Data-Driven Optimal Formation Control for Quadrotor Team With Unknown Dynamics

Wanbing Zhao, Hao Liu, *Member, IEEE*, Frank L. Lewis, *Life Fellow, IEEE*, and Xinlong Wang

***Abstract*—In this article, the data-driven optimal formation control problem is addressed for a heterogeneous quadrotor team with a virtual leader. Each quadrotor is considered as a highly nonlinear system with six degrees of freedom and the accurate dynamic information of the quadrotor is difficult to obtain in practical applications. An optimal cascade formation controller, including a position controller and an attitude controller, is proposed to track a virtual leader and form a predesigned formation. By using the reinforcement learning (RL) approach, the optimal formation controller is learned from the quadrotor system data without any knowledge of dynamic information of the quadrotors. Simulation results of a heterogeneous multiquadrotor system in a formation flight are given to show the effectiveness of the proposed controllers.**

***Index Terms*—Control, data driven, formation control, multiagent system, nonlinear system, quadrotor.**

List of Notations

$\hat{E}^I = \{\hat{e}_{Io}, \hat{e}_{Ix}, \hat{e}_{Iy}, \hat{e}_{Iz}\}$	Earth-fixed inertial frame.
$\hat{E}^B = \{\hat{e}_{Bo}, \hat{e}_{Bx}, \hat{e}_{By}, \hat{e}_{Bz}\}$	Body frame.
m	Quadrotor mass.
$R_f \in SO(3)$	Rotation matrix relating $\hat{E}^B$ to $\hat{E}^I$.
$J = \text{diag}(J_\phi, J_\theta, J_\psi)$	Quadrotor inertia matrix.
g	Gravity constant.
k_ω, k_t, l_τ	Positive constant parameters.
u_ϕ, u_θ, u_ψ, u_z	Quadrotor control inputs.
$I_N \in \mathbb{R}^{N\times N}$	$N \times N$ unit matrix.
$0_{m\times n} \in \mathbb{R}^{m\times n}$	$m \times n$ zero matrix.
$1_N \in \mathbb{R}^N$	$N \times 1$ vector with 1 as the elements.
$e_{N,j}$	$N \times 1$ vector with 1 in the j th element and 0s elsewhere.

Manuscript received 18 September 2020; accepted 2 January 2021. Date of publication 27 January 2021; date of current version 19 July 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61873012, Grant 61503012, and Grant 61633007; and in part by the Office of Naval Research under Grant N00014-17-1-2239. This article was recommended by Associate Editor A. Gasparri. *(Corresponding author: Hao Liu.)*

Wanbing Zhao and Xinlong Wang are with the School of Astronautics, Beihang University, Beijing 100191, China (e-mail: zhaowb27@buaa.edu.cn; xlwang@buaa.edu.cn).

Hao Liu is with the School of Astronautics, Beihang University, Beijing 100191, China, and also with the Institute of Artificial Intelligence, Beihang University, Beijing, 100191, China (e-mail: liuhao13@buaa.edu.cn).

Frank L. Lewis is with the University of Texas at Arlington Research Institute, University of Texas at Arlington, Fort Worth, TX 76118 USA (e-mail: lewis@uta.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCYB.2021.3049486.

Digital Object Identifier 10.1109/TCYB.2021.3049486

I. Introduction

RECENTLY, unmanned aerial vehicles (UAVs) have witnessed an increased interest in the research community, due to their potential applications in fields, such as disaster relief, forest fire detection, and agriculture industry. As typical UAVs, the quadrotors are popular for their capabilities to take-off and land vertically with simple structures (see [1]–[3]). Furthermore, multiple cooperative UAVs offer superiorities in achieving complicated tasks, for example, disaster monitoring, wilderness search, and cooperative transportation (see [4]–[6]). However, the coordination of a team of UAVs is challenging in uncertain environments. In the latest decade, multiple formation control strategies for quadrotors have been developed, such as the behavior-based controller in [7], the virtual structure approach in [8], and leader–follower methods in [9]–[12]. A dynamic leader is required in the leader–follower controller design and can be virtual, as illustrated in [13]. The virtual leader can either indicate a reference trajectory for the followers or a physical exosystem that is independent from the following agents. In the leader–follower formation control law design of [14], [15], the translational and rotational dynamics was studied for a team of quadrotors, but the nonlinear dynamics of the quadrotors was simplified. Actually, the quadrotor is a highly nonlinear system with six degrees of freedom (6DOF). However, the quadrotor dynamics in [14] and [15] was simplified, and the simplified model cannot accurately describe the nonlinear and coupled dynamics of the quadrotor system with aggressive maneuvers. Therefore, the formation control problem for multiple quadrotors with nonlinear and coupled dynamics remains to be solved.

Furthermore, the complete UAV dynamics is not available in practical applications and, thereby, the obtained vehicle models suffer from system uncertainties. The uncertain vehicle dynamics may deteriorate the formation flight performance of the traditional model-based controllers in [14] and [15]. To solve the uncertain system dynamics, adaptive control methods in [16]–[18] were developed for nonlinear systems, but these approaches could only achieve bounded tracking errors. In [12], [19], and [20], the robust methods were applied to counteract the effects of nonlinearities and system uncertainties on the vehicle dynamics. However, the optimality of the robust controllers in [12], [19], and [20] was generally not guaranteed. It is desirable to achieve data-driven optimal formation control for the quadrotor team without knowledge of the quadrotor dynamics in practical applications.

Reinforcement learning (RL) is a kind of learning method designed to obtain the optimal control policy without knowing the quadrotor dynamics. The RL method has been developed for a single quadrotor to achieve complicated flight tasks as shown in [21]–[25]. In [21], an RL algorithm was applied to identify the dynamic model by using the system data, and a model-based controller was designed to achieve the hover task of a quadrotor. In [22], an RL algorithm was used to plan trajectories to achieve the balancing task for a quadrotor with suspended load but the optimal control problem of the quadrotor was not discussed in [21] and [22]. In [23], a planning framework was designed to achieve construction tasks, and an optimal attitude controller was designed using the RL method. However, the optimal policy was derived offline in [23]. In [24], the neural networks (NNs) were trained using the RL method to control a single quadrotor, but the designed algorithm was conservative and the quadrotor model was simplified. In [25], the nonlinear dynamics was considered for a single quadrotor and an optimal position controller and an optimal attitude controller were trained using the RL approach. However, the coordination of quadrotors was not further investigated in [25]. In the latest years, an RL algorithm has become a promising data-driven optimal method for the synchronization control of multiquadrotor systems (see, e.g., [26] and [27]). In [26] and [27], the attitude synchronization for multiple quadrotors using RL algorithms was achieved, but the nonlinearities in the vehicle dynamics were not considered in [26] and the cascade data-driven formation control for heterogeneous quadrotors was not further studied in [27].

Despite several works on the RL-based control of a single quadrotor, to the best of our knowledge, the RL-based formation control problem for a heterogeneous multiquadrotor system with nonlinearities and couplings in the vehicle dynamics has not been investigated yet. In this article, a distributed formation controller, including a position controller and an attitude controller, is proposed. A data-driven optimal cascade formation controller, including an RL-based position formation controller and an RL-based attitude controller, is proposed. An optimal position controller is obtained by solving the algebraic Riccati equation (ARE), and an RL method is designed to learn the optimal position control law online. An optimal attitude control law is obtained by learning the solution to a Hamilton–Jacobi–Bellman (HJB) equation, similarly. Compared to previously relative studies, the main contributions of this article can be summarized as follows.

First, the optimal control problem of highly nonlinear multiquadrotor systems is solved. The optimal formation policies are obtained for the quadrotor team by using the system data. However, the traditional optimal approaches, such as the methods in [28] and [29], obtained the optimal policies by directly solving AREs, which is difficult for nonlinear quadrotor systems. In the traditional optimal controller design for quadrotors, the optimal policies are derived from an HJB equation. In fact, the HJB equation for the quadrotor system is nonlinear due to the existence of the nonlinear quadrotor dynamics. Therefore, it is difficult or impossible to obtain the optimal policies for the quadrotor team by using the traditional optimal approaches.

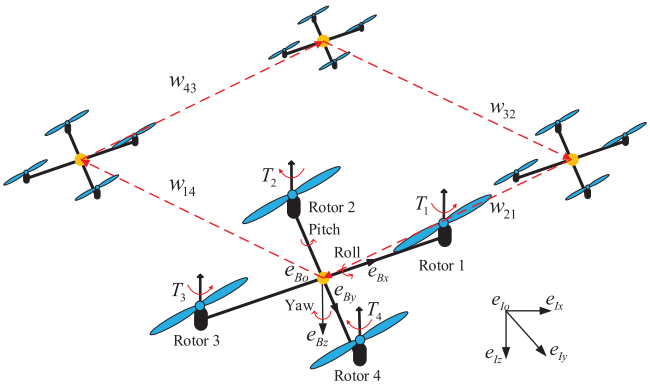

Fig. 1. Quadrotors in a formation team.

Second, the optimality of the formation controller is guaranteed for the heterogeneous multiquadrotor control system with unknown dynamics by using RL methods to learn the optimal policies from system data. In practical formation flight of multiple quadrotors, the quadrotors in the team are heterogeneous due to different payloads. Furthermore, as illustrated in [30], the payloads are unknown for the quadrotor team in each formation flight, which yields unknown dynamics in the quadrotor team. The unknown dynamics will result in an inaccurate HJB equation for the traditional optimal controllers, which makes the optimality of the traditional optimal controller questionable in practical formation flight.

Third, an adaptive cascade controller, including an RL-based position controller and an RL-based attitude controller, is proposed for multiple quadrotors. But the previous works in [21]–[25] only discussed the RL-based control for a single quadrotor, and the RL-based formation control problem for a heterogeneous quadrotor team was not further discussed. Yang *et al.* [26] only discussed the RL-based attitude synchronization control and the RL-based optimal formation control problem for multiple quadrotors was not further studied.

