2019 IEEE 58th Conference on Decision and Control (CDC)
Palais des Congrès et des Expositions Nice Acropolis
Nice, France, December 11-13, 2019

Heterogeneous Formation Control of Multiple Rotorcrafts with Unknown Dynamics using Reinforcement Learning*

Hao Liu[1], Fachun Peng[2], Hamidreza Modares[3], Bahare Kiumarsi[4], and Frank L. Lewis[5]

Abstract— **In this paper, a distributed model-free solution to the leader-follower formation control of heterogeneous multi-agent system is proposed using reinforcement learning. The multi-agent system consists of multiple rotorcrafts, including a virtual leader and multiple followers, and no knowledge of the dynamics of leaders and followers is assumed to be known a priori. The formation controller problem is first formulated as an optimal output regulation problem. A discounted performance function is then introduced to guarantee that the tracking error asymptotically converges to zero, and an online off-policy reinforcement learning algorithm is finally proposed to solve the optimal output problem online and using data generated along the agents' trajectories. A simulation example is provided to validate the effectiveness of the proposed control method.**

I. Introduction

In recent years, unmanned aerial vehicle (UAVs) formation control problem has attracted general concern in considerable fields, including communication relay, persistent surveillance, topographic survey, and large-scale search (see, [1]-[5]). Unmanned rotorcrafts are capable of taking off and landing vertically, and hovering at a certain height. Moreover, heterogeneous multiple rotorcraft systems can combine the advantages on different configurations of rotorcrafts (see, [6]), and have received increasing attention.

Multi-agent system control has converted from the centralized to the distributed in recent years. Each distributed formation agent just needs the limited knowledge about itself and its neighbors in the control protocol design(see, [7], [8]). Therefore, the distributed structure has the advantages of simple computation, easy realization, and strong adaptability (see, [9], [10]), and many researches on the distributed formation control methods have been developed. Heterogeneous distributed formation control focuses on the agents with different dynamics, causing great concern for researchers. The heterogeneity in formation makes control more parsimonious, budget less considered, and mission more diversified, as illustrated in [11]-[13]. Existing approaches in previous papers mainly depend on full dynamic information of the formation agents.

Furthermore, in practical applications, unknown dynamics and uncertainties cannot be easily ignored in the flight controller design of unmanned rotorcraft systems. Adaptive control protocol and robust control protocol are two of the methods to deal with the system uncertainties on the formation agent dynamics. In [14], an adaptive distributed time-varying gain controller was designed to solve the output formation-tracking problem with the systems considered as general linear heterogeneous. In [15], for a multi-agent system consisting of a leader and followers, an approximate optimal controller was designed to solve to the cooperative adaptive leader-follower control problem. In [16], an adaptive formation control law based on the smooth projection algorithm and the Lyapunov stability theory was developed to guarantee that the formation error could converge to a given neighborhood. However, the adaptive control approach only yields a bounded error and cannot guarantee the asymptotic synchronization. Besides, it is difficult to obtain an analytical solution to the output regulation problem, and the optimal convergence value of this distributed solution cannot be guaranteed.

Moreover, robust controllers can solve the control problems subject to uncertainties. In [17], a robust trajectory-tracking controller was designed by using neural network approximation to guarantee that the following robots could track the virtual vehicle.In [18], a suboptimal H_∞ controller was established for a leader-follower quadrotor formation to restrain the influence of external disturbances and parameter uncertainties. In [19], based on the artificial potential functions and the robust control theory, a decentralized adaptive control protocol was designed to solve the robot formation problem, effectively restricting approximation errors and external disturbances. However, the nominal system models of the leader and followers are necessary in the robust controller design.

In the heterogeneous multi-rotorcraft systems, the dynamics of each rotorcraft is complex and different, because the rotorcrafts can be installed with different loads to carry out different tasks. Therefore, it is difficult to obtain complete

*This work was supported by the National Natural Science Foundation of China under Grants 61873012, 61503012, and 61633007, and the Office of Naval Research under Grant N00014-17-1-2239.

[1]H. Liu is with the School of Astronautics, Beihang University, Beijing 100191, P.R. China, and also with the Key Laboratory of Spacecraft Design Optimization and Dynamic Simulation Technologies of Ministry of Education, Beihang University, Beijing 100191, P.R. China. liuhao13@buaa.edu.cn

[2]F. Peng is with the School of Astronautics, Beihang University, Beijing 100191, P.R. China, and also with the Key Laboratory of Spacecraft Design Optimization and Dynamic Simulation Technologies of Ministry of Education, Beihang University, Beijing 100191, P.R. China. pfc0321@buaa.edu.cn

[3]H. Modares is with the Department of Mechanical Engineering, Michigan State University, East Lansing, Michigan 48824, USA modaresh@msu.edu

[4]B. Kiumarsi is with the Department of Electrical and Computer Engineering, Michigan State University, East Lansing, Michigan 48824, USA kiumarsi@msu.edu

[5]F. L. Lewis is with the University of Texas at Arlington Research Institute, University of Texas at Arlington, Fort Worth, Texas 76118, USA lewis@uta.edu

978-1-7281-1398-2/19/$31.00 ©2019 IEEE

vehicle dynamics in real multi-rotorcraft formation systems. In this paper, the newly developed reinforcement learning (RL) algorithm, as shown in [20], [21], is proposed to address the heterogeneous multi-rotorcraft formation control problem with completely unknown dynamics. The rotorcraft system is composed of a virtual leader and heterogeneous followers, and the communication relationship between the rotorcrafts is described by a weighted directed graph. A discounted performance function is constructed for each rotorcraft, and an online off-policy RL law is developed to solve the discounted algebraic Riccati equations (AREs) without any knowledge of dynamic information of the virtual leader or the followers. The new contributions of this paper can be given as follows.

