follower distances itself from the boundary and gets closer to the leader, then the power term becomes larger and the cost function approaches zero. Then, to achieve the overall group flocking, we introduce a velocity-matching coupling function for the leader, enabling synchronization of velocities within their respective groups. Since the leader only has the pdf information of its followers, this information is employed in the cost function of the leader

$$\Phi_{\mathrm{VM,C}}(x_i, \rho_i) = \|x_i(t) - \mathbb{E}\{\rho_i(x_{i,F}, t)\}\|^2_{Q_{\mathrm{VM,C}}} \tag{11}$$

with $Q_{\mathrm{VM,C}} = \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} \in \mathbb{R}^{2m\times 2m}$, and $\rho_i(x_{i,F}, t)$ represents the pdf of the state of the large number of followers in group i. Note that every decomposed group leader has its respective followers' pdf information. These two coupled cost functions can validate the cohesiveness of the group leaders and the massive followers.

3) Flocking of Large-Scale Followers: The group leader serves as the flocking center for followers. The followers' flocking center error is represented as $e_{i,j}(t) = x_{i,j}(t) - x_i(t)$. Now, the tracking error dynamics can be obtained as follows:

$$\begin{aligned} de_{i,j}(t) &= dx_{i,j}(t) - dx_i(t) \\ &= [f'_{sr}(e_{i,j}) + g'_{sr}(e_{i,j})u_{i,j}]dt + \sigma_i d\omega^e_{i,j} \\ \text{with,}\quad f'_{sr}(e_{i,j}) &= f'_s(e_{i,j} + x_i) - f'_a(x_i) \\ g'_{sr}(e_{i,j}) &= g'_s(e_{i,j} + x_i) - g'_a(x_i)u_i/u_{i,j}\ ; \\ \omega^e_{i,j} &= \omega_{i,j} - \omega_i. \end{aligned} \tag{12}$$

Then, the tracking error cost function is defined as

$$\Phi_{\mathrm{FC,F}}(x_{i,j}) = \|x_{i,j} - x_i\|^2_{Q_{\mathrm{FC,F}}} \tag{13}$$

with $Q_{\mathrm{FC,F}}$ being the PDM. Now, the collision avoidance function of the followers [27] is defined as

$$\begin{aligned} &\Phi_{\mathrm{CO,F}}(x_{i,j}, \rho_i) \\ &= w_{\mathrm{CO,F}} \int_{x_{i,F}} \frac{\rho_i(x_{i,F}, t)}{\left[\varepsilon^2 + \|x_{i,j}(t) - \mathbb{E}\{\rho_i(x_{i,F}, t)\|^2_{Q_{\mathrm{CO,F}}}\right]^\beta} dx_{i,F} \end{aligned} \tag{14}$$

where $Q_{\mathrm{CO,F}} = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{2m\times 2m}$ with $I \in \mathbb{R}^{m\times m}$ and $w_{\mathrm{CO,F}}$ is a weighting parameter. Also, ε and β are positive parameters. This function serves the purpose of aiding followers in evading collisions with other followers within the same group. Given that followers possess solely pdf data about their group members, as opposed to precise state information, this cost function is employed to guide followers away from dense regions, thus diminishing the likelihood of collisions. Next, the velocity matching of the followers with their group members is achieved by the subsequent cost function

$$\Phi_{\mathrm{VM,F}}(x_{i,j}, \rho_i) = w_{\mathrm{VM,F}}\{\|x_{i,j} - \mathbb{E}\{\rho_i(x_{i,F}, t)\}\|^2_{Q_{\mathrm{VM,F}}}\} \tag{15}$$

where $Q_{\mathrm{VM,F}} = \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} \in \mathbb{R}^{2m\times 2m}$ is an identity matrix and $w_{\mathrm{VM,F}}$ denotes a positive weighting value. Now, the optimal flocking control cost function of this hybrid game-based leader–follower structure is defined as follows.

4) Optimal Flocking Control Cost Function: The overall flocking cost function for the leader system is derived as

$$\begin{aligned} V_i(x_i, x_{-i}, \rho_i) &= \mathbb{E}\left\{\int_0^\infty \begin{bmatrix} \Phi_{\mathrm{LD}}(x_i, x_{-i}, u_i) + \\ \Phi_{\mathrm{VM,C}}(x_i, \rho_i) \end{bmatrix} dt\right\} \\ \text{with, } \Phi_{\mathrm{LD}}(x_i, x_{-i}, u_i) &= \Phi_{\mathrm{FC,L}}(x_i) + \Phi_{\mathrm{CO,L}}(x_i, x_{-i}) \\ &\quad + \Phi_{\mathrm{VM,L}}(x_i, x_{-i}) + \|u_i\|^2_{R_i} \end{aligned} \tag{16}$$

where R_i is the weight matrix for the quadratic control input. The leader optimizes this cost function to achieve the flocking behavior with respect to its own group followers and other group leaders. Next, the jth follower cost function is as

$$\begin{aligned} V_{i,j}(x_{i,j}, x_i, \rho_i) &= \mathbb{E}\left\{\int_0^\infty \begin{bmatrix} \Phi_{\mathrm{FL}}(x_{i,j}, x_i, u_{i,j}) + \\ \Phi_{\mathrm{CO,F}}(x_{i,j}, \rho_i) + \\ \Phi_{\mathrm{VM,F}}(x_{i,j}, \rho_i) \end{bmatrix} dt\right\} \\ &\text{s.t.} \quad \|x_{i,j_o} - x_i\| < r_i \\ \text{with, } \Phi_{\mathrm{FL}}(x_{i,j}, x_i, u_{i,j}) &= \Phi_{\mathrm{FC,F}} + \Phi_{\mathrm{FC,C}} + \|u_{i,j}\|^2_{R_{i,j}} \end{aligned} \tag{17}$$

where x_{i,j_o} represents the follower jth initial state in group i. Note that the initial deployment of the follower satisfies the above constraint $\|x_{i,j_o} - x_i\| < r_i$, which indicates that the initial distance of the follower from the leader is less than the threshold distance r_i. After that, the coupling function $\Phi_{\mathrm{FC,C}}(x_{i,j}, x_i)$ ensures that the followers always stay close to the leader and do not venture beyond this threshold distance. By ensuring the follower cohesiveness with the leader, it is guaranteed that the leader can achieve the group collision avoidance behavior using cost function $\Phi_{\mathrm{CO,\,L}}(x_i, x_{-i})$. Even though, the leader does not have the individual state information of followers except the pdf term, these two cost functions still ensure the group collision avoidance and flocking behavior. The leader's overall cost function is

$$V_L(x_i, \rho) = \sum_{l=1}^{M} \mathbb{E}\left\{\int_0^\infty \begin{bmatrix} \Phi_{\mathrm{LD}}(x_i, x_{-i}, u_i) + \\ \Phi_{\mathrm{VM,C}}(x_i, \rho_i) \end{bmatrix} dt\right\}. \tag{18}$$

C. Hybrid Game -Distributed Flocking Control

The leaders and followers Hamiltonian can be obtained as follows, by incorporating optimal control theory [28] and Bellman's optimality [29]:

$$H_L[x_i, \partial_{x_i} V_L(x_i, \rho)] = \sum_{i=1}^{M} H_i[x_i, \partial_{x_i} V_i(x_i, x_{-i}, \rho_i)]. \tag{19}$$

Now, any individual leader i's distributed Hamiltonian is

$$\begin{aligned} &H_i[x_i, \partial_{x_i} V_i(x_i, x_{-i}, \rho_i)] \\ &= \mathbb{E}\{\Phi_{\mathrm{LD}}(x_i, x_{-i}, u_i) + \Phi_{\mathrm{VM,C}}(x_i, \rho_i) \\ &\qquad + \partial_{x_i} V_i{}^T(x_i, x_{-i}, \rho_i)[f'_a(x_i) + g'_a(x_i)u_i]\}. \end{aligned} \tag{20}$$

The Hamiltonian of the jth follower from ith group is

$$\begin{aligned} &H_{i,j}[x_{i,j}, \partial_{x_{i,j}} V_{i,j}(x_{i,j}, x_i, \rho_i)] \\ &= \mathbb{E}\big\{\Phi_{\mathrm{FL}}(x_{i,j}, x_i, u_{i,j}) + \Phi_{\mathrm{CO,F}}(x_{i,j}, \rho_i) \\ &\qquad + \Phi_{\mathrm{VM,F}}(x_{i,j}, \rho_i) + \partial_{x_{i,j}} V^T_{i,j}(x_{i,j}, x_i, \rho_i) \\ &\qquad + [f'_s(x_{i,j}) + g'_s(x_{i,j})u_{i,j}]\big\}. \end{aligned} \tag{21}$$