The remaining parts of this article are organized as follows. The preliminaries on the quadrotor model, the graph theory, and the problem formulation are given in Section II. The design method of the formation control protocol is illustrated in Section III. In Section IV, the simulation results are presented to show the effectiveness of the proposed control method. Concluding remarks are contained in Section V.

II. Preliminaries

A. Quadrotor Model

The quadrotors in a formation team are depicted in Fig. 1. As shown in Fig. 1, a rigid frame connects four rotors in each quadrotor system. Denote $p = [p_x \;\; p_y \;\; p_z]^T \in \mathbb{R}^3$ as the position of the quadrotor in $\hat{E}^I$, and $\Theta = [\phi \;\; \theta \;\; \psi]^T \in \mathbb{R}^3$ as the Euler angle (the roll angle ϕ, the pitch angle θ, and the yaw angle ψ). According to [12], the position and attitude dynamical model of the quadrotor system can be described as

$$\begin{aligned} m\ddot{p} &= R_f f \\ J\ddot{\Theta} &= -C\big(\Theta, \dot{\Theta}\big)\dot{\Theta} + \tau \end{aligned} \tag{1}$$

where $C(\Theta, \dot{\Theta}) = [c_{\Theta ij}] \in \mathbb{R}^{3\times 3}$ is the nonlinear Coriolis term depicted in [31]. $R_f \in SO(3)$ in (1) is the rotation matrix shown in [12]. $f \in \mathbb{R}^3$ is the force relative to $\hat{E}^B$ and can be given by $f = [0 \;\; 0 \;\; T_p]^T - R_f^T[0 \;\; 0 \;\; mg]^T$. $T_p \in \mathbb{R}$ indicates the total lift with $T_p = T_1 + T_2 + T_3 + T_4$ in Fig. 1 and $\tau = [\tau_\phi \;\; \tau_\theta \;\; \tau_\psi]^T \in \mathbb{R}^3$ the torque in $\hat{E}^B$ satisfying $T_p = k_\omega u_z$ and $\tau = [l_\tau k_\omega u_\phi \;\; l_\tau k_\omega u_\theta \;\; k_t u_\psi]^T$, respectively. It can be observed from (1) that each quadrotor is a nonlinear and coupled robotic system with 6DOF but 4-input. Besides, in practical flight, a power distribution board can be used to distribute the control inputs u_j ($j = z, \phi, \theta, \psi$) to the four rotors satisfying $u_z = \omega_1^2 + \omega_2^2 + \omega_3^2 + \omega_4^2$, $u_\phi = \omega_2^2 - \omega_4^2$, $u_\theta = \omega_1^2 - \omega_3^2$, and $u_\psi = \omega_1^2 - \omega_2^2 + \omega_3^2 - \omega_4^2$, where ω_j ($j = 1, 2, 3, 4$) is the rotational velocity of Rotor j. Then, the desired rotational velocities of the four rotors in real time can be obtained to generate the desired control inputs and further control the motion of the quadrotor in 6DOF.

B. Graph Theory

N quadrotors are considered in this article labeled from 1 to N. The communication among heterogeneous quadrotors is described by a directed graph $G = (V, E, W)$, where $V = \{v_i\}$ ($i = 1, 2, \ldots, N$) represents a set of nodes, $E \subseteq V \times V$ represents a set of edges, and $W = [w_{ij}] \in \mathbb{R}^{N\times N}$ represents a weighted adjacency matrix. The node v_i represents quadrotor i. If there exists an available channel from node j to node i to exchange information, that is, $(v_i, v_j) \subseteq E$, then $w_{ij} > 0$; otherwise, $w_{ij} = 0$. The set of neighbors of quadrotor i is denoted as $N_i = \{j | (v_i, v_j) \in E\}$. Let $\Phi = \{1, 2, \ldots, N\}$ and the direct graph G is assumed to be time invariant and $w_{ii} = 0$. Define the degree matrix of G as $D = \text{diag}(d_i) \in \mathbb{R}^{N\times N}$ with $d_i = \sum_{j=1}^{N} w_{ij}$ and the Laplacian matrix of G as $L = D - W$. A path of G from v_{i_1} to v_{i_k} is a sequence of edges in the form $\{(v_{i_1}, v_{i_2}), (v_{i_2}, v_{i_3}), \ldots, (v_{i_{k-1}}, v_{i_k})\}$. If there is a node, called the root, connected to any other nodes in G through paths, then G is said to have a spanning tree. Define $\sigma = \text{diag}(\varsigma_i) \in \mathbb{R}^{N\times N}$, where ς_i is the connection weight between quadrotor i and the virtual leader and satisfy that $\varsigma_i > 0$ if quadrotor i has access to the virtual leader, and $\varsigma_i = 0$, otherwise.

C. Problem Formulation

The main goal of this article is to obtain an optimal distributed control law for a team of heterogeneous quadrotors by using the system data to achieve the formation trajectory tracking with a predesigned formation pattern. The formation pattern of quadrotors is designed by vector $\delta_{ij} = [\delta_{ij,x} \;\; \delta_{ij,y} \;\; \delta_{ij,z}]^T \in \mathbb{R}^3$ ($i, j \in \Phi$). δ_{ij} indicates the desired position deviation between quadrotor i and quadrotor j. Let $p_i = [p_{i,x} \;\; p_{i,y} \;\; p_{i,z}]^T \in \mathbb{R}^3$ be the position of the quadrotor i in $\hat{E}^I$ and $p_0 = [p_{0,x} \;\; p_{0,y} \;\; p_{0,z}]^T \in \mathbb{R}^3$ be the desired position of the formation center in $\hat{E}^I$. p_0 can be regarded as the desired position reference of the virtual leader. Denote $\delta_i = [\delta_{i,x} \;\; \delta_{i,y} \;\; \delta_{i,z}]^T \in \mathbb{R}^3$ as the desired deviation of quadrotor i relative to p_0, yielding $\delta_{ij} = \delta_i - \delta_j$. Let $p_{ri} = [p_{ri,x} \;\; p_{ri,y} \;\; p_{ri,z}]^T \in \mathbb{R}^3$ be the position reference for quadrotor i satisfying $p_{ri} = p_0 + \delta_i$, and $e_{pi} = [e_{pi,x} \;\; e_{pi,y} \;\; e_{pi,z}]^T \in \mathbb{R}^3$ the position tracking error with $e_{pi} = p_i - p_{ri}$. In normal cases, the position reference p_{ri} is not directly available for quadrotor i, if quadrotor i has no access to the leader. In this article, the position reference $\hat{p}_{ri}$ is estimated for quadrotor i using the information of its neighbors and itself. From (1), one can rewrite the model for quadrotor i as

$$\begin{aligned} \ddot{p}_i &= m_i^{-1} k_{\omega i} u_{zi} R_{fi} c_{3,3} - g c_{3,3} \\ \ddot{\Theta}_i &= -J_i^{-1}\big(C\big(\Theta_i, \dot{\Theta}_i\big)\dot{\Theta}_i\big) + B_{\Theta i} u_{\Theta i} \end{aligned} \tag{2}$$

where $u_{\Theta i} = [u_{\phi i} \;\; u_{\theta i} \;\; u_{\psi i}]^T \in \mathbb{R}^3$ and $B_{\Theta i} = \text{diag}(b_{\Theta i,1}, b_{\Theta i,2}, b_{\Theta i,3}) = J_i^{-1}\text{diag}(l_{\tau i} k_{\omega i}, l_{\tau i} k_{\omega i}, k_{ti})$. In practical formation flight of the multiquadrotor system, mass m_i and inertia matrix J_i of each quadrotor are different due to different payloads each quadrotor carries, which makes the quadrotors heterogeneous. Furthermore, as illustrated in [30], the payloads are unknown for the quadrotor team in each formation flight, which makes mass m_i and inertia matrix J_i of each quadrotor unknown in practical applications.

III. Formation Protocol Design

In this section, the traditional optimal methods are applied to the position and attitude subsystem of the quadrotors to derive the optimal solutions. Then, it is explained that the optimal solutions are difficult to obtain in practical applications and the RL algorithms are designed to learn the optimal solutions for position and attitude subsystem by using system data.

A. Traditional Position Optimal Controller Design

As shown in Fig. 1, each quadrotor in practical formation is required to rapidly regulate the attitude to manipulate the direction of the total lift T_p and then achieve the desired position formation flight. Therefore, the position control input u_{pi} is virtual in practical applications and is normally considered as the projection of the total lift control input u_{zi} in the earth-fixed frame $\hat{E}^I$ with

$$u_{pi} = u_{zi} R_{fi} e_{3,3}. \tag{3}$$

Then, one can have the position dynamic system as

$$\ddot{p}_i = b_{pi} u_{pi} - g e_{3,3} \tag{4}$$

where $b_{pi} = m_i^{-1} k_{\omega i} I_3$. Define $r_{pi} = [p_{ri,x} \;\; \dot{p}_{ri,x} \; p_{ri,y} \; \dot{p}_{ri,y} \; p_{ri,z} \; \dot{p}_{ri,z}]^T \in \mathbb{R}^6$ as the position reference vector of quadrotor i, $x_{p0} = [p_{0,x} \;\; \dot{p}_{0,x} \;\; p_{0,y} \;\; \dot{p}_{0,y} \;\; p_{0,z} \;\; \dot{p}_{0,z}]^T \in \mathbb{R}^6$ as the leader state vector, and $x_{pi} = [p_{i,x} \;\; \dot{p}_{i,x} \;\; p_{i,y} \;\; \dot{p}_{i,y} \;\; p_{i,z} \;\; \dot{p}_{i,z}]^T \in \mathbb{R}^6$ as the quadrotor state vector. Under the assumption that there exists at least one spanning tree among the quadrotors, a distributed position controller for the quadrotor position system (4) is designed as

$$\begin{aligned} \dot{\hat{r}}_{pi} &= A_p \hat{r}_{pi} + \chi \sum_{j\in N_i} \big(w_{ij}\big(\hat{r}_{pj} - \hat{r}_{pi} + \bar{\delta}_{ij}\big) + \varsigma_i\big(x_{p0} + \bar{\delta}_i - \hat{r}_{pi}\big)\big) \\ u_{pi} &= K_{1i} x_{pi} + K_{2i} \hat{r}_{pi} \end{aligned} \tag{5}$$

where $A_p = \text{diag}(a_p, a_p, a_p) \in \mathbb{R}^{6\times 6}$, $a_p = [0_{2\times 1} \;\; e_{2,1}]$, $\bar{\delta}_i = [\delta_{i,x} \;\; 0 \;\; \delta_{i,y} \;\; 0 \;\; \delta_{i,z} \;\; 0]^T$, and $\bar{\delta}_{ij} = [\delta_{ij,x} \;\; 0 \;\; \delta_{ij,y} \;\; 0 \;\; \delta_{ij,z} \;\; 0]^T$.