First, formation control can be achieved for the multiple rotorcrafts, and the optimal tracking performance can be guaranteed by the proposed method. Second, different loads installed on each rotorcraft results in a heterogeneous system, and the proposed method can address the heterogeneous formation control problem without any dynamic vehicle information. Third, the proposed formation control law is distributed, and each rotorcraft just needs the limited knowledge about itself and its neighbors, which makes it comparatively easy to be implemented in practical applications.

This paper can be organized as follow. In Section 2, the theoretical background, preliminaries on graph theory, and heterogeneous formation model are introduced. In Section 3, a discounted performance function is constructed, and an online off-policy RL law is designed. Section 4 provides simulation examples and conclusion remarks are drawn in Section 5.

II. Background and Preliminaries

2.1 Graph Theory Notations

The communication among the heterogeneous rotorcrafts in formation can be described by weighted directed graph $\mathcal{G}=(V, A, W)$, Let $V = \{v_1, v_2, \cdots v_n\}$ denote a set of n nodes (rotorcrafts), $A \subset V \times V$ a set of edges, and $W = [w_{ij}] \in \mathbb{R}^{n\times n}$ a weighted adjacency matrix. Let $\Pi = \{1, 2, \cdots, n\}$. If $v_{i\to j} \in A$, the information can stream from v_i to v_j, and $w_{ij} > 0$. We assume that there is no self-loop, i.e., $w_{jj} = 0$. $n_i = \{j | v_{i\to j} \in A\}$ is used to describe the neighbor set of node i. Define matrix $D = \operatorname{diag}(d_i) = \operatorname{diag}\left(\sum_{j=1}^{n} w_{ij}\right) \in \mathbb{R}^{n\times n}$ as the in-degree matrix. The Laplacian L matrix can be defined as $L = D - W$. If there exists a sequence of edges, such as $(v_{i\to l}, v_{l\to m}, \cdots, v_{n\to j})$, then v_j is reachable for v_i. In a node that has at least a path to every other node is called the root node, and a spanning tree exists in $\mathcal{G}$ if it has at least one root node. The communication relationship between virtual leader and followers is described by a diagonal matrix $G = \operatorname{diag}(g_i) \in \mathbb{R}^{n\times n}$, and $g_i > 0$ if rotorcraft i has a path to the leader, and $g_i = 0$ otherwise.

Assumption 1: The graph has a spanning tree and the virtual leader is a root node.

2.2 Problem Formulation

Consider the rotorcrafts in the heterogeneous formation as rigid bodies. Let $S_g = \{S_{gx}, S_{gy}, S_{gz}\}$ denote the inertial frame fixed to the earth, and $S_{bi} = \{S_{bxi}, S_{byi}, S_{bzi}\}$ be the body fixed frame with its origin being fixed to the mass center of rotorcraft i. $P_i = \begin{bmatrix} p_{xi} & p_{yi} & p_{zi} \end{bmatrix}^T$ denotes the position vector of rotorcraft i in S_g, and $\Omega_i = \begin{bmatrix} \phi_i & \theta_i & \psi_i \end{bmatrix}^T$ denotes the Euler angle vector of rotorcraft i, where ϕ_i, θ_i, and ψ_i are the roll angle, pitch angle, and yaw angle, respectively.

The rotorcrafts considered in this paper use rotors to generate lifts to achieve different maneuvers. The rotorcrafts possess different configurations, such as helicopter, quadrotor, hexacopter, and octocopter. Generally, the rotorcraft systems can be divided into four subsystems including the longitudinal, lateral, height, and yaw subsystems. Each subsystem can be controlled by a corresponding control input. Let $u_i = \begin{bmatrix} u_{1i} & u_{2i} & u_{3i} & u_{4i} \end{bmatrix}^T \in \mathbb{R}^{4\times 1}$ denote the control input vector generated by the rotors of rotorcraft i to change the lift and torque in S_{bi}. The linear time-invariant dynamics of rotorcraft i is modelled for the four subsystems. The dynamics of the longitudinal is modeled as

$$\begin{aligned} \ddot{\theta}_i &= a_{1xi}\theta_i + a_{2xi}\dot{\theta}_i + b_{xi}u_{1i}, \\ \ddot{p}_{xi} &= a_{3xi}\theta_i + a_{4xi}p_{xi} + a_{5xi}\dot{p}_{xi}, \end{aligned} \tag{1}$$

the dynamics of the lateral as

$$\begin{aligned} \ddot{\phi}_i &= a_{1yi}\phi_i + a_{2yi}\dot{\phi}_i + b_{yi}u_{2i}, \\ \ddot{p}_{yi} &= a_{3yi}\phi_i + a_{4yi}p_{yi} + a_{5yi}\dot{p}_{yi}, \end{aligned} \tag{2}$$

the dynamics of the height as

$$\ddot{\psi}_i = a_{1\psi i}\psi_i + a_{2\psi i}\dot{\psi}_i + b_{\psi i}u_{3i}, \tag{3}$$