Using cooperative game theory [20], the leaders' HJB equation from multiple groups is defined in (24), as shown at the bottom of the page 6. Then, the HJB-FPK equations of followers are

derived in (25) and (26), as shown at the bottom of the next page. Once again, the leader–follower optimal flocking control can be obtained as follows:

$$u_i(x_i) = -1/2\ \mathbb{E}\{R_i^{-1} g_a'^T(x_i)\partial_{x_i} V_i(x_i, x_{-i}, \rho_i)\} \tag{22}$$

$$u_{i,j}(x_{i,j}) = -1/2\ \mathbb{E}\{R_{i,j}^{-1} g_s'^T(x_{i,j})\partial_{x_{i,j}} V_{i,j}(x_{i,j}, x_i, \rho_i)\}. \tag{23}$$

Remark 1: Solving the forward-in-time HJB equation for leaders, the interlinked HJB and FPK equations of followers is a complex task [28]. To address this challenge, the emerging reinforcement learning algorithm [23] has been embraced. This algorithm enables online learning of optimal control strategies and facilitates adaptation to environmental uncertainties.

III. Hierarchical Actor–Critic-Based NN Estimators

In the developed framework of hybrid game theory within the context of a multigroup mean field system, the primary objective of each agent is to attain a desired flocking behavior. This objective is pursued by optimizing the control strategies of leader–followers within each group, achieved through the minimization of corresponding cost functions. To accomplish this, a set of n critic and n actor NNs are employed, aimed at estimating optimal cost functions and control actions for the n group leaders. These leaders operate cooperatively, exchanging information to collaboratively achieve distributed flocking control. Meanwhile, each follower within any group i maintains an actor–critic-mass NN. The NN structure is shown in Fig. 2. This network serves the purpose of estimating the follower's control, cost, and pdf.

Next, the leaders' optimal flocking cost and control function are defined as follows:

1) Ideal NNs Design for Leader:

$$\begin{aligned}\text{Critic: } V_i^*(x_i, x_{-i}, \rho_i) &= \mathbb{E}\{W_{i,V}^T \phi_{i,V} + \varepsilon_{\text{HJB}_i}\}\\ \text{Actor: } u_i^*(x_i, x_{-i}, \rho_i) &= \mathbb{E}\{W_{i,u}^T \phi_{i,u} + \varepsilon_{i,u}\}.\end{aligned} \tag{27}$$

Also, the followers' optimal flocking cost, control and PDF functions are as follows:

2) Ideal NNs Design for Followers:

$$\begin{aligned}\text{Critic: } V_{i,j}^*(x_{i,j}, x_i, \rho_i) &= \mathbb{E}\{W_{i,j,V}^T \phi_{i,j,V} + \varepsilon_{\text{HJB}_{i,j}}\}\\ \text{Actor: } u_{i,j}^*(x_{i,j}, x_i, \rho_i) &= \mathbb{E}\{W_{i,j,u}^T \phi_{i,j,u} + \varepsilon_{i,j,u}\}\\ \text{Mass: } \rho_i(x_{i,j}, t) &= \mathbb{E}\{W_{i,j,\rho}^T \phi_{i,j,\rho} + \varepsilon_{\text{FPK}_{i,j}}\}\end{aligned} \tag{28}$$

where $W_{i,V}$, $W_{i,u}$, $W_{i,j,V}$, $W_{i,j,u}$, and $W_{i,j,\rho}$ correspond weight of the ith group leader–follower. The activation functions of these networks are represented by $\phi(.)$, while ε denotes the NN reconstruction errors. Moving forward, we can approximate the leaders' optimal cost and control as follows:

3) Estimated NN Design for the Leader:

$$\begin{aligned}\text{Critic: } \hat{V}_i(x_i, x_{-i}, \bar{\hat{\rho}}_i) &= \mathbb{E}\{\hat{W}_{i,V}{}^T \hat{\phi}_{i,V}\}\\ \text{Actor: } \hat{u}_i(x_i, x_{-i}, \bar{\hat{\rho}}_i) &= \mathbb{E}\{\hat{W}_{i,u}^T \hat{\phi}_{i,u}\}.\end{aligned} \tag{29}$$

Note that decentralized followers maintain an estimation over their intragroup pdf $\rho_{i,j}(x_{i,j}, t)$ and their group leader collects the estimations from followers. The leader then calculates the statistical mean of collected pdf, that is, $\bar{\hat{\rho}}_i = \sum_{i,j} \hat{\rho}_{i,j}$. Note that this communication link is highly fault-tolerant and does not require low latency. This is because followers transmit similar pdf functions to the leader in the same group. Substituting the approximation (29) into (24) and incorporating the reconstruction error impact, the residual errors of the leader critic–actor NNs are obtained as

$$\begin{aligned}&\text{Critic: } \mathbb{E}\{e_{\text{HJB}_i}\}\\ &= \mathbb{E}\{\Phi_{\text{VM,C}}(x_i, \bar{\tilde{\rho}}_i) - \tilde{W}_{i,V}^T \hat{\Psi}_{i,V}(x_i, x_{-i}, \bar{\hat{\rho}}_i)\\ &\quad - W_{i,V}^T \tilde{\Psi}_{i,V}(x_i, x_{-i}, \bar{\tilde{\rho}}_i) - \varepsilon_{\text{HJB}_i}\}\end{aligned} \tag{30}$$

$$\begin{aligned}&\text{Actor: } \mathbb{E}\{e_{i,u}\}\\ &= \mathbb{E}\{-\tilde{W}_{i,u}^T \hat{\phi}_{i,u}(x_i, x_{-i}, \bar{\hat{\rho}}_i) - W_{i,u}^T \tilde{\phi}_{i,u}(x_i,\\ &\quad x_{-i}, \bar{\tilde{\rho}}_i) - 1/2\ R_i^{-1} g_a'^T(x_i)\partial_{x_i} \tilde{V}_i - \varepsilon_{i,u}\}\end{aligned} \tag{31}$$

with, $\hat{H}_i = H_i[x_i, \partial_{x_i}\hat{\phi}_{i,V}] = \hat{W}_{i,V}^T \hat{H}_{i,W}$

$$\mathbb{E}\{\Psi_{i,V}(x_i, x_{-i}, \bar{\hat{\rho}}_i)\} = \mathbb{E}\{\partial_t \hat{\phi}_{i,V} + 0.5\sigma_i^2 \Delta\hat{\phi}_{i,V} - \hat{H}_{i,W}\}$$

$$\mathbb{E}\{\Phi_{\text{VM,C}}(x_i, \bar{\hat{\rho}}_i)\} = \mathbb{E}\{\Phi_{\text{VM,C}}(x_i, \bar{\tilde{\rho}}_i) + \Phi_{\text{VM,C}}(x_i, \rho_i)\}.$$

Also, $\tilde{W}_{i,V} = W_{i,V} - \hat{W}_{i,V}$, $\tilde{W}_{i,u} = W_{i,u} - \hat{W}_{i,u}$. The approximated functions representation is as follows:

$$\begin{aligned}&\mathbb{E}\{\Psi_{i,V}(x_i, x_{-i}, \bar{\tilde{\rho}}_i)\}\\ &\quad = \mathbb{E}\{\Psi_{i,V}(x_i, x_{-i}, \rho_i) - \hat{\Psi}_{i,V}(x_i, x_{-i}, \bar{\hat{\rho}}_i)\}\\ &\mathbb{E}\{\tilde{\phi}_{i,u}(x_i, x_{-i}, \bar{\tilde{\rho}}_i)\}\\ &\quad = \mathbb{E}\{\tilde{\phi}_{i,u}(x_i, x_{-i}, \rho_i) - \hat{\phi}_{i,u}(x_i, x_{-i}, \bar{\hat{\rho}}_i)\}.\end{aligned}$$