$\hat{r}_{pi} \in \mathbb{R}^6$ in (5) is the estimated value of r_{pi}, χ is a positive scaling factor to guarantee that $\hat{r}_{pi}$ converge to r_{pi}, and $K_i = [K_{1i} \ K_{2i}] \in \mathbb{R}^{1\times 12}$ is the position controller parameter. $\bar{\delta}_{ij}$ in (5) is predesigned according to the desired formation. It can be observed from (5) that the proposed position controller is distributed for each quadrotor because only the information of its neighbors and itself is used as the feedback of the controller. Combining the estimated leader state and the quadrotor state, one can obtain a position augmented subsystem as follows:

$$\begin{aligned} \dot{X}_{pi} &= \bar{A}_{pi}X_{pi} + \bar{B}_{pi}u_{pi} + Z_{pi}\varepsilon_{pi} \\ e_{pi} &= \bar{C}_{pi}X_{pi} \end{aligned} \tag{6}$$

where $X_{pi} = [x_{pi}^T \ \hat{r}_{pi}^T]^T$, $\bar{A}_{pi} = \text{diag}(A_p, A_p)$, $\bar{B}_{pi} = m_i^{-1}k_{\omega i}[e_{12,2} \ e_{12,4} \ e_{12,6}]$, $Z_{pi} = [0_{6\times 6} \ \chi I_6]^T$, $\bar{C}_{pi} = [C_{pi} \ -C_{pri}]$, and $C_{pi}, C_{pri} \in \mathbb{R}^{3\times 6}$ are the output matrices of the quadrotor and the leader with $C_{pi} = C_{pri} = [e_{6,1} \ e_{6,3} \ e_{6,5}]^T$. $\varepsilon_{pi} \in \mathbb{R}^6$ is the local neighborhood error of quadrotor i satisfying $\varepsilon_{pi} = \sum_{j\in N_i}(w_{ij}(\hat{r}_{pj} - \hat{r}_{pi} + \bar{\delta}_{ij}) + \varsigma_i(x_{p0} + \bar{\delta}_i - \hat{r}_{pi}))$. ε_{pi} will converge to 0 if χ is selected properly, which will be discussed in the following section. To achieve the optimal control of quadrotor position system, a performance function for the augmented system (6) can be given by

$$V_{pi}(X_{pi}) = \int_t^\infty X_{pi}^T(\bar{Q}_{pi} + K_i^T R_{pi}K_i)X_{pi}d\tau = X_{pi}^T\Pi_{pi}X_{pi} \tag{7}$$

where $\bar{Q}_{pi} = \bar{C}_{pi}^T Q_{pi}\bar{C}_{pi}$ and Π_{pi} is a matrix depending on X_{pi}. $Q_{pi} > 0$ and $R_{pi} = R_{pi}^T > 0$ are the weighting matrices. Define $K_{1i} \in \mathbb{R}^{1\times 6}$ and $K_{2i} \in \mathbb{R}^{1\times 6}$ as the control gains of the controller. Then, by minimizing the performance function in (7), one can obtain the optimal control input u_{pi}^* as

$$u_{pi}^* = K_i X_{pi} \tag{8}$$

where $K_i = [K_{1i} \ K_{2i}] = -R_{pi}^{-1}\bar{B}_{pi}^T\Pi_{pi}$ and Π_{pi} is the solution to the following ARE as:

$$\bar{A}_{pi}^T\Pi_{pi} + \Pi_{pi}\bar{A}_{pi} + \bar{Q}_{pi} - \Pi_{pi}\bar{B}_{pi}R_{pi}^{-1}\bar{B}_{pi}^T\Pi_{pi} = 0. \tag{9}$$

The ARE in (9) requires the knowledge of the vehicle position dynamics. However, due to the unknown dynamics resulted from the unknown payloads, it is difficult to obtain the optimal position control policy by simply solving the ARE in (9). In the following section, an RL algorithm is designed to learn the optimal policy without requirement of the vehicle position dynamics.

B. RL Algorithm for Position Controller

In order to obtain the solution to the ARE in (9), a data-driven RL method is used to learn the optimal control u_{pi}^* on-line. The augmented system (6) can be rewritten as

$$\dot{X}_{pi} = S_{pi}X_{pi} + \bar{B}_{pi}\left(u_{pi}^0 - K_i^n X_{pi}\right) + Z_{pi}\varepsilon_{pi} \tag{10}$$

where $S_{pi} = \bar{A}_{pi} + \bar{B}_{pi}K_i^n$, u_{pi}^0 is a stabilizing and exploring control input, and K_i^n is the updated control parameter in the nth iteration. Mathematical derivatives of (10) yield the Bellman equation as

$$\begin{aligned} &X_{pi}^T(t+\Delta t)\Pi_{pi}^n X_{pi}(t+\Delta t) - X_{pi}^T(t)\Pi_{pi}^n X_{pi}(t) \\ &= -\int_t^{t+\Delta t} X_{pi}^T\bar{Q}_{pi}X_{pi}d\tau - \int_t^{t+\Delta t}\left(K_i^n\right)^T R_{pi}K_i^n d\tau \\ &\quad + 2\int_t^{t+\Delta t}\left(K_i^{n+1}X_{pi}\right)^T R_{pi}\left(u_{pi}^0 - K_i^n X_{pi}\right)d\tau \\ &\quad + \int_t^{t+\Delta t}\left(\Delta V_{pi}^n\right)^T Z_{pi}\varepsilon_{pi}d\tau. \end{aligned} \tag{12}$$

Algorithm 1 Data-Driven RL Position Control Algorithm

1. Initialization

Let $u_{pi}^0 = K_i^0 X_{pi} + u_{\text{pei}}$, where K_i^0 is stabilizing and u_{pei} is an exploring input noise. Collect the quadrotor system data and select a positive threshold $\varepsilon_{\lim 1}$.

2. Policy evaluation

For K_i^n, one can obtain Π_{pi}^n and K_i^{n+1} simultaneously by solving the Bellman equation

$$\begin{aligned} &X_{pi}^T(t+\Delta t)\Pi_{pi}^n X_{pi}(t+\Delta t) - X_{pi}^T(t)\Pi_{pi}^n X_{pi}(t) \\ &= -\int_t^{t+\Delta t} X_{pi}^T\bar{Q}_{pi}X_{pi}d\tau - \int_t^{t+\Delta t}\left(K_i^n\right)^T R_{pi}K_i^n d\tau \\ &\quad + 2\int_t^{t+\Delta t}\left(K_i^{n+1}X_{pi}\right)^T R_{pi}\left(u_{pi}^0 - K_i^n X_{pi}\right)d\tau \\ &\quad + \int_t^{t+\Delta t}\left(\Delta V_{pi}^n\right)^T Z_{pi}\varepsilon_{pi}d\tau \end{aligned} \tag{11}$$

where Δt indicates the time interval. The control parameter K_i^n can be improved by solving (11) by using the vehicle system data, which yields a data-driven RL algorithm in Algorithm 1.

3. Policy improvement

Set $K_i^n = K_i^{n+1}$ and go to 2 until $\left\|K_i^n - K_i^{n+1}\right\| < \varepsilon_{\lim 1}$.

The stabilizing and exploring control input u_{pi}^0 is required for the RL methods to generate the system data, as shown in [32]–[35]. In order to obtain an initial stabilizing policy for the quadrotor system, the traditional proportional-integral-derivative (PID) controller can be implemented to stabilize the quadrotor system. Algorithm 1 is based on the sequential evaluation of the parameters Π_{pi}^n and K_i^{n+1} by solving (12) and the improvement of the control parameter K_i^n. It can be observed from (12) that the RL approach in Algorithm 1 differs from the RL method in [32] due to the term $(\Delta V_{pi}^n)^T Z_{pi}\varepsilon_{pi}$ in (12). In fact, according to Theorem 1, the local neighborhood error ε_{pi} ($i \in \Phi$) will converge to 0 and the least squares (LS) method shown in [32] can be applied to obtain Π_{pi}^n and K_i^{n+1} for quadrotor i under the persistence of excitation (PE) condition. The convergence of Algorithm 1 to the optimal policy can be proven in the way similar to [32]. Denote $\Theta_{ri} = [\phi_{ri} \ \ \theta_{ri} \ \ \psi_{ri}]^T$ as the attitude reference for quadrotor i. After the virtual control input u_{pi} is determined, the desired total lift control input u_{zi}, the desired roll angle ϕ_{ri}, and the desired pitch angle θ_{ri} can be obtained from (3) as

$$\begin{aligned} u_{zi} &= u_{pi,z}/\cos\theta_i/\cos\phi_i \\ \phi_{ri} &= \arcsin\left(\left(\cos\phi_i\sin\theta_i\sin\psi_i - u_{pi,y}/u_{zi}\right)/\cos\psi_i\right) \\ \theta_{ri} &= \arcsin\left(\left(u_{pi,x}/u_{zi} - \sin\phi_i\sin\psi_i\right)/\cos\psi_i/\cos\phi_i\right). \end{aligned} \tag{13}$$

The yaw angle reference ψ_{ri} is set as a fixed constant for the quadrotor team to synchronize in practical applications.