and the dynamics of the yaw as

$$\ddot{p}_{zi} = a_{1zi}p_{zi} + a_{2zi}\dot{p}_{zi} + b_{zi}u_{4i}, \tag{4}$$

where the rotorcraft parameters depend on the identification process. Let $(A_{j_1 i}, B_{j_1 i}), j_1 = x, y$ and $(A_{j_2 i}, B_{j_2 i}), j_2 = z, \psi$ denote the dynamics of the four subsystems of rotorcraft i, where

$$A_{j_1 i} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ a_{1j_1 i} & a_{2j_1 i} & 0 & 0 \\ 0 & 0 & 0 & 1 \\ a_{3j_1 i} & 0 & a_{4j_1 i} & a_{5j_1 i} \end{bmatrix}, B_{j_1 i} = \begin{bmatrix} 0 \\ b_{j_1 i} \\ 0 \\ 0 \end{bmatrix},$$

and

$$A_{j_2 i} = \begin{bmatrix} 0 & 1 \\ a_{1j_2 i} & a_{2j_2 i} \end{bmatrix}, B_{j_2 i} = \begin{bmatrix} 0 \\ b_{j_2 i} \end{bmatrix}.$$

Therefore, one can obtain the dynamics of rotorcraft i as

$$\begin{aligned} \dot{x}_i &= A_i x_i + B_i u_i, \\ y_i &= C_i x_i, \end{aligned} \tag{5}$$

where $A_i = \operatorname{diag}(A_{xi}, A_{yi}, A_{zi}, A_{\psi i}) \in \mathbb{R}^{12\times 12}$, $B_i = \operatorname{diag}(B_{xi}, B_{yi}, B_{zi}, B_{\psi i}) \in \mathbb{R}^{12\times 4}$, $C_i = \operatorname{diag}(C_{xi}, C_{yi}, C_{zi}, C_{\psi i}) \in \mathbb{R}^{4\times 12}$, $C_{xi} = C_{yi} = \begin{bmatrix} 0 & 0 & 0 & 1 \end{bmatrix}$, and $C_{zi} = C_{\psi i} = \begin{bmatrix} 0 & 1 \end{bmatrix}$.

Assumption 2: The dynamics of each follower vehicle is stabilizable and observable.

For heterogeneous systems, different types and configurations of the rotorcrafts make (A_i, B_i) different from the others, which poses a challenge for the formation controller design. Let $\varsigma_i \in \mathbb{R}^{8\times 1}$ denote the desired state of rotorcraft i, which includes the translational states P_i, $\dot{P}_i$, and the yaw states ψ_i, $\dot{\psi}_i$. $\varsigma_0 \in \mathbb{R}^{8\times 1}$ is the state of the virtual leader. Let know express the dynamics of the desired state of rotorcraft i as

$$\begin{aligned} \dot{\varsigma}_i &= A_s \varsigma_i, \\ y_{ri} &= C_s \varsigma_i, \end{aligned} \tag{6}$$

where (A_s, C_s) is the same as the dynamics of the virtual leader, and satisfies that

$$A_s = \begin{bmatrix} 0_{4\times4} & I_{4\times4} \\ 0_{4\times4} & 0_{4\times4} \end{bmatrix}, C_s = \begin{bmatrix} I_{3\times3} & 0_{3\times5} \\ 0_{5\times3} & 0_{5\times5} \end{bmatrix}.$$

where $I_{N\times N}$ indicates a $N \times N$ unit matrix and $0_{M\times N}$ a $M \times N$ zero matrix.

The objective of distributed formation control is to guarantee that rotorcrafts form a specific desired formation determined by a set of offset values between agents' desired states, and track the trajectories. Let now

$$\varsigma_i = c_i \left(\sum\nolimits_{j\in n_i} w_{ij} \left(\varsigma_j + \Delta\varsigma_{ij}\right) + g_i \left(\varsigma_0 + \Delta\varsigma_i\right) \right), \tag{7}$$

where $\Delta\varsigma_{ij}$ is the deviation between the desired state of rotorcraft i and j, $\Delta\varsigma_i$ is the deviation between the desired state of rotorcraft i and the state of the virtual leader, and c_i is a scalar.

Problem 1 (Formation Control of Rotorcraft): Consider that the system consists of a virtual leader and multiple followers. Design the control inputs $u_i (i \in \Pi)$ such that the specified states of followers synchronize to the desired states obtained by (7).

To solve *Problem 1*, by selecting appropriate parameters C_i and C_s, one can have that if $y_i \rightarrow y_{ri}$ $(i \in \Pi)$, then the rotorcraft formation can be achieved. In this case, the formation problem is converted to the output synchronization of each rotorcraft. Combining (5) and (6), one can formulate the output regulation equations as follows:

$$\begin{aligned} A_i \chi_i + B_i \Lambda_i &= \chi_i A_s, \\ C_i \chi_i &= C_s, \end{aligned} \tag{8}$$

where $\chi_i \in \mathbb{R}^{12\times 8}$ and $\Lambda_i \in \mathbb{R}^{4\times 8}$ are the solution to equations (8). From [22], one can obtain the following control law to achieve the trajectory tracking of the heterogeneous system as:

$$u_i = \hat{K}_{fi} \left(x_i - \chi_i \varsigma_i\right) + \Lambda_i \varsigma_i, \tag{9}$$

where $\hat{K}_{fi} \in \mathbb{R}^{4\times 12}$ is the state feedback gain to guarantee that $A_i + B_i \hat{K}_{fi}$ is a Hurwitz matrix. To obtain the control input u_i in (9), one requires the complete knowledge of dynamic of the heterogeneous system.