The updated laws for leaders' actor–critic NN are given as

$$\mathbb{E}\{\dot{\hat{W}}_{i,V}\} = \mathbb{E}\left\{-\alpha_{i,V} \frac{\hat{\Psi}_{i,V}(x_i, x_{-i}, \bar{\hat{\rho}}_i) e_{\text{HJB}_i}^T}{1 + \|\hat{\Psi}_{i,V}(x_i, x_{-i}, \hat{\rho}_i))\|^2}\right\} \tag{32}$$

$$\mathbb{E}\{\dot{\hat{W}}_{i,u}\} = \mathbb{E}\left\{-\alpha_{i,u} \frac{\hat{\phi}_{i,u}(x_i, x_{-i}, \bar{\hat{\rho}}_i) e_{i,u}^T}{1 + \|\hat{\phi}_{i,u}(x_i, x_{-i}, \hat{\rho}_i)\|^2}\right\}. \tag{33}$$

Now, the followers' optimal flocking cost, control and mass (PDF) are estimated as follows:

1) Estimated NN Design for Followers:

$$\begin{aligned}\text{Critic: } \hat{V}_{i,j}(x_{i,j}, x_i, \hat{\rho}_{i,j}) &= \mathbb{E}\{\hat{W}_{i,j,V}^T \hat{\phi}_{i,j,V}\}\\ \text{Actor: } \hat{u}_{i,j}(x_{i,j}, x_i, \hat{\rho}_{i,j}) &= \mathbb{E}\{\hat{W}_{i,j,u}^T \hat{\phi}_{i,j,u}\}\\ \text{Mass: } \hat{\rho}_{i,j}(x_{i,j}, t) &= \mathbb{E}\{\hat{W}_{i,j,\rho}^T \hat{\phi}_{i,j,\rho}\}.\end{aligned} \tag{34}$$

By substituting (34) into (25) and (26), residual errors are defined as

$$\begin{aligned}&\text{Critic: } \mathbb{E}\{e_{\text{HJB}_{i,j}}\}\\ &= \mathbb{E}\{\Phi_{\text{CO,F}}(x_{i,j}, \tilde{\rho}_{i,j}) + \Phi_{\text{VM,F}}(x_{i,j}, \tilde{\rho}_{i,j})\\ &\quad - \tilde{W}_{i,j,V}^T \hat{\Psi}_{i,j,V}(x_{i,j}, x_i, \hat{\rho}_{i,j}) - W_{i,j,V}^T \tilde{\Psi}_{i,j,V}(x_{i,j}, x_i, \tilde{\rho}_{i,j})\\ &\quad - \varepsilon_{\text{HJB}_{i,j}}\}\end{aligned} \tag{35}$$

$$\begin{aligned}&\text{Mass: } \mathbb{E}\{e_{\text{FPK}_{i,j}}\}\\ &= \mathbb{E}\{-\tilde{W}_{i,j,\rho}^T \hat{\Psi}_{i,j,\rho}(x_{i,j}, \hat{V}_{i,j}, t) - W_{i,j,\rho}^T\\ &\quad \tilde{\Psi}_{i,j,\rho}(x_{i,j}, \tilde{V}_{i,j}, t) - \varepsilon_{\text{FPK}_{i,j}}\}\end{aligned} \tag{36}$$

$$\begin{aligned}&\text{Actor: } \mathbb{E}\{e_{i,j,u}\}\\ &= \mathbb{E}\{-\tilde{W}_{i,j,u}^T \hat{\phi}_{i,j,u}(x_{i,j}, x_i, \hat{\rho}_{i,j}) - W_{i,j,u}^T\\ &\quad \tilde{\phi}_{i,j,u}(x_{i,j}, x_i, \tilde{\rho}_{i,j}) - \frac{1}{2} R_{i,j}^{-1} g_s'^T(x_{i,j})\partial_{x_{i,j}} \tilde{V}_{i,j} - \varepsilon_{i,j,u}\}\end{aligned} \tag{37}$$

with, $\hat{H}_{i,j} = H_{i,j}[x_{i,j}, \partial_{x_{i,j}}\hat{\phi}_{i,j,V}] = \hat{W}_{i,j,V}^T \hat{H}_{i,j,W}$

$$\mathbb{E}\{\Psi_{i,j,V}(x_{i,j}, x_i, \hat{\rho}_{i,j})\} = \mathbb{E}\left\{\begin{bmatrix}\partial_t \hat{\phi}_{i,j,V} + 0.5\sigma_i^2\\ -\Delta\hat{\phi}_{i,j,V}\hat{H}_{i,j,W}\end{bmatrix}\right\}$$

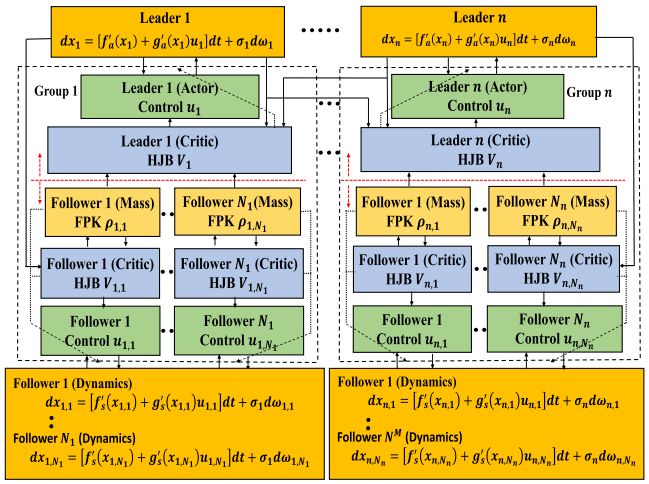


Fig. 2. Hierarchical actor–critic-based flocking control structure.

$$\mathbb{E}\{\Psi_{i,j,\rho}(x_{i,j},\hat{V}_{i,j},t)\} = \mathbb{E}\{\partial_t\hat{\phi}_{i,j,\rho} - 0.5\sigma_i^2\Delta\hat{\phi}_{i,j,\rho} - \mathrm{div}(\hat{\phi}_{i,j,\rho})D_p\hat{H}_{i,j}\}.$$

Also, $\tilde{W}_{i,j,V} = W_{i,j,V} - \hat{W}_{i,j,V}$, $\tilde{W}_{i,j,u} = W_{i,j,u} - \hat{W}_{i,j,u}$ and $\tilde{W}_{i,j,\rho} = W_{i,j,\rho} - \hat{W}_{i,j,\rho}$. The updated laws for followers' actor–critic-mass NNs are given as follows:

$$\mathbb{E}\{\dot{\hat{W}}_{i,j,V}\} = \mathbb{E}\left\{-\alpha_{i,j,V}\frac{\hat{\Psi}_{i,j,V}(x_{i,j},x_i,\hat{\rho}_{i,j})e^T_{\mathrm{HJB}_{i,j}}}{1+\|\hat{\Psi}_{i,j,V}(x_{i,j},x_i,\hat{\rho}_{i,j})\|^2}\right\} \quad (38)$$

$$\mathbb{E}\{\dot{\hat{W}}_{i,j,u}\} = \mathbb{E}\left\{-\alpha_{i,j,u}\frac{\hat{\phi}_{i,j,u}(x_{i,j},x_i,\rho_{i,j})e^T_{i,j,u}}{1+\|\hat{\phi}_{i,j,u}(x_{i,j},x_i,\rho_{i,j})\|^2}\right\} \quad (39)$$

$$\mathbb{E}\{\dot{\hat{W}}_{i,j,\rho}\} = \mathbb{E}\left\{-\alpha_{i,j,\rho}\frac{\hat{\Psi}_{i,j,\rho}(x_{i,j},\hat{v}_{i,j},t)e^T_{\mathrm{FPK}_{i,j}}}{1+\|\hat{\Psi}_{i,j,\rho}(x_{i,j},\hat{V}_{i,j},t)\|^2}\right\} \quad (40)$$

where the learning rates are $\alpha_{i,V}$, $\alpha_{i,u}$, $\alpha_{i,j,V}$, $\alpha_{i,j,u}$, and $\alpha_{i,j,\rho}$.

Remark 1: The functions $\hat{\Psi}_{i,V}(x_i,x_{-i},\hat{\rho}_i)$, $\hat{\phi}_{i,u}(x_i,x_{-i},\hat{\rho}_i)$, $\hat{\Psi}_{i,j,V}(x_{i,j},x_i,\hat{\rho}_{i,j})$, $\hat{\phi}_{i,j,u}(x_{i,j},x_i,\rho_{i,j})$, and $\hat{\Psi}_{i,j,\rho}(x_{i,j},\hat{v}_{i,j},t)$ satisfiy the persistency excitation (PE) condition when the NN weights are converging.