After the desired attitudes are obtained from (13), an attitude controller is required to achieve the desired attitude reference tracking.

C. Traditional Attitude Optimal Controller Design

Define $x_{\Theta i} = [\phi_i \ \ \dot{\phi}_i \ \ \theta_i \ \ \dot{\theta}_i \ \ \psi_i \ \ \dot{\psi}_i]^T$ as the state of quadrotor i. From (1), the quadrotor attitude model can be rewritten as

$$\begin{aligned} \dot{x}_{\Theta i} &= F_{\Theta i}(x_{\Theta i}) + B_{\Theta i} u_{\Theta i} \\ y_{\Theta i} &= C_{\Theta i} x_{\Theta i}, \ \ i \in \Phi \end{aligned} \tag{14}$$

where $y_{\Theta i} \in \mathbb{R}^3$ is the output of quadrotor i and $F_{\Theta i}(x_{\Theta i}) \in \mathbb{R}^6$ the nonlinear term with

$$F_{\Theta i}(x_{\Theta i}) = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & -c_{\Theta 11} J_{\phi i}^{-1} & 0 & -c_{\Theta 12} & 0 & -c_{\Theta 13} \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & -c_{\Theta 21} & 0 & -c_{\Theta 22} J_{\theta i}^{-1} & 0 & -c_{\Theta 23} \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & -c_{\Theta 31} & 0 & -c_{\Theta 32} & 0 & -c_{\Theta 33} J_{\psi i}^{-1} \end{bmatrix} x_{\Theta i}.$$

$B_{\Theta i} \in \mathbb{R}^{6\times 3}$ and $C_{\Theta i} \in \mathbb{R}^{3\times 6}$ are the input and output matrices satisfying $B_{\Theta i} = [e_{6,2} b_{\Theta i,1} \ \ e_{6,4} b_{\Theta i,2} \ \ e_{6,6} b_{\Theta i,3}]$ and $C_{\Theta i} = [e_{6,1} \ \ e_{6,3} \ \ e_{6,5}]^T$, respectively. Let $x_{\Theta ri} = [\phi_{ri} \ \ \dot{\phi}_{ri} \ \ \theta_{ri} \ \ \dot{\theta}_{ri} \ \ \psi_{ri} \ \ \dot{\psi}_{ri}]^T$ be the attitude reference vector of quadrotor i, satisfying

$$\begin{aligned} \dot{x}_{\Theta ri} &= F_{\Theta ri}(x_{\Theta ri}) \\ y_{\Theta ri} &= C_{\Theta ri} x_{\Theta ri} \end{aligned} \tag{15}$$

where $C_{\Theta ri} = C_{\Theta i}$, $y_{\Theta ri} \in \mathbb{R}^3$ is the output and $F_{\Theta ri}(x_{\Theta ri}) \in \mathbb{R}^6$ is a smooth function. From (14) and (15), one can have the quadrotor attitude augmented subsystem as

$$\begin{aligned} \dot{X}_{\Theta i} &= \bar{F}_{\Theta}(X_{\Theta i}) + \bar{B}_{\Theta i} u_{\Theta i}, \\ e_{\Theta i} &= \bar{C}_{\Theta i} X_{\Theta i} \end{aligned} \tag{16}$$

where $X_{\Theta i} = [x_{\Theta i}^T \ \ x_{\Theta ri}^T]^T \in \mathbb{R}^{12}$, $\bar{F}_{\Theta}(X_{\Theta i}) = [F_{\Theta i}(x_{\Theta i})^T \ \ F_{\Theta ri}(x_{\Theta ri})^T]^T \in \mathbb{R}^{12}$, $\bar{B}_{\Theta i} = [B_{\Theta i}^T \ \ 0]^T$, $\bar{C}_{\Theta i} = [\, C_{\Theta i} \ \ -C_{\Theta ri} \,]$, and $e_{\Theta i} = [e_{\Theta i,x} \ \ e_{\Theta i,y} \ \ e_{\Theta i,z}]^T \in \mathbb{R}^3$ is the attitude tracking error. The control input $u_{\Theta i}$ is designed to guarantee that $e_{\Theta i}$ converges to 0. Define the performance function of the augmented subsystem in (16) as

$$V_{\Theta i}(X_{\Theta i}, u_{\Theta i}) = \int_t^{\infty} X_{\Theta i}^T \bar{Q}_{\Theta i} X_{\Theta i} + u_{\Theta i}^T R_{\Theta i} u_{\Theta i} d\tau \tag{17}$$

where $\bar{Q}_{\Theta i} = \bar{C}_{\Theta i}^T Q_{\Theta i} \bar{C}_{\Theta i}$, $Q_{\Theta i} = Q_{\Theta i}^T > 0$ and $R_{\Theta i} = R_{\Theta i}^T > 0$. Using (16) and differentiating (17), one can have the Bellman equation as

$$\begin{aligned} H_{\Theta i}(V_{\Theta i}, u_{\Theta i}) \triangleq{}& X_{\Theta i}^T \bar{Q}_{\Theta i} X_{\Theta i} + u_{\Theta i}^T R_{\Theta i} u_{\Theta i} \\ &+ \Delta V_{\Theta i}^T \big(\bar{F}_{\Theta i} + \bar{B}_{\Theta i} u_{\Theta i}\big) = 0 \end{aligned} \tag{18}$$

where $\Delta V_{\Theta i} = \partial V_{\Theta i} / \partial X_{\Theta i}$. Let $V_{\Theta i}^*$ be an optimal solution to (18). An HJB equation can be obtained as

$$\min_{u_{\Theta i}} H_{\Theta i}\big(V_{\Theta i}^*, u_{\Theta i}\big) = 0. \tag{19}$$

By differentiating (19) with respect to the control input $u_{\Theta i}$, one can derive the optimal control input $u_{\Theta i}^*$ as follows:

$$u_{\Theta i}^* = -\frac{1}{2} R_{\Theta i}^{-1} \bar{B}_{\Theta i}^T \Delta V_{\Theta i}^*. \tag{20}$$

Substituting $u_{\Theta i}^*$ in (20) into (19) yields the following HJB equation as:

$$\begin{aligned} &X_{\Theta i}^T \bar{Q}_{\Theta i} X_{\Theta i} + \big(\Delta V_{\Theta i}^*\big)^T \bar{F}_{\Theta i} \\ &\quad - \frac{1}{4} \big(\Delta V_{\Theta i}^*\big)^T \bar{B}_{\Theta i} R_{\Theta i}^{-1} \bar{B}_{\Theta i}^T \Delta V_{\Theta i}^* = 0. \end{aligned} \tag{21}$$

It should be noted that the HJB equation in (21) is nonlinear due to the nonlinearities in the vehicle dynamics. However, as illustrated in [36], it is difficult or impossible to obtain the solution to the nonlinear HJB equation. Moreover, the HJB equation (21) requires the knowledge of the dynamic model of the multiquadrotor system $\bar{F}_{\Theta i}$ and $\bar{B}_{\Theta i}$. However, the accurate system dynamics $\bar{F}_{\Theta i}$ and $\bar{B}_{\Theta i}$ of quadrotors are difficult to obtain in practical applications. Therefore, the following data-driven RL method is used to learn the optimal control law $u_{\Theta i}^*$ online.

D. RL Algorithm for Attitude Controller

One can rewrite the augmented system in (16) as

$$\dot{X}_{\Theta i} = \bar{F}_{\Theta}(X_{\Theta i}) + \bar{B}_{\Theta i} u_{\Theta i}^n + \bar{B}_{\Theta i}\Big(u_{\Theta i}^0 + u_{\Theta ei} - u_{\Theta i}^n\Big), \ \ i \in \Phi \tag{22}$$

where $u_{\Theta i}^0$ is a stabilizing control input and $u_{\Theta ei}$ is an exploring control input noise. Differentiating $V_{\Theta i}$ in (17), one has that

$$\begin{aligned} \dot{V}_{\Theta i}^n &= \big(\Delta V_{\Theta i}^n\big)^T \big(\bar{F}_{\Theta}(X_{\Theta i}) + \bar{B}_{\Theta i} u_{\Theta i}^n\big) \\ &\quad + \big(\Delta V_{\Theta i}^n\big)^T \bar{B}_{\Theta i}\Big(u_{\Theta i}^0 + u_{\Theta ei} - u_{\Theta i}^n\Big) \\ &= -X_{\Theta i}^T \bar{Q}_{\Theta i} X_{\Theta i} - \big(u_{\Theta i}^n\big)^T R_{\Theta i} u_{\Theta i}^n \\ &\quad - 2\Big(u_{\Theta i}^{n+1}\Big)^T R_{\Theta i}\Big(u_{\Theta i}^0 + u_{\Theta ei} - u_{\Theta i}^n\Big). \end{aligned} \tag{23}$$

It can be seen that the quadrotor dynamics is replaced by the quadrotor system data in (23). By integrating (23), one has that

$$\begin{aligned} &V_{\Theta i}^n(X_{\Theta i}(t + \Delta t)) - V_{\Theta i}^n(X_{\Theta i}(t)) \\ &\quad = \int_t^{t+\Delta t} -X_{\Theta i}^T \bar{Q}_{\Theta i} X_{\Theta i} - \big(u_{\Theta i}^n\big)^T R_{\Theta i} u_{\Theta i}^n d\tau \\ &\qquad - \int_t^{t+\Delta t} 2\Big(u_{\Theta i}^{n+1}\Big)^T R_{\Theta i}\Big(u_{\Theta i}^0 + u_{\Theta ei} - u_{\Theta i}^n\Big) d\tau. \end{aligned} \tag{24}$$

It can be observed from (24) that the value function $V_{\Theta i}^n$ and the control protocol $u_{\Theta i}^{n+1}$ can be updated simultaneously by using system data, which yields a data-driven RL algorithm in Algorithm 2. In Algorithm 2, the control protocol $u_{\Theta i}^{n+1}$ is iteratively updated by solving (28) until the convergence is achieved. The convergence of Algorithm 2 to the optimal policy $V_{\Theta i}^*$ and $u_{\Theta i}^*$ can be proven through contradiction shown in [37].