III. Design of Formation Controller via Reinforcement Learning

The objective of this paper is to obtain an optimal control law to guarantee that the formation can be formulated and the center of the formation can follow the trajectory of the virtual leader. Consider a discounted performance function in [23] for rotorcraft i as

$$\begin{aligned} L\left(x_i, u_i\right) = \textstyle\int_t^\infty e^{-\gamma_i(\tau - t)} \big(& \Delta y_i^T Q_i \Delta y_i \\ & + u_i^T R_i u_i \big) \, d\tau, \end{aligned} \tag{10}$$

where $\Delta y_i = C_i x_i - C_s \varsigma_i$. Q_i and R_i are both symmetric and positive definite weight matrices, and the discount factor satisfies $\gamma_i > 0$. γ_i is used to guarantee that the performance. Define the augmented state of rotorcraft i by combining (8) and (10) as

$$X_i = \begin{bmatrix} x_i{}^T & \varsigma_i^T \end{bmatrix}^T \in \mathbb{R}^{20\times 1}. \tag{11}$$

Therefore, the state feedback control law can be designed as

$$u_i = K_{fsi} X_i = K_{fi} x_i + K_{si} \varsigma_i, \tag{12}$$

where $K_{fsi} = \begin{bmatrix} K_{fi} & K_{si} \end{bmatrix}$. Substituting (12) into (10), one can derive the quadratic form of (10) as

$$\begin{aligned} L\left(x_i\right) &= \textstyle\int_t^\infty e^{-\gamma_i(\tau - t)} X_i^T \Big(K_{fsi}^T R_i K_{fsi} \\ & \quad + C_{fsi}^T Q_i C_{fsi} \Big) X_i d\tau \\ &= X_i^T D_i X_i, \end{aligned} \tag{13}$$

where $C_{fsi} = \begin{bmatrix} C_i & -C_s \end{bmatrix}$. Then, the augmented dynamic can be obtained from (5) and (6) as

$$\dot{X}_i = A_{fsi} X_i + B_{fsi} u_i, \tag{14}$$

where

$$A_{fsi} = \begin{bmatrix} A_i & 0 \\ 0 & A_s \end{bmatrix}, B_{fsi} = \begin{bmatrix} B_i \\ 0 \end{bmatrix}. \tag{15}$$

The optimal control input K_i can be obtained by $K_i = -R_i^{-1} B_{fsi}^T D_i$, where D_i is the solution to the following discounted algebraic Riccati equation (ARE) as:

$$\begin{aligned} A_{fsi}^T D_i - \gamma_i D_i + D_i A_{fsi} + C_{fsi}^T Q_i C_{fsi} & \\ - D_i B_{fsi} R_i^{-1} B_{fsi}^T D_i &= 0. \end{aligned} \tag{16}$$

It can be seen that solving (16) depends the complete knowledge of (13).

Theorem 1: If $A_i + B_i K_{1i}$ is a Hurwitz matrix, there exists a positive constant γ_i^* such that for any $\gamma_i \leq \gamma_i^*$, the formation tracking error can be converge to zero asymptotically.

Proof: Let

$$D_i = \begin{bmatrix} D_{11i} & D_{12i} \\ D_{21i} & D_{22i} \end{bmatrix}. \tag{17}$$

From (15) and (17), one can rewrite the discounted ARE (13) as follows:

$$\begin{aligned} A_i^T D_{11i} - \gamma_i D_{11i} + D_{11i} A_i + C_{fsi}^T Q_i C_{fsi} & \\ - D_{11i} B_{fsi} R_i^{-1} B_{fsi}^T D_{11i} &= 0. \end{aligned} \tag{18}$$

Then, one can obtain K_{fi} in (12) as:

$$K_{fi} = -R_i^{-1} B_i^T D_{11i}. \tag{19}$$

From [24], if γ_i satisfies that

$$\gamma_i \leq \gamma_i^* = 2 \left\| \sqrt{B_i R_i^{-1} B_i^T Q_i} \right\|, \tag{20}$$

then $A_i+B_iK_{fi}$ is Hurwitz. Multiplying by X_i^T and X_i on both sides of the equation (16), one has that

$$\begin{aligned} 2X_i^TA_{fsi}^TD_iX_i-\gamma_iX_i^TD_iX_i+X_i^TC_{fsi}^TQ_iC_{fsi}X_i \\ -(D_iX_i)^TB_{fs_i}R_iB_{fsi}^T\left(D_iX_i\right)=0. \end{aligned} \tag{21}$$

$D_iX_i{=}\,0$ yields that $X_i^TC_{fsi}^TQ_iC_{fsi}X_i{=}\,0$ in (21). It means $X_i^TD_iX_i{=}\,0$ that results in $X_i^TC_{fsi}^TQ_iC_{fsi}X_i{=}\,0$. In this case, $(y_i-y_{ri})^TQ_i\left(y_i-y_{ri}\right)=0$, and thereby $y_i-y_{ri}=0$. Consider the Lyaounov function as follows:

$$L_i\left(X_i\right)=X_i^TD_iX_i\geq 0. \tag{22}$$

Its derivative of Lyaounov function in (22) can be given by

$$\dot{L}_i\left(X_i\right)=X_i^T\left(D_i\hat{A}_{fsi}+\hat{A}_{fsi}^TD_i\right)X_i\geq 0, \tag{23}$$

with

$$\hat{A}_{fsi}=\left[\begin{array}{cc} A_i+B_iK_{fi} & B_iK_{si} \\ 0 & A_s \end{array}\right].$$