Theorem 1: The critic NN weight's $\mathbb{E}\{\hat{W}_{i,V}\}$ and $\mathbb{E}\{\hat{W}_{i,j,V}\}$ are tuned using update laws (32) and (38). The approximated weight errors $\mathbb{E}\{\tilde{W}_{i,V}\}$ and $\mathbb{E}\{\tilde{W}_{i,j,V}\}$ and cost estimation errors $\mathbb{E}\{\tilde{V}_i\}$ and $\mathbb{E}\{\tilde{V}_{i,j}\}$ are uniformly ultimately bounded (UUB). In addition, with a minimal reconstruction errors, $\mathbb{E}\{\hat{W}_{i,V}\}$, $\mathbb{E}\{\tilde{W}_{i,j,V}\}$, $\mathbb{E}\{\tilde{V}_i\}$, and $\mathbb{E}\{\tilde{V}_{i,j}\}$ are asymptotically stable. The bounds $\tilde{V}_i$ and $\tilde{V}_{i,j}$ are defined as

$$\begin{aligned}&\mathbb{E}\left\{\|\tilde{V}_i(t)\|\right\}\\ &= \mathbb{E}\left\{\|\tilde{W}^T_{i,V}\hat{\phi}_{i,V} + W^T_{i,V}\tilde{\phi}_{i,V} + \varepsilon_{\mathrm{HJB}_i}\|\right\}\\ &\le B_{i,V,W}\mathbb{E}\left\{\|\hat{\phi}_{i,V}\|\right\} + l_{\phi_{i,V}}\mathbb{E}\left\{\|W_{i,V}\|\right\}\bar{b}_{i,j,\rho} + \mathbb{E}\left\{\|\varepsilon_{\mathrm{HJB}_i}\|\right\}\\ &\equiv b_{i,V}(t). \end{aligned} \quad (41)$$

Algorithm 1 Adaptive Flocking Control

Initialize the state of group_l leader x_i in the flock
Initialize the state of jth follower $x_{i,j}$ from group i
Exchanging the state information between leader i and other neighbor leaders $i_{\mathrm{nei}} \in \mathcal{M}_i$
Initialize NN weights $\hat{W}_{i,V}$, $\hat{W}_{i,u}$ randomly
Initialize NN weights _j from group_i: $\hat{W}_{i,j,V}$, $\hat{W}_{i,j,u}$, $\hat{W}_{i,j,\rho}$ randomly
Initialize the leader-follower errors-
    Leader: e_{HJB_i}, $e_{i,u} \leftarrow \infty$
    Follower: $e_{\mathrm{HJB}_{i,j}}$, $e_{\mathrm{FPK}_{i,j}}$, $e_{i,j,u} \leftarrow \infty$
Initialize NNs thresholds δ_{HJB_i}, $\delta_{\mathrm{HJB}_{i,j}}$, $\delta_{i,u}$, $\delta_{i,j,u}$, $\delta_{\mathrm{FPK}_{i,j}}$
while TRUE **do**
    while $e_{\mathrm{HJB}_i} \ge \delta_{\mathrm{HJB}_i}, e_{i,u} \ge \delta_{i,u}$, **do**
        Use (32) and (33) to update NN weights
        Use (30) and (31) to update errors
    end while
    $\hat{u}_i(x_i,x_{-i},\bar{\hat{\rho}}_i) \leftarrow \hat{W}^T_{i,u}\hat{\phi}_{i,u}(x_i,x_{-i},\bar{\hat{\rho}}_i)$
    Apply the control $\hat{u}_i$
    Calculate new state x_i
    while $e_{\mathrm{HJB}_{i,j}} \ge \delta_{\mathrm{HJB}_{i,j}}, e_{i,j,u} \ge \delta_{i,j,u}, e_{\mathrm{FPK}_{i,j}} \ge \delta_{\mathrm{FPK}_{i,j}}$ **do**
        Use (38), (39) and (40) to update NN weights
        Use (35), (36) and (37) to update errors
    end while
    $\hat{u}_{i,j}(x_{i,j},x_i,\hat{\rho}_{i,j}) \leftarrow \hat{W}^T_{i,j,u}\hat{\phi}_{i,j,u}(x_{i,j},x_i,\hat{\rho}_{i,j})$
    Apply the control $\hat{u}_{i,j}$
    Calculate new state $x_{i,j}$
end while

Similarly

$$\begin{aligned}\mathbb{E}\left\{\|\tilde{V}_{i,j}(t)\|\right\} &= \mathbb{E}\left\{\|\tilde{W}^T_{i,j,V}\hat{\phi}_{i,j,V} + W^T_{i,j,V}\tilde{\phi}_{i,j,V} + \varepsilon_{\mathrm{HJB}_{i,j}}\|\right\}\\ &\le b_{i,j,V,W}\mathbb{E}\left\{\|\hat{\phi}_{i,j,V}\|\right\} + l_{\phi_{i,j,V}}\mathbb{E}\left\{\|W_{i,j,V}\|\right\}b_{i,j,\rho}\\ &\quad + \mathbb{E}\left\{\|\varepsilon_{\mathrm{HJB}_{i,j}}\|\right\} \equiv b_{i,j,V}(t)\end{aligned} \quad (42)$$

with the Lipschitz constants are $l_{\phi_{i,V}}$ and $l_{\phi_{i,j,V}}$.

Proof: Excluded because of page restrictions.

Theorem 2: The mass NN weight $\mathbb{E}\{\hat{W}_{i,j,\rho}\}$ is tuned by (40). The approximated weight error $\mathbb{E}\{\tilde{W}_{i,j,\rho}\}$ and pdf approximation error $\mathbb{E}\{\tilde{\rho}_{i,j}\}$ are UUB. Also, with a minimal reconstruction errors, $\mathbb{E}\{\tilde{W}_{i,j,\rho}\}$ and $\mathbb{E}\{\tilde{\rho}_{i,j}\}$ are asymptotically stable. The bound of $\tilde{\rho}_{i,j}$ is as follows:

$$\begin{aligned}&\mathbb{E}\left\{\|\tilde{\rho}_{i,j}(t)\|\right\}\\ &= \mathbb{E}\left\{\|\tilde{W}^T_{i,j,\rho}\hat{\phi}_{i,j,\rho} + \varepsilon_{\mathrm{FPK}_{i,j}}\|\right\}\\ &\le b_{i,j,\rho,W}\mathbb{E}\left\{\|\hat{\phi}_{i,j,\rho}\|\right\} + \mathbb{E}\left\{\|\varepsilon_{\mathrm{FPK}_{i,j}}\| \equiv b_{i,j,\rho}(t).\right.\end{aligned} \quad (43)$$

Proof: Excluded because of page restrictions.

$$\text{HJB(Leader): } \mathbb{E}\{\Phi_{\mathrm{VM,C}}(x_i,\rho_i)\} = \mathbb{E}\{-\partial_t V_i(x_i,x_{-i},\rho_i) - 0.5{\sigma_i}^2\Delta V_i(x_i,x_{-i},\rho_i) + H_i[x_i,\partial_{x_i}V_i(x_i,x_{-i},\rho_i)]\} \quad (24)$$

$$\text{HJB(Follower): } \mathbb{E}\{\Phi_{\mathrm{CO,F}}(x_{i,j},\rho_i) + \Phi_{\mathrm{VM,F}}(x_{i,j},\rho_i)\} = \mathbb{E}\left\{\begin{bmatrix}-\partial_t V_{i,j}(x_{i,j},x_i,\rho_i) - 0.5{\sigma_i}^2\Delta V_{i,j}(x_{i,j},x_i,\rho_i)\\ +H_{i,j}[x_{i,j},\partial_{x_{i,j}}V_{i,j}(x_{i,j},x_i,\rho_i)]\end{bmatrix}\right\} \quad (25)$$

$$\text{FPK(Follower): } \mathbb{E}\{\partial_t\rho_i(x_{i,j},t) - 0.5{\sigma_i}^2\Delta\rho_i(x_{i,j},t) - div(\rho_i D_p H_{i,j}[x_{i,j},\partial_{x_{i,j}}V_{i,j}(x_{i,j},x_i,\rho_i)])\} = 0 \quad (26)$$