Algorithm 2 Data-Driven RL Attitude Control Algorithm

1. Initialization

Start with a stabilizing control policy $u^0_{\Theta i}$ and an exploring control policy $u_{\Theta ei}$ for quadrotor i. Collect the quadrotor system data and determine a small positive threshold $\varepsilon_{\lim 2}$.

2. Policy evaluation

For any given $u^n_{\Theta i}$, one can have $V^n_{\Theta i}$ and $u^{n+1}_{\Theta i}$ simultaneously by solving the Bellman equation

$$\begin{aligned} &V^n_{\Theta i}(X_{\Theta i}(t+\Delta t)) - V^n_{\Theta i}(X_{\Theta i}(t)) = \\ &\int_t^{t+\Delta t} -X^T_{\Theta i}\bar{Q}_{\Theta i}X_{\Theta i} - \left(u^n_{\Theta i}\right)^T R_{\Theta i}u^n_{\Theta i}d\tau \\ &-\int_t^{t+\Delta t} 2\left(u^{n+1}_{\Theta i}\right)^T R_{\Theta i}\left(u^0_{\Theta i} + u_{\Theta ei} - u^n_{\Theta i}\right)d\tau \end{aligned} \tag{28}$$

3. Policy improvement

Set $u^{n+1}_{\Theta i} = u^n_{\Theta i}$ and go to 2 until $\left\|u^{n+1}_{\Theta i} - u^n_{\Theta i}\right\| < \varepsilon_{\lim 2}$.

In order to approximate $V^n_{\Theta i}$ and $u^{n+1}_{\Theta i}$ in (28), critic–actor NN are constructed as follows:

$$\hat{V}^n_{\Theta i}(X_{\Theta i}) = W_{vi}\varpi_{vi}(X_{\Theta i}) \tag{25}$$

$$\hat{u}^{n+1}_{\Theta i}(X_{\Theta i}) = W_{ui}\varpi_{ui}(X_{\Theta i}),\ i \in \Phi \tag{26}$$

where $\varpi_{vi}(X_{\Theta i}) \in \mathbb{R}^{l_1}$ and $\varpi_{ui}(X_{\Theta i}) \in \mathbb{R}^{l_2}$ are the activation functions and the superscripts l_1 and l_2 represent the numbers of neurons. $W_{vi} \in \mathbb{R}^{1\times l_1}$ and $W_{ui} \in \mathbb{R}^{3\times l_2}$ represent the weighting vectors. $\hat{V}^n_{\Theta i}(X_{\Theta i})$ and $\hat{u}^{n+1}_{\Theta i}(X_{\Theta i})$ are approximations of $V^n_{\Theta i}(X_{\Theta i})$ and $u^{n+1}_{\Theta i}(X_{\Theta i})$, respectively. Let $e_{\Theta ui} = [e_{\Theta ui,1}\ e_{\Theta ui,2}\ e_{\Theta ui,3}]^T = u^0_{\Theta i} + u_{\Theta ei} - u^n_{\Theta i}$ ($i \in \Phi$), $R_{\Theta i} = \text{diag}(R_{\Theta i,1}, R_{\Theta i,2}, R_{\Theta i,3})$, $x_{\Theta i,1} = \varpi_{vi}(X_{\Theta i}(t+\Delta t)) - \varpi_{vi}(X_{\Theta i}(t))$, $x_{\Theta i,2} = 2R_{\Theta i,1}\int_t^{t+\Delta t}\varpi_{ui}(X_i(t))e_{\Theta ui,1}d_\tau$, $x_{\Theta i,3} = 2R_{\Theta i,2}\int_t^{t+\Delta t}\varpi_{ui}(X_i(t))e_{\Theta ui,2}d_\tau$, and $x_{\Theta i,4} = 2R_{\Theta i,3}\int_t^{t+\Delta t}\varpi_{ui}(X_i(t))e_{\Theta ui,3}d_\tau$. Then, the Bellman equation (24) can be rewritten as

$$e_{B\Theta i}(t) + y_{\Theta i} = \bar{W}_{\Theta i}x_{\Theta i} \tag{27}$$

where $y_{\Theta i} = \int_t^{t+\Delta t} -X^T_{\Theta i}\bar{Q}_{\Theta i}X_{\Theta i} - (u^n_{\Theta i})^T R_{\Theta i}u^n_{\Theta i}d_\tau$, $\bar{W}_{\Theta i} = [W_{vi}\ W_{2i,1}\ W_{2i,2}\ W_{2i,3}]$, and $x_{\Theta i} = [x^T_{\Theta i,1}\ x^T_{\Theta i,2}\ x^T_{\Theta i,3}\ x^T_{\Theta i,4}]^T$. $e_{B\Theta i}(t)$ is the approximation error and $W_{2i,m}$ ($m = 1, 2, 3$) is the mth row of the matrix W_{2i}. By using the least-square method, the approximation error $e_{B\Theta i}(t)$ can be minimized under the PE condition.

E. Stability Analysis

Define $\tilde{r}_{pi} = [\tilde{r}_{pi,x}\ \tilde{r}_{vi,x}\ \tilde{r}_{pi,y}\ \tilde{r}_{vi,y}\ \tilde{r}_{pi,z}\ \tilde{r}_{vi,z}]^T \in \mathbb{R}^6$ as the estimation error for quadrotor i with $\tilde{r}_{pi} = r_{pi} - \hat{r}_{pi}$. In this section, the stability of the heterogeneous multiquadrotor system using the proposed optimal formation controller will be analyzed.

Theorem 1: Using the distributed optimal formation controller, including the position control policy (5), (8) and the attitude control policy (20), the estimation errors $\tilde{r}_{pi}$ will converge to 0 if χ in (5) is sufficiently large and the multiquadrotor system is asymptotically stable.

Proof: From (5), one can obtain that

$$\dot{\bar{\tilde{r}}}_p = P_0\bar{\tilde{r}}_p \tag{29}$$

where $\bar{\tilde{r}}_p = [\tilde{r}^T_{p1}\ \tilde{r}^T_{p2}\ \cdots\ \tilde{r}^T_{pN}]^T$ and $P_0 = I_N \otimes A_p - \chi(L+\sigma)\otimes I_6$. Since the spanning tree exists in the quadrotor communication graph G, all the real parts of the eigenvalues of the matrix $L+\sigma$ are positive constants. Then, one can obtain from (29) that P_0 is Hurwitz if χ is large enough. Therefore, one can see that the estimation error $\tilde{r}_{pi}$ ($i \in \Phi$) will converge to 0 on condition that χ is sufficiently large. By differentiating the performance function (7) and using (6) one can obtain that

$$\begin{aligned} \dot{V}_{pi}(X_{pi}) = &X^T_{pi}\Pi_{pi}\left(\bar{A}_{pi}X_{pi} + \bar{B}_{pi}u_{pi} + Z_{pi}\varepsilon_{pi}\right) \\ &+ \left(\bar{A}_{pi}X_{pi} + \bar{B}_{pi}u_{pi} + Z_{pi}\varepsilon_{pi}\right)^T\Pi_{pi}X_{pi}. \end{aligned} \tag{30}$$

Substituting (8) into (30) yields

$$\begin{aligned} \dot{V}_{pi}(X_{pi}) = &X^T_{pi}\Big(\Pi_{pi}\bar{A}_{pi} - \Pi_{pi}\bar{B}_{pi}R^{-1}_{pi}\bar{B}^T_{pi}\Pi_{pi} + \bar{A}^T_{pi}\Pi_{pi}\Big)X_{pi} \\ &- X^T_{pi}\Big(\Pi^T_{pi}\bar{B}_{pi}R^{-T}_{pi}\bar{B}^T_{pi}\Pi_{pi}\Big)X_{pi} + X^T_{pi}\Pi_{pi}Z_{pi}\varepsilon_{pi} \\ &+ \varepsilon^T_{pi}Z^T_{pi}\Pi_{pi}X_{pi}. \end{aligned} \tag{31}$$

From (9) and (31), one can obtain that

$$\begin{aligned} \dot{V}_{pi}(X_{pi}) = &X^T_{pi}\Big(-\bar{Q}_{pi} - \Pi^T_{pi}\bar{B}_{pi}R^{-T}_{pi}\bar{B}^T_{pi}\Pi_{pi}\Big)X_{pi} \\ &+ X^T_{pi}\Pi_{pi}Z_{pi}\varepsilon_{pi} + \varepsilon^T_{pi}Z^T_{pi}\Pi_{pi}X_{pi}. \end{aligned} \tag{32}$$

By selecting the positive scaling factor χ with a sufficiently large value, one can obtain from (32) that ε_{pi} converges to 0 and $\dot{V}_{pi}(X_{pi}) < 0$. Therefore, it can be concluded that the position tracking error e_{pi} converges to 0, that is, the position dynamic subsystem for quadrotor i is asymptotically stable. From (16) and (21), one can obtain that

$$\begin{aligned} \dot{V}^*_{\Theta i} &= \left(\Delta V^*_{\Theta i}\right)^T\left(\bar{F}_{\Theta i} + \bar{B}_{\Theta i}u^*_{\Theta i}\right) \\ &= \frac{1}{4}\left(\Delta V^*_{\Theta i}\right)^T\bar{B}_{\Theta i}R^{-1}_{\Theta i}\bar{B}^T_{\Theta i}\Delta V^*_{\Theta i} - X^T_{\Theta i}\bar{Q}_{\Theta i}X_{\Theta i} \\ &\quad + \left(\Delta V^*_{\Theta i}\right)^T\bar{B}_{\Theta i}u^*_{\Theta i}. \end{aligned} \tag{33}$$

Substituting (20) into (33), one can have that

$$\dot{V}^*_{\Theta i} = -\frac{1}{4}\left(\Delta V^*_{\Theta i}\right)^T\bar{B}_{\Theta i}R^{-1}_{\Theta i}\bar{B}^T_{\Theta i}\Delta V^*_{\Theta i} - X^T_{\Theta i}\bar{Q}_{\Theta i}X_{\Theta i} < 0. \tag{34}$$

It can be seen from (34) that the attitude tracking error $e_{\Theta i}$ converges to 0 and thus the attitude augmented system is asymptotically stable. ■

IV. Simulation Results

Consider the formation control problem for six heterogeneous quadrotors modeled as (1) with the quadrotor parameters shown in Table I.