If γ_i satisfies the inequation (20), $A_i+B_iK_{fi}$ can be guaranteed to be Hurwitz. Since all eigenvalues of A_s are on the imaginary axis, $\hat{A}_{fsi}$ has marginally stability. Therefore, $Q_i\geq 0$ can guarantee that $\dot{L}_i\left(X_i\right)=-X_i^TQ_iX_i\leq 0$. Then, the convergence of the augmented state X_i is the largest invariant subspace, where $\dot{L}_i\left(X_i\right)=0$ from the LaSalle's invariance principle. From (23), if $D_iX_i{=}\,0$, then $\dot{L}_i\left(X_i\right)=0$ and the null space of D_i is attractive. Therefore, $A_i+B_iK_{fi}$ is Hurwitz, and if γ_i satisfies (20) and $X_i^TD_iX_i\neq 0$, then $\dot{L}_i\left(X_i\right)<0$. ■

3.1 Online Off-policy RL Algorithm for Discounted ARE

In this subsection, an online off-policy RL algorithm is used to solve (16) without any knowledge on (14). The augmented dynamics (14) is firstly rewritten as follows:

$$\begin{aligned} \dot{X}_i&=A_{fsi}X_i+B_{fsi}u_i \\ &=A_{fsi}^kX_i+B_{fsi}\left(-K_{fsi}^kX_i+u_i\right), \end{aligned} \tag{24}$$

with $A_{fsi}^k=A_{fsi}+B_{fsi}K_{fsi}^k$. Consider that

$$\begin{aligned} V\left(X_i,D_i\right)=e^{-\gamma_i\Delta t}X_i(t+\Delta t)^TD_i^kX_i\left(t+\Delta t\right) \\ -X_i(t)^TD_i^kX_i\left(t\right). \end{aligned} \tag{25}$$

The right-hand side of (25) can be converted to an integral form as follows:

$$V\left(X_i,D_i\right)=\int_t^{t+\Delta t}\frac{d}{d\tau}\left(e^{-\gamma_i(\tau-t)}{X_i}^TD_i^kX_i\right)d\tau. \tag{26}$$

Let $J\left(X_i\right)=e^{-\gamma_it}X_i(t)^TD_iX_i\left(t\right)$. Considering the exploration noise v_i, one can have that the control input in $[t,t+\Delta t]$ satisfies that $u_i^k=K_{fsi}^kx_i+v_i$. Then combining (16) and (24), one can have the derivative of $J\left(X_i\right)$ from [23] as follows:

$$\begin{aligned} \tfrac{dJ(X_i)}{dt}=2e^{-\gamma t}\left(u_i{-}K_{fsi}^kX_i\right)^T(B_{fsi})^TD_i^kX \\ -e^{-\gamma t}X_i^T\left(C_{fsi}^TQ_iC_{fsi}{+}\left(K_{fsi}^k\right)^TR_iK_{fsi}^k\right)X_i, \end{aligned} \tag{27}$$

where $K_{fsi}^k=R_i^{-1}B_{fsi}^TD_i^{k-1}$. Let $Q_i=C_{fsi}^TQ_iC_{fsi}+\left(K_{fsi}^k\right)^TR_iK_{fsi}^k$. Then, $(B_{fsi})^TD_i^k$ can be substituted by $R_iK_{fsi}^{k+1}$. Now, from (25), (26) and (27), the Bellman equation can be given by

$$\begin{aligned} X_i(t)^TD_i^kX_i\left(t\right)=&e^{-\gamma_i\Delta t}X_i(t+\Delta t)^TD_i^kX_i\left(t+\Delta t\right) \\ &+\int_t^{t+\Delta t}e^{-\gamma_i(\tau-t)}\Big({X_i}^TQ_iX_i \\ &+2\big(K_{fsi}^kX_i-u_i\big)^TR_iK_{fsi}^{k+1}X_i\Big)\,d\tau \end{aligned} \tag{28}$$

From (28), D_i^k and K_i^{k+1} are obtained via the iteration process simultaneously. In the iteration process, both the knowledge of dynamics of the followers (A_i,B_i,C_i) and leader (A_s,C_s) is not required. Although Q_i in (28) is related to C_{fsi} and thus related to the dynamics of followers and virtual leader, Q_i can only be calculated by the output measurement equation satisfying ${X_i}^TQ_iX_i=(y_i-y_{ri})^TQ_i\left(y_i-y_{ri}\right)$. The steps of the off-policy RL algorithm are shown in *Algorithm 1*. The proposed algorithm simultaneously solves the discounted ARE equation (16) and yields K_{fsi}.