Theorem 3: The actor NN's weights $\mathbb{E}\{\hat{W}_{i,u}\}$ and $\mathbb{E}\{\hat{W}_{i,j,u}\}$ are tuned in (33) and (39). The actor approximated weight errors $\mathbb{E}\{\tilde{W}_{i,u}\}$ and $\mathbb{E}\{\tilde{W}_{i,j,u}\}$ and the approximation errors $\mathbb{E}\{\tilde{u}_i\}$ and $\mathbb{E}\{\tilde{u}_{i,j}\}$ are UUB. Also, with a minimal reconstruction errors, $\mathbb{E}\{\tilde{W}_{i,u}\}$, $\mathbb{E}\{\tilde{W}_{i,j,u}\}$, $\mathbb{E}\{\tilde{u}_i\}$, and $\mathbb{E}\{\tilde{u}_{i,j}\}$ are asymptotically stable. The bounds $\tilde{u}_i$ and $\tilde{u}_{i,j}$ are as follows:

$$\begin{aligned}&\mathbb{E}\Big\{\|\tilde{u}_i(t)\|\Big\}\\&\quad=\mathbb{E}\Big\{\|\tilde{W}_{i,u}^T(t)\hat{\phi}_{i,u}+W_{i,u}^T\tilde{\phi}_{i,u}+\varepsilon_{i,u}\|\Big\}\\&\quad\le b_{i,u,W}\mathbb{E}\Big\{\|\hat{\phi}_{i,u}\|\Big\}+l_{\phi_{i,u}}\mathbb{E}\Big\{\|W_{i,u}\|\|\Big\}\bar{b}_{i,j,\rho}+\mathbb{E}\Big\{\|\varepsilon_{i,u}\|\Big\}\\&\quad\equiv b_{i,u}(t)\end{aligned}\tag{44}$$

and

$$\begin{aligned}\mathbb{E}\Big\{\|\tilde{u}_{i,j}(t)\|\Big\}&=\mathbb{E}\Big\{\|\tilde{W}_{i,j,u}^T\hat{\phi}_{i,j,u}+W_{i,j,u}^T\tilde{\phi}_{i,j,u}+\varepsilon_{i,j,u}\|\Big\}\\&\quad+\mathbb{E}\Big\{\|\varepsilon_{i,j,u}\|\Big\}\\&\le b_{i,j,u,W}\mathbb{E}\Big\{\|\hat{\phi}_{i,j,u}\|\Big\}+l_{\phi_{i,j,u}}\mathbb{E}\Big\{\|W_{i,j,u}\|\Big\}b_{i,j,\rho}\\\mathbb{E}\Big\{\|\varepsilon_{i,j,u}\|\Big\}&\equiv b_{i,j,u}(t)\end{aligned}\tag{45}$$

with the Lipschitz constants $l_{\phi_{i,u}}$ and $l_{\phi_{i,j,u}}$.

Proof: Excluded because of page restrictions.

Moreover, to analyze the close loop stability, we need to consider the leader flock centering and velocity error. The velocity error is represented as

$$e_{i,v}=\sum_{i_{\text{nei}}\in\mathcal{M}_i}a_{i,i_{\text{nei}}}[A_2x_i(t)-A_2x_{i_{\text{nei}}}(t)]\tag{46}$$

where $A_2=\begin{bmatrix}0 & I\end{bmatrix}\in\mathbb{R}^{m\times 2m}$ with $I\in\mathbb{R}^m$ is the identity matrix. Next, the dynamic of the velocity error is defined as

$$\begin{aligned}de_{i,v}&=[f'_{av}(e_{i,v})+g'_{av}(e_{i,v})u_i]dt+\sigma_{i,v}d\omega^l_{i,v}\\&\text{with, } f'_{av}(e_{i,v})\\&\quad=\sum_{i_{\text{nei}}\in\mathcal{M}_i}a_{i,i_{\text{nei}}}A_2\\&\qquad\left[f'_a\Big(\frac{1}{n_{ia}}(e_{i,v}+\sum_{i_{\text{nei}}\in\mathcal{M}_i}a_{i,i_{\text{nei}}}x_{i_{\text{nei}}})\Big)-f'_a(x_{i_{\text{nei}}})\right]\\g'_{av}(e_{i,v})&=\sum_{i_{\text{nei}}\in\mathcal{M}_i}a_{i,i_{\text{nei}}}A_2\\&\qquad\left[g'_a\left(\frac{1}{n_{ia}}\left(e_{i,v}+\sum_{i_{\text{nei}}\in\mathcal{M}_i}a_{i,i_{\text{nei}}}x_{i_{\text{nei}}}\right)\right)-g'_a(x_{i_{\text{nei}}})\frac{u_i}{u_{i_{\text{nei}}}}\right]\end{aligned}\tag{47}$$

where n_{ia} is the ith leaders' neighborhood number.

Lemma 1: There exist optimal control policy [30], that is, u_i and $u_{i,j}$, for the stochastic dynamics (5), (12) and (47), that satisfy

$$\mathbb{E}\Big\{e_i^T\Big[f'_{ar}(e_i)+g'_{ar}(e_i)u_i+\tfrac{\sigma_i d\omega_i}{dt}\Big]\Big\}\le-\gamma_1\mathbb{E}\Big\{\|e_i\|^2\Big\}$$

$$\mathbb{E}\Big\{e_{i,j}^T\Big[f'_{sr}(e_{i,j})+g'_{sr}(e_{i,j})u_{i,j}+\tfrac{\sigma_i d\omega^e_{i,j}}{dt}\Big]\Big\}\le-\gamma_2\mathbb{E}\Big\{\|e_{i,j}\|^2\Big\}$$

$$\mathbb{E}\Big\{e_{i,v}^T\Big[f'_{av}(e_{i,v})+g'_{av}(e_{i,v})u_i+\tfrac{\sigma_{i,v}d\omega_{i,v}}{dt}\Big]\Big\}\le-\gamma_3\mathbb{E}\Big\{\|e_{i,v}\|^2\Big\}$$

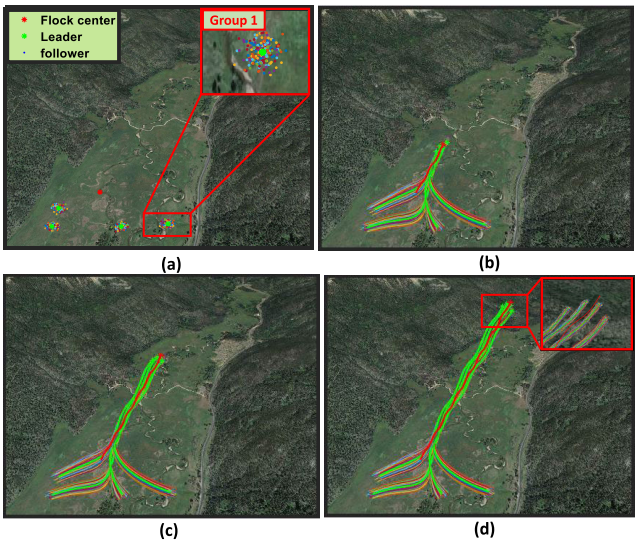


Fig. 3. Multigroup leader–follower flocking. The red curve depicts the trajectory of the flocking center, the leader trajectory is depicted in green color, and other colors are for followers. (a) $t=0$ s. (b) $t=8$ s. (c) $t=14$ s. (d) $t=20$ s.

Theorem 4: The NN weights are updated by the tuning rule (32)–(40). The learning rate of the respective NNs, that is, $\alpha_{i,V}$, $\alpha_{i,u}$, $\alpha_{i,j,V}$, $\alpha_{i,j,\rho}$, and $\alpha_{i,j,u}$ are all positive number. Next, $\mathbb{E}\{\tilde{W}_{i,V}\}$, $\mathbb{E}\{\tilde{W}_{i,u}\}$, $\mathbb{E}\{\tilde{W}_{i,j,V}\}$, $\mathbb{E}\{\tilde{W}_{i,j,\rho}\}$, $\mathbb{E}\{\tilde{W}_{i,j,u}\}$, $\mathbb{E}\{\tilde{V}_i\}$, $\mathbb{E}\{\tilde{V}_{i,j}\}$, $\mathbb{E}\{\tilde{\rho}_{i,j}\}$, $\mathbb{E}\{\tilde{u}_i\}$, $\mathbb{E}\{\tilde{u}_{i,j}\}$, $\mathbb{E}\{e_i\}$, $\mathbb{E}\{e_{i,j}\}$, and $\mathbb{E}\{e_{i,v}\}$ are UUB. Again, $\mathbb{E}\{\tilde{W}_{i,V}\}$, $\mathbb{E}\{\tilde{W}_{i,j,V}\}$, $\mathbb{E}\{\tilde{W}_{i,u}\}$, $\mathbb{E}\{\tilde{W}_{i,j,u}\}$, $\mathbb{E}\{\tilde{W}_{i,j,\rho}\}$, $\mathbb{E}\{\tilde{V}_i\}$, $\mathbb{E}\{\tilde{V}_{i,j}\}$, $\mathbb{E}\{\tilde{u}_i\}$, $\mathbb{E}\{\tilde{u}_{i,j}\}$, $\mathbb{E}\{\tilde{\rho}_{i,j}\}$, $\mathbb{E}\{e_i\}$, $\mathbb{E}\{e_{i,j}\}$, and $\mathbb{E}\{e_{i,v}\}$ are asymptotically stable for the perfectly selected NNs structures.