The quadrotors are required to track the virtual leader with the heading synchronize to 0° and form a regular plane hexagon with the position deviation designed as: $\delta_{21} = [-0.15\ \ -0.52\ \ 0]^T$ m, $\delta_{32} = [-0.30\ \ 0\ \ 0]^T$ m, $\delta_{43} = [-0.15\ \ 0.52\ \ 0]^T$ m, $\delta_{54} = [0.15\ \ 0.52\ \ 0]^T$ m, $\delta_{65} = [0.30\ \ 0\ \ 0]^T$ m, $\delta_{16} = [0.15\ \ -0.52\ \ 0]^T$ m, and $\delta_1 = [-0.30\ \ 0\ \ 0]^T$ m. The information exchange of quadrotors is modeled by a directed graph with the weights as $w_{21} = w_{32} = w_{43} = w_{54} = w_{65} = w_{16} = 1$. The root v_1 has access to the virtual leader, which means that $\varsigma_1 = 1$. The virtual leader is initialized by $p_{r0}(t) =$

TABLE I
HETEROGENEOUS QUADROTOR SYSTEM PARAMETERS

Parameters	$m(kg)$	$J(10^{-3}kg \cdot m^2)$	k_ω	k_t	$g(m/s^2)$	$l_\tau(m)$
Quadrotor1	1.7	$diag(4.8, 4.8, 8.8)$	1	1	9.8	0.2
Quadrotor2	1.8	$diag(4.9, 4.9, 10.0)$	1	1	9.8	0.2
Quadrotor3	1.9	$diag(5.0, 5.0, 10.1)$	1	1	9.8	0.2
Quadrotor4	2.0	$diag(5.1, 5.1, 10.2)$	1	1	9.8	0.2
Quadrotor5	2.1	$diag(5.2, 5.2, 10.3)$	1	1	9.8	0.2
Quadrotor6	2.2	$diag(5.3, 5.3, 10.4)$	1	1	9.8	0.2

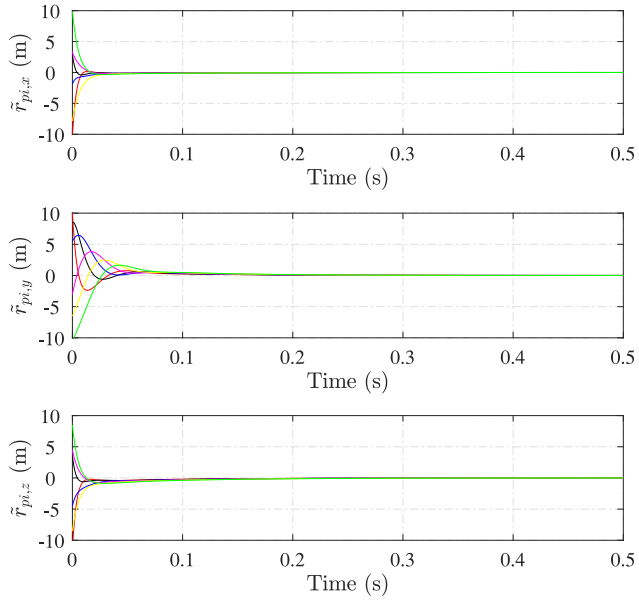

Fig. 2. Position estimation errors.

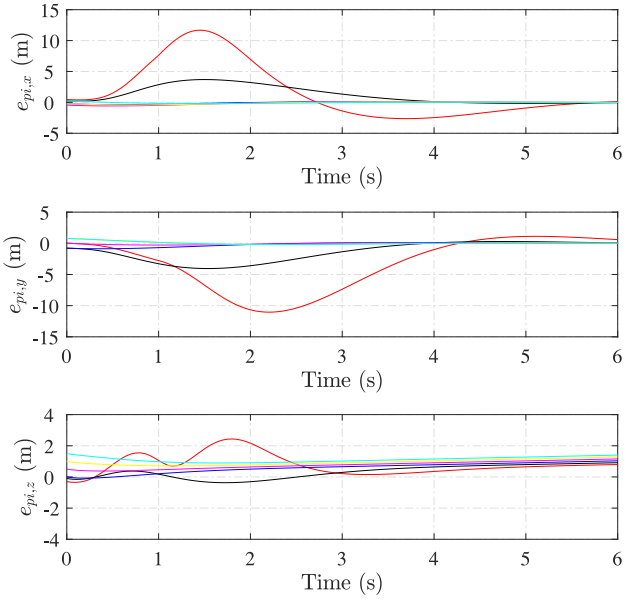

Fig. 3. Position tracking errors of quadrotors using the initial controller.

$[0.5t \;\; 0.5t \;\; 1.5+t]^T$ m and the initial states of the quadrotors are selected as: $p_1(0) = [0.75 \;\; 0 \;\; 1.20]^T$ m, $p_2(0) = [0.38 \;\; -1.30 \;\; 1.40]^T$ m, $p_3(0) = [-0.38 \;\; -1.30 \;\; 1.50]^T$ m, $p_4(0) = [-0.75 \;\; 0 \;\; 2]^T$ m, $p_5(0) = [-0.38 \;\; 1.30 \;\; 2.50]^T$ m, $p_6(0) = [0.38 \;\; 1.30 \;\; 3]^T$ m, $\dot{p}_i = 0_{3\times1}$, $\Theta_i = 0_{3\times1}$, and $\dot{\Theta}_i = 0_{3\times1}$.

First, the position reference of each quadrotor is estimated from (5). The position estimation errors are illustrated in Fig. 2. The six solid lines colored with red, black, blue, pink, yellow, and brilliant blue represent quadrotors 1–6, respectively. It can be observed that the estimation errors for each quadrotor converge to 0 within one 0.1 s. After the position estimation errors converge to 0, the RL-based formation controller is implemented to track the virtual leader and form a predesigned formation pattern. A simple proportional-derivative (PD) controller is selected and manually tuned to obtain an initial stabilizing controller for the quadrotor team with the position tracking errors depicted in Fig. 3 and the attitude tracking errors shown in Fig. 4. From Figs. 3 and 4, one can see that the formation performance of the quadrotor team using the initial manually tuned controller is undesirable. The initial controller is nonoptimal and the objective of the proposed RL algorithms is to learn the optimal control policies by using the initial nonoptimal controller. Then, the

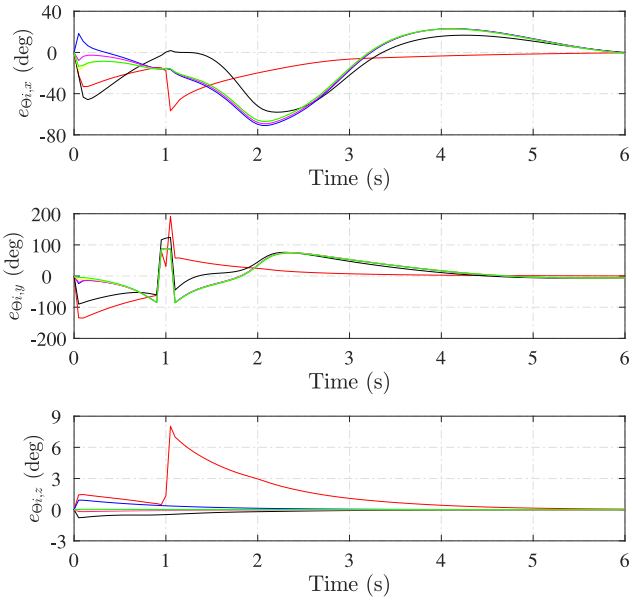

Fig. 4. Attitude tracking errors of quadrotors using the initial controller.

initial stabilizing controller is used to generate system data for Algorithms 1 and 2 with the exploring noises u_{pei} and $u_{\Theta ei}$ selected as the sum of multiple sine waves satisfying $u_{\text{pei}} = 10 \cdot \sum_{m=1}^{10} \sin(\omega_m t)$ and $u_{\Theta ei} = 40 \cdot \sum_{m=1}^{10} \sin(\omega_m t)$,

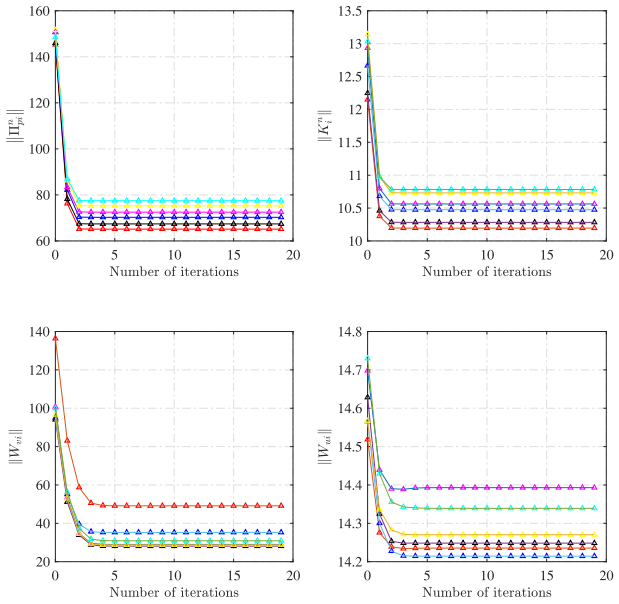

Fig. 5. Convergences of weights for heterogeneous quadrotor team in Algorithms 1 and 2.