Algorithm 1 Online Off-policy RL Algorithm

```
Step 1.   Initialization: $k=0$ and $K_{fsi}^k$ is stabilizing. Then, let $u_i=K_{fsi}^kX_i+\nu$, where $\nu$ is the exploration noise.
Step 2.   Solve $D_i^k$ and $K_{fsi}^{k+1}$ from (28) with ${X_i}^TQ_iX_i=(y_i-y_{ri})^TQ_i\left(y_i-y_{ri}\right)$.
Step 3.   Let $k=k{+}1$, and repeat Steps 2 and 3 until the convergence is achieved.
Step 4.   Let $u_i=K_{fsi}X_i=-R_i^{-1}B_{fsi}^TD_i^kX_i$ as the optimal control input.
```

For *Algorithm 1*, it can be observed that the least squares theory to obtain the solution of (28) can be used to reduce the numerical error as shown in [25]. A fixed control input can be used to acquire enough samples under the persistence of excitation beforehand, which contains the system state and input information of each rotorcraft.

3.2 Online Off-policy RL Algorithm for Heterogeneous Formation

In this subsection, by combining the consensus theory and RL algorithm for optimal trajectory tracking, one can obtain an online off-policy RL algorithm for multi-rotorcraft systems to solve the heterogeneous formation problem without any requirement of knowledge of the global system dynamics. The controller input K_{fsi} in (12) can be obtained from (16). Algorithm 1 is used to obtain the numerical solution of the global discounted ARE equation without the requirement of the dynamics of followers (A_i,B_i,C_i) and the dynamics of the leader (A_s,C_s). Solving the output regulation equation (8) is necessary for the optimal formation control problem, and the output regulation equation (8) can be implicitly solved by Algorithm 1.

IV. SIMULATION RESULTS

In this section, a heterogeneous formation system is studied, consisting of a virtual leader and five heterogeneous followers. For the rotorcraft formation, each rotorcraft is required to track the reference trajectory and keep the yaw angle at 0°. Therefore, the dynamics of the virtual leader does

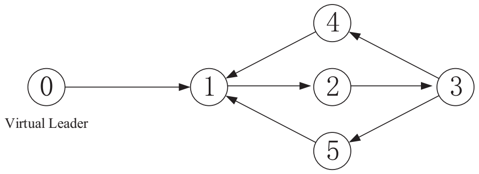

Fig. 1. Communication graph.

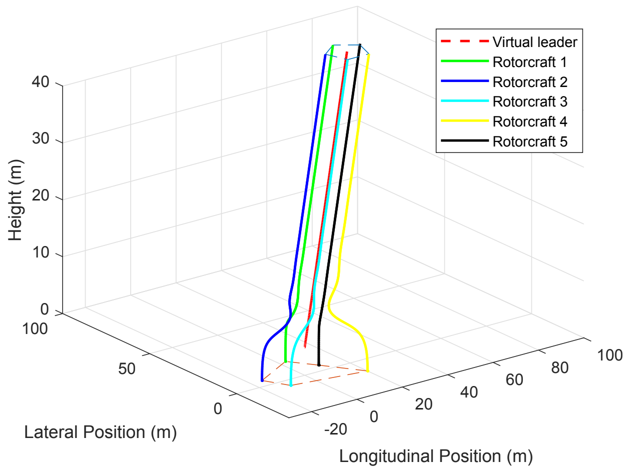

Fig. 2. Evaluation of trajectory tracking of heterogeneous rotorcrafts formation.

not involve the pitching and rolling motions. The initial states of the virtual leader are chosen as $P_0(0) = \begin{bmatrix} 0 & 0 & 5 \end{bmatrix}^T$, $\dot{P}_0 = \begin{bmatrix} 10 & 10 & 4 \end{bmatrix}^T$, and $\begin{bmatrix} \psi_0 & \dot{\psi}_0 \end{bmatrix}^T = \begin{bmatrix} 0 & 0 \end{bmatrix}^T$. The initial states of the followers are selected as $P_1(0) = \begin{bmatrix} 3 & 15 & 0 \end{bmatrix}^T$, $P_2(0) = \begin{bmatrix} -15 & 5 & 0 \end{bmatrix}^T$, $P_3(0) = \begin{bmatrix} -10 & -5 & 0 \end{bmatrix}^T$, $P_4(0) = \begin{bmatrix} 20 & -10 & 0 \end{bmatrix}^T$, $P_5(0) = \begin{bmatrix} 10 & 5 & 0 \end{bmatrix}^T$, $\dot{P}_1(0) = \dot{P}_2(0) = \dot{P}_3(0) = \dot{P}_4(0) = \dot{P}_5(0) = \begin{bmatrix} 0 & 0 & 5 \end{bmatrix}^T$, $\Omega_i(0) = 0_{1\times 3}$, and $\dot{\Omega}_i(0) = 0_{1\times 3}$. The directed weighted graph of the formation system is shown in Fig. 1. The node 0 represents the virtual leader, while the others represent the heterogeneous follower rotorcrafts. The weighted adjacency matrix $W = I_{n\times n}$. The matrices Q_i and R_i are chosen as $Q_i = 100 I_{4\times 4}$ and $R_i = I_{4\times 4}$, respectively. The optimal state feedback gain can be obtained via *Algorithm 1*.

The trajectory tracking control of heterogeneous rotorcrafts formation is shown in Fig. 2. The rotorcrafts can form a formation in a desired pentagon shape. The three-dimensional trajectory tracking error of each follower is shown in Fig. 3. The velocity response of the follower rotorcrafts is depicted in Fig. 4 and the attitude response in Fig. 5. It can be seen that the proposed method can achieve the desired formation control for the heterogeneous rotorcraft system without any knowledge of the virtual leader and the followers.