Proof: Excluded because of page restrictions.

IV. SIMULATION RESULTS

In this part of the study, the developed algorithm is implemented in multigroup large-scale unmanned aerial vehicle (LS-UAV) systems.

A. Hybrid Game-Based Flocking Control

In this experiment, a total number of four groups have been deployed to an area. Each group has a leader with over 300 followers. Besides that, a virtual flocking center trajectory is given to all the leaders. Each leader shares information about their state with the neighborhood leaders. In addition, the states of each leader are broadcast to its corresponding followers and each follower tracks its respective leader to achieve the desired flocking behavior. Let the initial state, that is, the leaders' position and velocity, be $x_1=\begin{bmatrix}3.8 & 1.2 & 0 & 0\end{bmatrix}^T$, $x_2=\begin{bmatrix}4.1 & 1 & 0.3 & 0.15\end{bmatrix}^T$, $x_3=\begin{bmatrix}4.27 & 1.05 & 0.1 & 0.12\end{bmatrix}^T$ and $x_4=\begin{bmatrix}4.4 & 1.3 & 0.5 & 0.3\end{bmatrix}^T$. Now the normal distributions, that are used to produce followers' initial state–Group 1 position: $\mathcal{N}(\mu=[3.8,1.2],\sigma=0.15\times I_2)$, velocity: $\mathcal{N}(\mu=[0,0],\sigma=0.02,\times I_2)$; Group 2 position: $\mathcal{N}(\mu=[4.1,1],\sigma=0.15\times I_2)$, velocity: $\mathcal{N}(\mu=[0.3,0.15],\sigma=0.02\times I_2)$; Group 3 position: $\mathcal{N}(\mu=[4.27,1.05],\sigma=0.15\times I_2)$, velocity: $\mathcal{N}(\mu=[0.1,0.12],\sigma=0.02\times I_2)$; and Group 4 position: $\mathcal{N}(\mu=[4.4,1.3],\sigma=0.15\times I_2)$, velocity: $\mathcal{N}(\mu=[0.5,0.3],\sigma=0.02\times I_2)$. The intrinsic dynamics of the leaders are similar to [28]. Also, the followers' intrinsic dynamics are

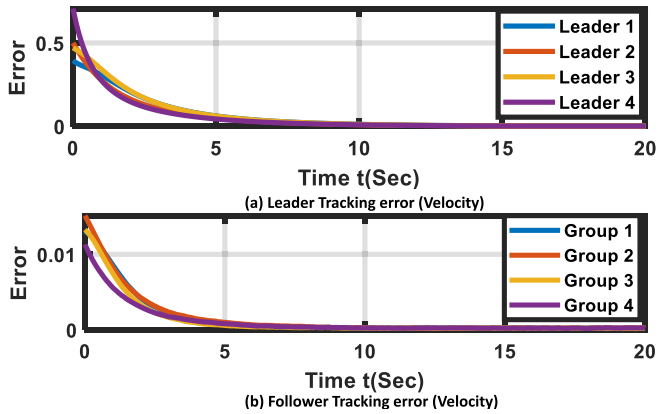


Fig. 4. Leaders and followers' velocity errors (different groups). (a) Leader tracking error (velocity). (b) Follower tracking error (velocity).

given as $f_s'(x_{i,j}) = \left[1/2\, x_{i,j,2} - x_{i,j,1}\ x_{i,j,4} - x_{i,j,3}\ Y\ Z\right]^T$, with $Y = x_{i,j,2}[(\cos(2x_{i,j,1})^2 - 1)] - x_{i,j,1}/2$ and $Z = x_{i,j,4}/2\ [(\cos(2x_{i,j,3} + 2)^2 - 1)] - x_{i,j,3}/2$. Also, $g_s'(x_{i,j}) = \left[0\ 0\ \cos(2x_{i,j,1})\ \cos(2x_{i,j,3}) + 2\right]^T$, with $x_{i,j}(t) = \left[x_{i,j,1}\ x_{i,j,2}\ x_{i,j,3}\ x_{i,j,4}\right]^T$. Furthermore, $h = 1$, $R_i = R_{i,j} = 1$, $w_{\mathrm{C,L}} = 0.8$, $w_{\mathrm{VM,L}} = 1$, $w_{\mathrm{CO,F}} = 1$, $w_{\mathrm{VM,F}} = 1$. The total simulation time of the experiments is 20 s. A hierarchical actor–critic-based NN for estimating the cost, control, and pdf of all the leaders and followers is constructed. The learning rates are $\alpha_{i,V} = 2 \times 10^{-5}$, $\alpha_{i,u} = 2 \times 10^{-4}$, $\alpha_{i,j,V} = 2 \times 10^{-5}$, $\alpha_{i,j,u} = 2 \times 10^{-4}$, and $\alpha_{i,j,\rho} = 2 \times 10^{-3}$. The thresholds parameters are $\delta_{\mathrm{HJB}_i} = 0.00001$, $\delta_{\mathrm{HJB}_{i,j}} = 0.00001$, $\delta_{i,u} = 0.000001$, $\delta_{i,j,u} = 0.000001$, and $\delta_{\mathrm{FPK}_{i,j}} = 0.001$. Fig. 3 shows the trajectory motion of LS-UAVs from different groups with flocking behavior at time $t = 0$ s, $t = 8$ s, $t = 14$ s, and $t = 20$ s. The starting positions of the leaders, followers, and virtual trajectory are depicted in Fig. 3(a). A small window is plotted to show the details of the group $l = 1$ agents starting position. As time progresses, the leaders and their associated followers start their movement in the lower part of the figure and advance upward. Fig. 3 demonstrates that every single leader tracks the virtual flocking center and all of the followers track their corresponding leaders to attain the overall flocking behavior. The velocity matching of the leaders and followers has been verified in Fig. 4. Fig. 4(a) demonstrates the normed velocity error of all four leaders from the respective four groups. The velocity error plotted in this figure is the error between the corresponding leader and the virtual flocking center. It is clear from this demonstration that the velocity error converges to zero with time. This implies that, as time progresses, the leaders from all the groups match their velocity with the flocking center. Similarly, Fig. 4(b) represents the followers' average normed velocity errors with respect to their associated leaders. The four curves shown in the figure represent the average velocity error of the four groups' followers. This figure demonstrates that the velocity errors converge to zero along with time, which implies that the followers match their velocity with their leaders to attain the group flocking behavior. To demonstrate the follower velocity errors in detail and without loss of generality, the followers' error from the group $i = 1$ has been plotted in Fig. 5. This

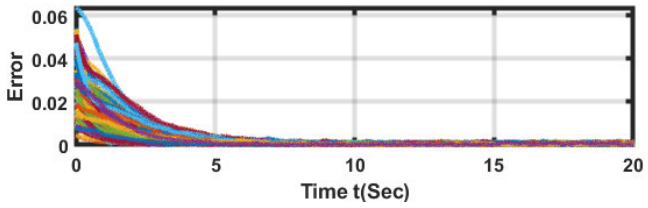


Fig. 5. Follower (group 1) velocity error.