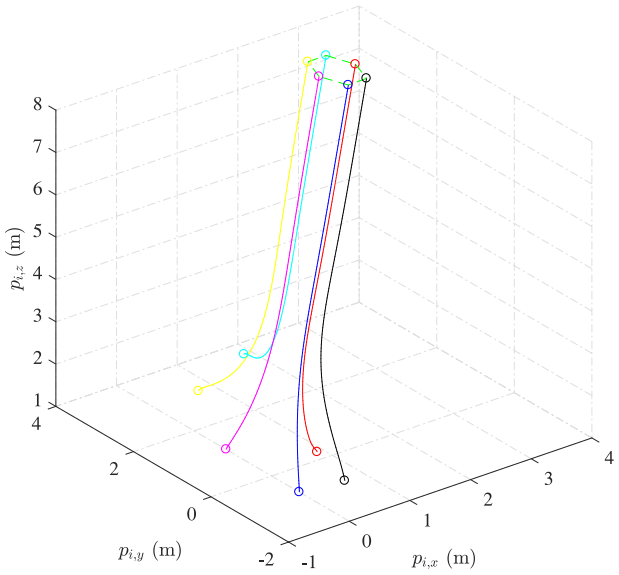

Fig. 6. 3-D trajectories of quadrotors using learned optimal controller.

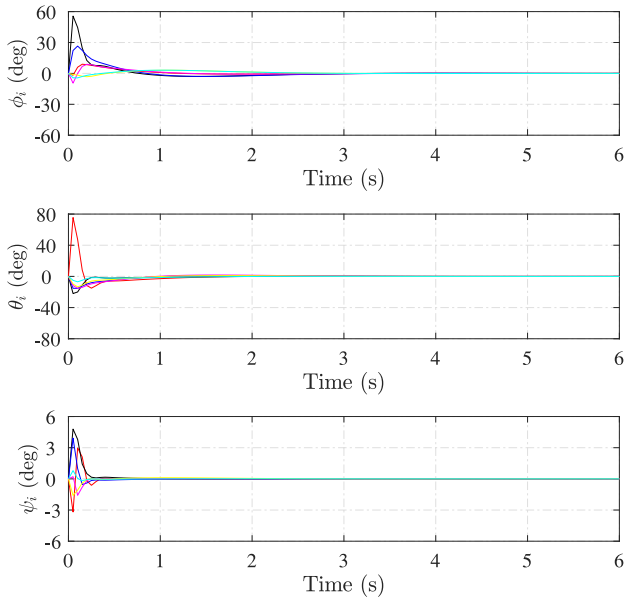

Fig. 7. Attitude responses of quadrotors using the learned optimal controller.

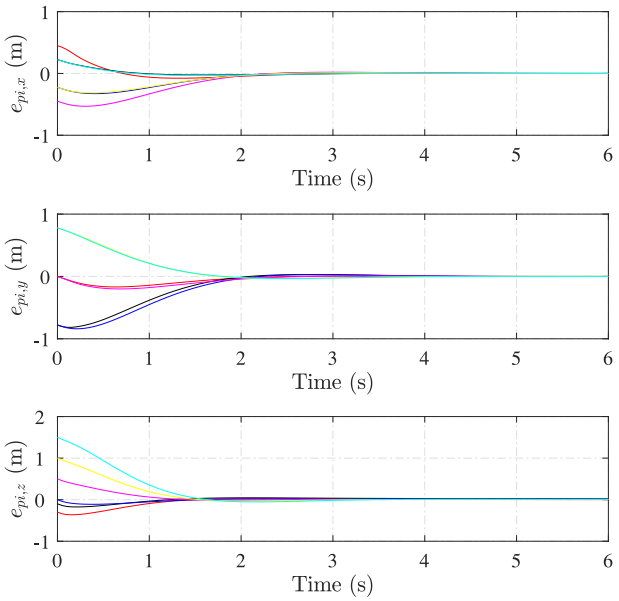

Fig. 8. Position tracking errors of quadrotors using the learned optimal controller.

where ω_i $(m = 1, 2, \ldots, 10)$ are random constants. The time interval Δt is set as $\Delta t = 0.05$ s. The designing matrices of Algorithms 1 and 2 are selected as $Q_{pi} = 20I_6$, $Q_{\Theta i} = 100I_6$, $R_{pi} = I_3$, and $R_{\Theta i} = I_3$. The basis functions of the critic NN in (25) are selected as fourth-order polynomials and the basis function of the actor NN in (26) are selected as first-order polynomials. The convergences of the parameters for Algorithms 1 and 2 are shown in Fig. 5. It can be seen from Fig. 5 that the convergences of all the parameters in the heterogeneous quadrotor team are achieved within five iterations. The 3-D trajectories of the six quadrotors using the learned optimal controller are shown in Fig. 6. The dashed lines in Fig. 6 represent the formation pattern. Fig. 7 shows the Euler angle $\Theta_i(t)$ of the quadrotors using the learned optimal controller. The position tracking errors $e_{pi}(t)$ and the attitude tracking error $e_{\Theta i}(t)$ for the quadrotor team using the learned optimal controller are depicted in Figs. 8 and 9, respectively. As shown in Figs. 6–9, the formation performance of the quadrotor team is greatly improved by applying the learned optimal formation controller. It should be noted that this article is not interested in comparing the performance of the proposed controller with other finely tuned control methods. The advantage of this article is to learn an optimal control policy using a nonoptimal initial controller without knowledge of quadrotor dynamics and consequently improve the formation performance of the quadrotor team.

V. Conclusion

In this article, a data-driven optimal formation control protocol was proposed for heterogeneous quadrotors to track a

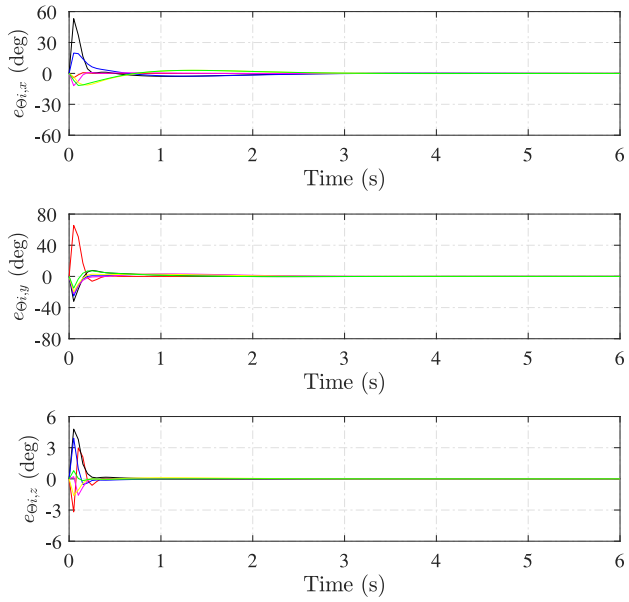

Fig. 9. Attitude tracking errors of quadrotors using the learned optimal controller.

virtual leader and form a predesigned formation. The overall system was composed of a group of quadrotors considered as multiple-input–multiple-output systems with completely and highly nonlinear dynamics. The resulted distributed optimal formation controller includes an optimal position controller and an optimal attitude controller. The position controller was designed to track the virtual leader and form a predesigned formation pattern. Then, the attitude controller was constructed to track the desired attitude. Both the position and attitude controllers were designed based on RL method by learning the solutions to the ARE and the HJB equation respectively, by using the system data. A simulation of heterogeneous multiquadrotor system was conducted and the simulation results verified the effectiveness of the proposed formation controller.

REFERENCES

[1] H. Du, W. Zhu, G. Wen, Z. Duan, and J. L*ü*, "Distributed formation control of multiple quadrotor aircraft based on nonsmooth consensus algorithms," *IEEE Trans. Cybern.*, vol. 49, no. 1, pp. 342–353, Jan. 2019.

[2] H. Liu, Y. Tian, F. L. Lewis, Y. Wan, and K. P. Valavanis, "Robust formation tracking control for multiple quadrotors under aggressive maneuvers," *Automatica*, vol. 105, pp. 179–185, Jul. 2019.

[3] N. Wang, S.-F. Su, M. Han, and W.-H. Chen, "Backpropagating constraints-based trajectory tracking control of a quadrotor with constrained actuator dynamics and complex unknowns," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 49, no. 7, pp. 1322–1337, Jul. 2019.

[4] S. He, M. Wang, S.-L. Dai, and F. Luo, "Leader–follower formation control of USVs with prescribed performance and collision avoidance," *IEEE Trans. Ind. Informat.*, vol. 15, no. 1, pp. 572–581, Jan. 2019.

[5] Y. Cao, W. Yu, W. Ren, and G. Chen, "An overview of recent progress in the study of distributed multi-agent coordination," *IEEE Trans. Ind. Informat.*, vol. 9, no. 1, pp. 427–438, Feb. 2013.

[6] K.-K. Oh, M.-C. Park, and H.-S. Ahn, "A survey of multi-agent formation control," *Automatica*, vol. 53, pp. 424–440, Mar. 2015.

[7] T. Balch and R. C. Arkin, "Behavior-based formation control for multirobot teams," *IEEE Trans. Robot. Autom.*, vol. 14, no. 6, pp. 926–939, Dec. 1998.

[8] M. A. Lewis and K.-H. Tan, "High precision formation control of mobile robots using virtual structures," *Auton. Robots*, vol. 4, pp. 387–403, Oct. 1997.

[9] M. Egerstedt and X. Hu, "Formation constrained multi-agent control," *IEEE Trans. Robot. Autom.*, vol. 17, no. 6, pp. 947–951, Dec. 2001.

[10] M. Mehran and F. Y. Hadaegh, "Formation flying control of multiple space-craft via graphs, matrix inequalities, and switching," *J. Guid. Control Dyn.*, vol. 24, no. 2, pp. 369–377, 2001.