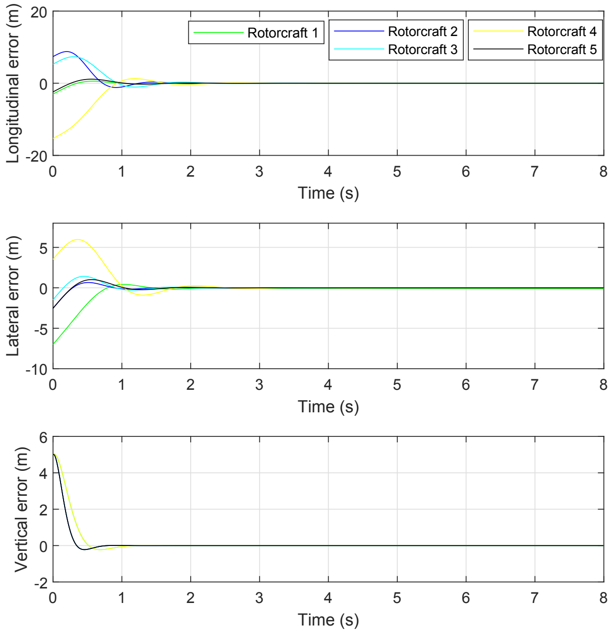

Fig. 3. Position error of trajectory tracking of heterogeneous rotorcrafts formation.

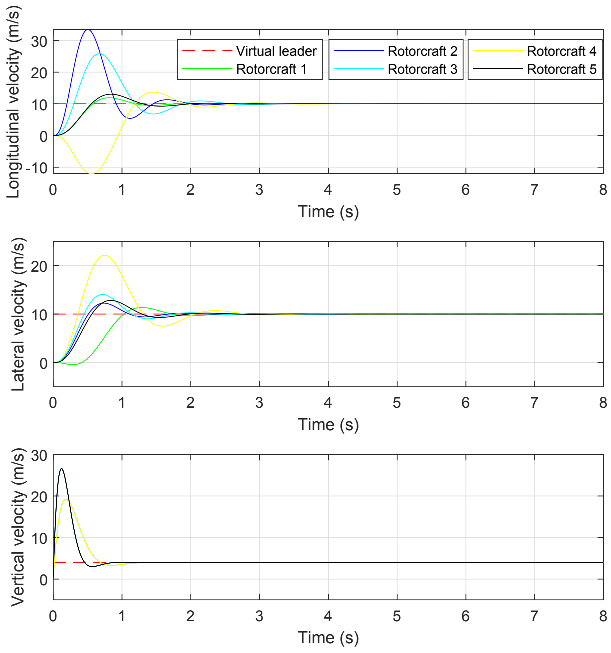

Fig. 4. Velocity response of trajectory tracking of heterogeneous rotorcrafts formation.

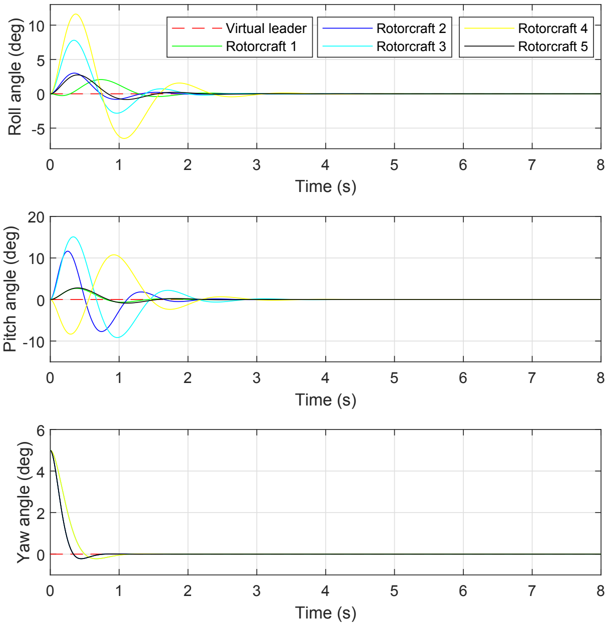

Fig. 5. Attitude response of trajectory tracking of heterogeneous rotorcrafts formation.

V. Conclusions

In this paper, a distributed formation control law based on the reinforcement learning is proposed for the heterogeneous rotorcraft systems without any knowledge of dynamics. A discounted performance function is introduced for each rotorcraft, and it is proven that the discounted performance function can be minimized and the solution of discounted algebraic Riccati equations (AREs) can be obtained. An online off-policy reinforcement learning algorithm is proposed to obtain the numerical solution of the AREs independently of the dynamics of leader or followers. A simulation example is provided to validate the effectiveness of the proposed control method.

References

[1] J. Wang and M. Xin,"Integrated optimal formation control of multiple unmanned aerial vehicles," *IEEE Transactions on Control Systems Technology*, vol. 21, no. 5, pp. 1731 - 1744, Nov. 2012.

[2] N. Nigam, S. Bieniawski, I. Kroo, and J. Vian, "Control of multiple UAVs for persistent surveillance: Algorithm and flight test results," *IEEE Transactions on Control Systems Technology*, vol. 20, no. 5, pp. 1236-1251, Sep. 2012.

[3] W. Meng, Z. He, R. Su, P. Yadav, R. Teo, and L. Xie, "Decentralized multi-UAV flight autonomy for moving convoys search and track," *IEEE Transactions on Control Systems Technology*, vol. 25, no. 4, pp. 1480-1487, Sep. 2016.