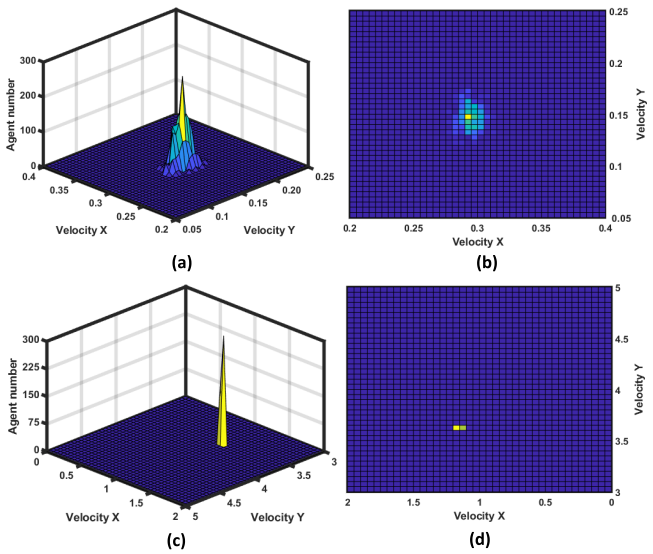


Fig. 6. Velocity of all followers from group 2 with respect to time. (a) and (c) 3-D views; (b) and (d) 2-D Views.

figure demonstrates that the velocity errors of followers within group 1 tend to approach zero as time progresses. Moreover, in Fig. 6, the pdf of the group 2 followers' velocity at different times, is plotted. Fig. 6(a) and (b) and (c) and (d) shows the follower's velocity pdf at time $t = 1$ s and $t = 15$ s in 3-D and 2-D view, respectively. From this figure, it is perfectly shown that the velocity of all followers is matched along with time. The efficacy of the NNs has been evaluated by demonstrating the HJB and FPK errors of the leader and follower. Without sacrificing generality, we take into consideration the leader HJB error and the follower_$i = 1$ HJB error in the group $l = 1$. In Fig. 7(a) and 7(b), the leader–follower error of the HJB equation from group 1 is demonstrated. The figure certainly illustrates that the error converges to zero after a particular period of time. Fig. 7 confirms the optimality of the leader and follower in terms of the cost function. Furthermore, in Fig. 7(c), the mean-field error convergence is verified by plotting the FPK error of follower number one from group one. Here, the FPK error's convergence indicates that the follower's mean field FPK equation is solved. Therefore, the ε-Nash Equilibrium has been obtained.

B. Comparison Result

In this section, we conducted a simulation to demonstrate the significance of our hybrid game-based distributed flocking control against the cooperative control method [31], [32]. Each group includes 1 leader and 200 followers. The agents' states

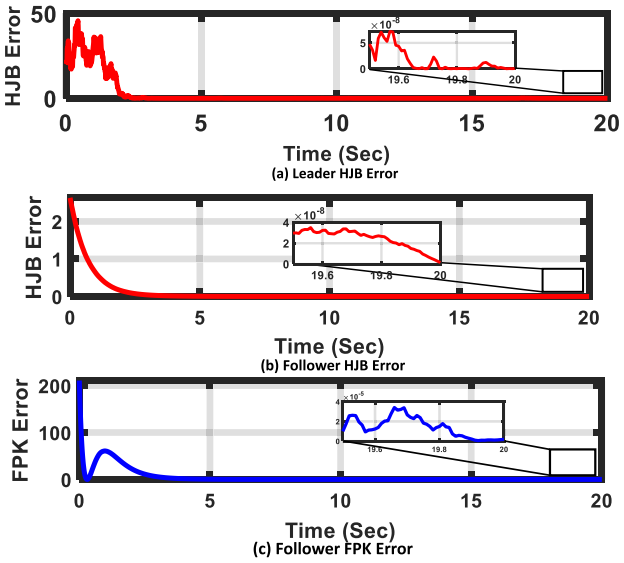


Fig. 7. HJB equation errors. (a) Leader from group_1 HJB error. (b) Follower_1 from group 1 HJB error. (c) Follower FPK error.

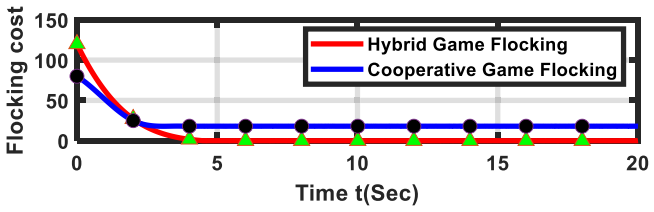


Fig. 8. Leader flocking cost for hybrid and cooperative game.

have been initiated using the same normal distribution provided in the hybrid game-based flocking control scheme. The cost of the leader and follower flocking control is redefined

$$V_i(t) = \mathbb{E}\Big\{\int_0^\infty \begin{bmatrix} \Phi_{\mathrm{LD}}(x_i, x_{-i}, u_i) + \\ \Phi_{\mathrm{VM,C}}(x_i, \rho_i) + w_F N_i. \end{bmatrix} dt\Big\}. \tag{48}$$

The newly added term $w_F N_i$ represents the leader's cost of communication with its corresponding group followers. Here, $w_F = 0.2$ denotes the weight of the communication and N_i represents the followers' number in group i. The performance comparison of the developed algorithm against the traditional algorithm is shown in Fig. 8. Here, the leader from Group 1 is used for this comparison study. This figure clearly demonstrates that the hybrid game-based flocking algorithm outperforms the conventional cooperative flocking approach after a certain period. The new flocking cost of the follower

$$V_{i,j}(t) = \mathbb{E}\Big\{\int_0^\infty \begin{bmatrix} \Phi_{\mathrm{FL}}(x_{i,j}, x_i, u_{i,j}) + \Phi_{\mathrm{CO,F}}(x_{i,j}, \\ \rho_i) + \Phi_{\mathrm{VM,F}}(x_{i,j}, \rho_i) + w_{F,c} N_i. \end{bmatrix} dt\Big\}. \tag{49}$$

Here, the weight $w_{F,c}$ is selected as $w_{F,c} = 0.2$. Fig. 9 clearly shows that the developed method outperforms the traditional cooperative flocking control as time progresses. We also performed another comparison study to show the efficiency of our multigroup decomposition method. In this experiment, we demonstrate the performance analysis of our developed method using the follower agent's velocity pdf for two different systems. The first system consists of a total of

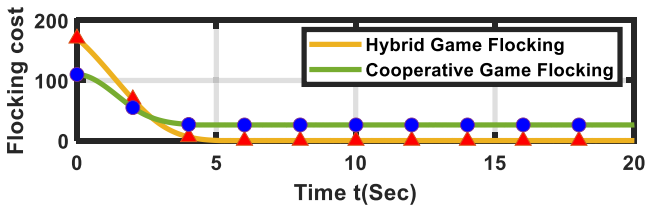


Fig. 9. Follower cost for hybrid and cooperative game.

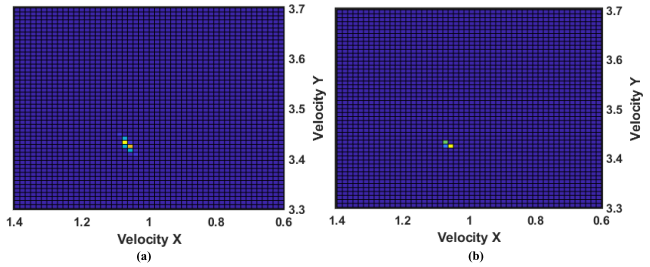


Fig. 10. Followers pdf of velocity for single group and multigroup systems. (a) Single-group velocity pdf (t = 15 s). (b) Multigroup velocity pdf (t = 15 s).

1200 agents, while the second system is the decomposition of the first system into four groups, with each group comprising 300 follower agents. The initial states of single group agents are generated as, position: $\mathcal{N}(\mu = [4.05, 1.10], \sigma = 0.15 \times I_2)$ and velocity: $\mathcal{N}(\mu = [0.2, 0.15], \sigma = 0.02 \times I_2)$. Also, the initial states of the agents from multiple groups are generated as same as in the previous experiment shown in hybrid game-based flocking control. The other parameters are identical to the previous simulation. Fig. 10(a) shows the pdf of followers' velocity in the single group. Similarly, Fig. 10(b) presents the velocity pdf of the 1200 followers, albeit decomposed into four groups. These figures clearly demonstrate that the followers, when divided into four groups, exhibit better velocity matching compared to the single group.

V. Conclusion

This article presents a new hybrid game-based approach to achieve flocking behavior in an LS-MAS under uncertain environments. The proposed method integrates MFG theory and distributed control, decomposing the LS-MAS into subgroups to relax the dependence on a single MFG pdf. A hierarchical structure with cooperative leaders and noncooperative followers is introduced to attain overall flocking behavior. Cooperative games among leaders are developed at a higher level, while a mean field population of noncooperative followers tracks leaders' behavior at a lower level. A Stackelberg game establishes connections between group leaders and followers. A multiactor critic NN architecture is used to solve associated HJB and FPK equations. Lastly, numerical simulation including comparison study and Lyapunov analysis has been performed to demonstrate the efficiency of the developed methodologies. In the near future, the pdf decomposition method to divide the agents into multiple groups will be improved to achieve better efficiency in terms of agent interaction. In addition, the safety of this large-scale system

will be studied in detail and the assumption on stability and convergence theorems will be relaxed.