[11] D. A. Mercado, R. Castro, and R. Lozano, "Quadrotors flight formation control using a leader-follower," in *Proc. Eur. Control Conf.*, Jul. 2013, pp. 3858–3863.

[12] H. Liu, T. Ma, F. L. Lewis, and Y. Wan, "Robust formation control for multiple quadrotors with nonlinearities and disturbances," *IEEE Trans. Cybern.*, vol. 50, no. 4, pp. 1362–1371, Apr. 2020.

[13] M. Defoort, A. Polyakov, G. Demesure, M. Djemai, and K. Veluvolu, "Leader-follower fixed-time consensus for multi-agent systems with unknown non-linear inherent dynamics," *IET Control Theory Appl.*, vol. 9, no. 14, pp. 2165–2170, 2015.

[14] E. Zhao, T. Chao, S. Wang, and M. Yang, "Finite-time formation control for multiple flight vehicles with accurate linearization model," *Aerosp. Sci. Technol.*, vol. 71, no. 1, pp. 90–98, Dec. 2017.

[15] C. Hua, J. Chen, and Y. Li, "Leader-follower finite-time formation control of multiple quadrotors with prescribed performance," *Int. J. Syst. Sci.*, vol. 48, no. 12, pp. 2499–2508, Aug. 2017.

[16] I. Palunko, P. Cruz, and R. Fierro, "Agile load transportation: Safe and efficient load manipulation with aerial robots," *IEEE Robot. Autom. Mag.*, vol. 19, no. 3, pp. 69–79, Sep. 2012.

[17] Z. T. Dydek, A. M. Annaswamy, and E. Lavretsky, "Adaptive configuration control of multiple UAVs," *Control Eng. Pract.*, vol. 21, no. 8, pp. 1043–1052, Aug. 2013.

[18] Z. Zhu, Y. Pan, Q. Zhou, and C. Lu, "Event-triggered adaptive fuzzy control for stochastic nonlinear systems with unmeasured states and unknown backlash-like hysteresis," *IEEE Trans. Fuzzy Syst.*, early access, Feb. 13, 2020, doi: 10.1109/TFUZZ.2020.2973950.

[19] B. Zhu, H. H.-T. Liu, and Z. Li, "Robust distributed attitude synchronization of multiple three-DOF experimental helicopters," *Control Eng. Pract.*, vol. 36, no. 3, pp. 87–99, Mar. 2015.

[20] X.-L. Lin, C.-F. Wu, and B.-S. Chen, "Robust H_∞ adaptive fuzzy tracking control for MIMO nonlinear stochastic poisson jump diffusion systems," *IEEE Trans. Cybern.*, vol. 49, no. 8, pp. 3116–3130, Aug. 2019.

[21] N. O. Lambert, D. S. Drew, J. Yaconelli, S. Levine, R. Calandra, and K. S. J. Pister, "Low-level control of a quadrotor with deep model-based reinforcement learning," *IEEE Robot. Autom. Lett.*, vol. 4, no. 4, pp. 4224–4230, Oct. 2019.

[22] A. Faust, I. Palunko, P. Cruz, R. Fierro, and L. Tapia, "Automated aerial suspended cargo delivery through reinforcement learning," *Artif. Intell.*, vol. 247, pp. 381–398, Jun. 2017.

[23] S. R. B. dos Santos, S. N. Givigi, and C. L. Nascimento, "Autonomous construction of multiple structures using learning automata: Description and experimental validation," *IEEE Syst. J.*, vol. 9, no. 4, pp. 1376–1387, Dec. 2015.

[24] J. Hwangbo, I. Sa, R. Siegwart, and M. Hutter, "Control of a quadrotor with reinforcement learning," *IEEE Robot. Autom. Lett.*, vol. 2, no. 4, pp. 2096–2103, Oct. 2017.

[25] C. X. Mu and Y. Zhang, "Learning-based robust tracking control of quadrotor with time-varying and coupling uncertainties," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 1, pp. 259–273, Jan. 2020.

[26] Y. Yang, H. Modares, D. C. Wunsch, and Y. Yin, "Leader–follower output synchronization of linear heterogeneous systems with active leader using reinforcement learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2139–2153. Jun. 2018.

[27] H. Liu, W. Zhao, F. L. Lewis, Z.-P. Jiang, and H. Modares, "Attitude synchronization for multiple quadrotors using reinforcement learning," in *Proc. Chin. Control Conf.*, Jul. 2019, pp. 2480–2483.

[28] H. Zhang, F. L. Lewis, and A. Das, "Optimal design for synchronization of cooperative systems: State feedback, observer and output feedback," *IEEE Trans. Autom. Control*, vol. 56, no. 8, pp. 1948–1952, Aug. 2011.

[29] Nusawardhana, S. H. Zak, and W. A. Crossley, "Nonlinear synergetic optimal controllers," *J. Guid. Control Dyn.*, vol. 30, no. 4, pp. 1134–1147, 2007.

[30] H. Lee and H. J. Kim, "Constraint-based cooperative control of multiple aerial manipulators for handling an unknown payload," *IEEE Trans. Ind. Informat.*, vol. 13, no. 6, pp. 2780–2790, Dec. 2017.

[31] G. V. Raffo, M. G. Ortega, and F. R. Rubio, "An integral predictive/nonlinear H_∞ control structure for a quadrotor helicopter," *Automatica*, vol. 46, no. 1, pp. 29–39, Jan. 2010.

[32] Y. Jiang and Z.-P. Jiang, "Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics," *Automatica*, vol. 48, no. 10, pp. 2699–2704, 2012.

[33] W. Gao and Z.-P. Jiang, "Adaptive dynamic programming and adaptive optimal output regulation of linear systems," *IEEE Trans. Autom. Control*, vol. 61, no. 12, pp. 4164–4169, Dec. 2016.
[34] F. A. Yaghmaie and D. J. Braun, "Reinforcement learning for a class of continuous-time input constrained optimal control problems," *Automatica*, vol. 99, pp. 221–227, Jan. 2019.
[35] B. Pang, Z.-P. Jiang, and I. Mareels, "Reinforcement learning for adaptive optimal control of continuous-time linear periodic systems," *Automatica*, vol. 118, Aug. 2020, Art. no. 109035.
[36] K. Vamvoudakis and F. Lewis, "Online actor–critic algorithm to solve the continuous-time infinite horizon optimal control problem," *Automatica*, vol. 46, no. 5, pp. 878–888, May 2010.
[37] H. -N. Wu and B. Luo, "Neural network based online simultaneous policy update algorithm for solving the HJI equation in nonlinear control," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 12, pp. 1884–1895, Dec. 2012.

Wanbing Zhao received the B.S. degree in guidance, navigation, and control from Beihang University, Beijing, China, in 2016, where he is currently pursuing the Ph.D. degree with the School of Astronautics.

His research interest is primarily in robust control, nonlinear control, reinforcement learning, multiagent systems, and quadrotor flight control.

Hao Liu (Member, IEEE) received the B.E. degree in control science and engineering from Northwestern Polytechnical University, Xi'an, China, in 2008, and the Ph.D. degree in automatic control from Tsinghua University, Beijing, China, in 2013.

In 2012, he was a visiting student with the Research School of Engineering, Australian National University, Canberra, ACT, Australia. Since 2013, he has been with the School of Astronautics, Beihang University, Beijing, where he is currently an Associate Professor. From 2017 to 2018, he was a Visiting Scholar with the University of Texas at Arlington Research Institute, Fort Worth, TX, USA. His research interests include formation control, reinforcement learning, robust control, nonlinear control, unmanned aerial vehicles, unmanned underwater vehicles, and multiagent systems.

Dr. Liu received the Best Paper Award on IEEE ICCA 2018. He serves as an Associate Editor for the *Transactions of the Institute of Measurement and Control*, and the *Advanced Control for Applications: Engineering and Industrial Systems*.

Frank L. Lewis (Life Fellow, IEEE) received the bachelor's degree in physics/EE and the M.S.E.E. degree from Rice University, Houston, TX, USA, in 1971, the M.S. degree in aeronautical engineering from the University of West Florida, Pensacola, FL, USA, in 1977, and the Ph.D. degree in electrical engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 1981.

He is the Moncrief-O'Donnell Chair of the University of Texas at Arlington (UTA) Research Institute, Fort Worth, TX, USA. He has authored seven U.S. patents, numerous journal special issues, journal papers, and 20 books, including *Optimal Control, Aircraft Control, Optimal Estimation,* and *Robot Manipulator Control*, which are used as university textbooks worldwide. His current research interests include feedback control, intelligent systems, cooperative control systems, and nonlinear systems.

Dr. Lewis received the Fulbright Research Award, the NSF Research Initiation Grant, the ASEE Terman Award, the International Neural Network Society Gabor Award, the U.K. Institute of Measurement and Control Honeywell Field Engineering Medal, the IEEE Computational Intelligence Society Neural Networks Pioneer Award, the AIAA Intelligent Systems Award, the Texas Regents Outstanding Teaching Award 2013, the Outstanding Service Award from Dallas IEEE Section, and selected as an Engineer of the year by the Fort Worth IEEE Section. He was listed in Fort Worth Business Press Top 200 Leaders in Manufacturing. He is a member of the National Academy of Inventors, a Fellow of the IFAC and U.K. Institute of Measurement and Control, a Professional Engineer in Texas, a U.K. Chartered Engineer, a UTA Distinguished Scholar Professor, and a UTA Distinguished Teaching Professor.

Xinlong Wang received the B.E. degree in mechanical design and automation from, Beijing University of Aeronautics and Astronautics, Beijing, China, in 1999, and the Ph.D. degree in precision instrument and machinery from Beijing Aerospace Science, Beijing, in 2002.

He is currently a Professor and a Doctoral Supervisor with the School of Astronautics, Beijing University of Aeronautics and Astronautics. His research interests include inertial navigation, celestial navigation, global positioning systems, integrated navigation, as well as aircraft guidance, navigation, and control.