[4] P. Sujit and D. Ghose, "Search using multiple UAVs with flight time constraints," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 40, no. 2, pp. 491-509, Apr. 2004.

[5] X. Wang, V. Yadav, and S. N. Balakrishnan, "Cooperative UAV formation flying with obstacle collision avoidance," *IEEE Transactions on Control Systems Technology*, vol. 15, no. 4, pp. 672-679, Jul. 2007.

[6] A. Oller and I. Maza, *Multiple heterogeneous unmanned aerial vehicles*. Springer Berlin Heidelberg, 2007.

[7] F. L. Lewis, H. Zhang, and K. Hengster-Movric, *Cooperative control of multi-agent systems: optimal and adaptive design approaches*. Springer-Verlag, 2014.

[8] Z. Lin, L. Wang, Z. Han, and M. Fu. "Distributed formation control of multi-agent systems using complex laplacian," *IEEE Transactions on Automatic Control*, vol. 59, no. 7, pp. 1765-1777, Jul. 2014.

[9] D. V. Dimaroganas, E. Frazzoli, and K. H. Johansson, "Distributed Event-Triggered Control for Multi-Agent Systems," *IEEE Transaction on Automatic Control*, vol. 57, no. 5, pp. 1291-1297, May. 2012.

[10] A. Jadbabaie, J. Lin, and A. Morse, "Coordination of groups of mobile autonomous agents using nearest neighbor rules," *IEEE Transactions on Automatic Control*, vol. 48, no. 6, pp. 988-1001, Jun. 2003.

[11] W. Hu and L. Liu, "Cooperative output regulation of heterogeneous linear multi-agent systems by event-triggered control," *IEEE Transactions on Cybernetics*, vol. 47, no. 1, pp. 105 - 116, Jan. 2017.

[12] S. Li, J. Zhang, X. Li, F. Wang, X. Luo, and X. Guan, "Formation control of heterogeneous discrete-time nonlinear multi-agent systems with uncertainties," *IEEE Transactions on Industrial Electronics*, vol. 64, no. 6, pp. 4730-4740, Jun. 2017.

[13] F. Gao, S. Li, Y. Zheng, D. Kum, "Robust control of heterogeneous vehicular platoon with uncertain dynamic and communication delay," *IET Intelligent Transport Systems*, vol. 10, no. 7, pp. 503-513, Sep. 2016.

[14] S. Zuo, Y. Song, F. L. Lewis, and A. Davoudi, "Adaptive output formation-tracking of heterogeneous multi-agent systems using time-varying L_2-gain design," *IEEE Control Systems Letters*, vol. 2, no. 2, pp. 236-241, Mar. 2018.

[15] W. Gao, Z. Jiang, F. L. Lewis, and Y. Wang, "Leader-to-formation stability of multiagent systems: an adaptive optimal control approach," *IEEE Transactions on Automatic Control*, vol. 63, no. 10, pp. 3581-3587, Jan. 2018.

[16] K. Choi, S. Yoo, J. park, and Y. Choi, "Adaptive formation control in absence of leader's velocity information," *IET Control Theory and Applications*, vol. 4, no. 4, pp.521-528, Apr. 2010.

[17] J. Ghommam, H. Mehrjerdi, and M. Saad, "Robust formation control without velocity measurement of the leader robot," *Control Engineering Practice*, vol. 21, no. 8, pp.1143-1156, Aug. 2013.

[18] W. Jasim, and D. Gu, "Robust team formation control for quadrotors," *IEEE Transactions on Control Systems Technology*, vol. 26, no. 4, pp.1516-1523, Jul. 2018.

[19] B. Ranjbar-Sahraei, F. Shabaninia, A. Nemati, and S. Stan, "A novel robust decentralized adaptive fuzzy control for swarm formation of multiagent systems," *IEEE Transactions on Industrial Electronics*, vol. 59, no. 8, pp. 3124-3134, Aug. 2012.

[20] F. L. Lewis and D. Vrabie, "Reinforcement learning and adaptive dynamic programming for feedback control," *IEEE Circuits and Systems Magazine*, vol. 9, no. 3, pp.32-50, 2009.

[21] H. Modares, S. P. Nageshrao, G. Lopes, R. Babuska, and F. L. Lewis, "Optimal model-free output synchronization of heterogeneous systems using off-policy reinforcement learning," *Automatica*, vol. 71, pp. 334-341, Sep. 2016.

[22] T. Yang, A. Saberi, A. Stroorvogel, and H. Grip, "Output synchronization for heterogeneous networks of introspective right-invertible agents," *International Journal of Robust and Nonlinear Control*, vol. 24, no. 13, pp.1821-1844, Sep. 2014.

[23] H. Modares and F. L. Lewis, "Linear quadratic tracking control of partially-unknown continuous-time systems using reinforcement learning," *IEEE Transactions on Automatic Control*, vol. 59, no. 11, pp. 3051-3056, Nov. 2014.

[24] H. Modares, F. L. Lewis, and Z.-P. Jiang, "H_∞ tracking control of completely unknown continuous-time systems via off-policy reinforcement learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 10, pp. 2550-2562, Oct. 2015.

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

[26] H. W. Knobloch, A. Isidori, and D. Flockerzi, *Topics in Control Theory*, Springer, 1993.