REFERENCES

[1] A. Okubo, "Dynamical aspects of animal grouping: Swarms, schools, flocks, and herds," *Adv. Biophys.*, vol. 22, pp. 1–94, Jan. 1986. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0065227X86900031

[2] J. T. Emlen, "Flocking behavior in birds," *Auk*, vol. 69, no. 2, pp. 160–170, 1952.

[3] A. A. Paranjape, S.-J. Chung, K. Kim, and D. H. Shim, "Robotic herding of a flock of birds using an unmanned aerial vehicle," *IEEE Trans. Robot.*, vol. 34, no. 4, pp. 901–915, Aug. 2018.

[4] J. Toner and Y. Tu, "Flocks, herds, and schools: A quantitative theory of flocking," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 58, no. 4, pp. 4828–4858, Oct. 1998, doi: 10.1103/PhysRevE.58.4828.

[5] C. W. Reynolds, "Flocks, herds and schools: A distributed behavioral model," *Model. Comput. Graph.*, vol. 21, no. 4, pp. 25–34, 1987, doi: 10.1145/37402.37406.

[6] J. Zhu, J. Lu, and X. Yu, "Flocking of multi-agent non-holonomic systems with proximity graphs," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 60, no. 1, pp. 199–210, Jan. 2013.

[7] F. Wang and Y. Chen, "Fast convergent flocking control of multi-agent systems with switching communication topology," in *Proc. Amer. Control Conf. (ACC)*, Jul. 2020, pp. 695–700.

[8] S. Martin, "Multi-agent flocking under topological interactions," *Syst. Control Lett.*, vol. 69, pp. 53–61, Jul. 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167691114000851

[9] H. Pei, S. Chen, and Q. Lai, "A local flocking algorithm of multi-agent dynamic systems," *Int. J. Control*, vol. 88, no. 11, pp. 2242–2249, Nov. 2015, doi: 10.1080/00207179.2015.1039595.

[10] H.-T. Zhang, Z. Cheng, G. Chen, and C. Li, "Model predictive flocking control for second-order multi-agent systems with input constraints," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 62, no. 6, pp. 1599–1606, Jun. 2015.

[11] R. Olfati-Saber, "Flocking for multi-agent dynamic systems: Algorithms and theory," *IEEE Trans. Autom. Control*, vol. 51, no. 3, pp. 401–420, Mar. 2006.

[12] H. Su, X. Wang, and Z. Lin, "Flocking of multi-agents with a virtual leader," *IEEE Trans. Autom. Control*, vol. 54, no. 2, pp. 293–307, Feb. 2009.

[13] J.-M. Lasry and P.-L. Lions, "Mean field games," *Jpn. J. Math.*, vol. 2, no. 1, pp. 229–260, 2007, doi: 10.1007/s11537-007-0657-8.

[14] P. Cardaliaguet, *Notes on Mean Field Games*. Dauphine, France: Univ. of Paris, 2012.

[15] P. Cardaliaguet and A. Porretta, "An introduction to mean field game theory," in *Mean Field Games*. Berlin, Germany: Springer, 2020, pp. 1–158.

[16] A. Khan, B. Rinner, and A. Cavallaro, "Cooperative robots to observe moving targets: Review," *IEEE Trans. Cybern.*, vol. 48, no. 1, pp. 187–198, Jan. 2018.

[17] S. Zhang, T. Li, X. Cheng, J. Li, and B. Xue, "Multi-group formation tracking control for second-order nonlinear multi-agent systems using adaptive neural networks," *IEEE Access*, vol. 9, pp. 168207–168215, 2021.

[18] Z. Wu, X. Liu, J. Sun, and X. Wang, "Multi-group formation tracking control via impulsive strategy," *Neurocomputing*, vol. 411, pp. 487–497, Oct. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231220310171

[19] G. Jing, Y. Zheng, and L. Wang, "Flocking of multi-agent systems with multiple groups," *Int. J. Control*, vol. 87, no. 12, pp. 2573–2582, Dec. 2014, doi: 10.1080/00207179.2014.935485.

[20] J. Wang and M. Xin, "Flocking of multi-agent systems using a unified optimal control approach," *J. Dyn. Syst., Meas., Control*, vol. 135, no. 6, Aug. 2013, Art. no. 061005, doi: 10.1115/1.4024903.

[21] F. L. Lewis, H. Zhang, K. Hengster-Movric, and A. Das, *Cooperative Control of Multi-Agent Systems: Optimal and Adaptive Design Approaches*. New York, NY, USA: Springer-Verlag, 2014.

[22] Z. Zhou and H. Xu, "Decentralized adaptive optimal tracking control for massive autonomous vehicle systems with heterogeneous dynamics: A Stackelberg game," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 12, pp. 5654–5663, Dec. 2021.

[23] V. François-Lavet, P. Henderson, R. Islam, M. G. Bellemare, and J. Pineau, "An introduction to deep reinforcement learning," *Found. Trends Mach. Learn.*, vol. 11, nos. 3–4, pp. 219–354, 2018, doi: 10.1561/2200000071.

[24] O. Tutsoy, D. E. Barkana, and K. Balikci, "A novel exploration-exploitation-based adaptive law for intelligent model-free control approaches," *IEEE Trans. Cybern.*, vol. 53, no. 1, pp. 329–337, Jan. 2023.

[25] O. Tutsoy and M. Brown, "An analysis of value function learning with piecewise linear control," *J. Experim. Theor. Artif. Intell.*, vol. 28, no. 3, pp. 529–545, May 2016.

[26] J. J. Murray, C. J. Cox, G. G. Lendaris, and R. Saeks, "Adaptive dynamic programming," *IEEE Trans. Syst., Man, Cybern. C, (Appl. Rev.)*, vol. 32, no. 2, pp. 140–153, May 2002.

[27] H. Shiri, J. Park, and M. Bennis, "Massive autonomous UAV path planning: A neural network based mean-field game theoretic approach," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2019, pp. 1–6.

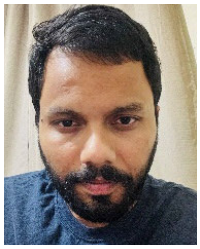
[28] Z. Zhou and H. Xu, "A novel mean-field-game-type optimal control for very large-scale multiagent systems," *IEEE Trans. Cybern.*, vol. 52, no. 6, pp. 5197–5208, Jun. 2022.

[29] F. L. Lewis and D. Vrabie, "Reinforcement learning and adaptive dynamic programming for feedback control," *IEEE Circuits Syst. Mag.*, vol. 9, no. 3, pp. 32–50, Aug. 2009.

[30] H. Xu, S. Jagannathan, and F. L. Lewis, "Stochastic optimal control of unknown linear networked control system in the presence of random delays and packet losses," *Automatica*, vol. 48, no. 6, pp. 1017–1030, Jun. 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0005109812001033

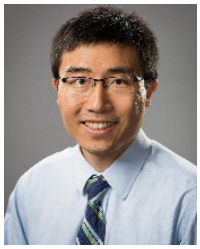
[31] J. Wu et al., "Autonomous cooperative flocking for heterogeneous unmanned aerial vehicle group," *IEEE Trans. Veh. Technol.*, vol. 70, no. 12, pp. 12477–12490, Dec. 2021.

[32] A. Oroojlooy and D. Hajinezhad, "A review of cooperative multi-agent deep reinforcement learning," *Appl. Intell.*, vol. 53, no. 11, pp. 13677–13722, 2023.



Shawon Dey received the M.S. degree in electrical engineering from South Dakota Mines, Rapid City, SD, USA, in 2020. He is currently pursuing the Ph.D. degree with the Electrical and Biomedical Engineering Department, University of Nevada, Reno, Reno, NV, USA.

He is currently a Research Assistant with AIMS Lab, UNR. His research interests include artificial intelligence, autonomous systems, and game theory.



Hao Xu (Member, IEEE) was born in Nanjing, China. He received the master's degree in electrical engineering from Southeast University, Nanjing, in 2009, and the Ph.D. degree in electrical engineering from the Missouri University of Science and Technology, Rolla, MO, USA, in 2012.

He is currently an Associate Professor with the Department of Electrical and Biomedical Engineering, University of Nevada, Reno, Reno, NV, USA. His current research interests include game theory, large-scale multiagent systems, optimization, trusted AI, intelligent control design for advanced power systems, smart grids, autonomous unmanned aircraft systems, and wireless passive sensor networks.