Containment Control of Heterogeneous Systems with Non-Autonomous Leaders: A Distributed Optimal Model Reference Approach

YONGLIANG YANG¹, (Member, IEEE), SHUSEN CHENG², YIXIN YIN², (MEMBER, IEEE), DONALD C. WUNSCH, II.³(Fellow, IEEE)

¹School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, P. R. China. (e-mail: yangyongliang@ieee.org; yyx@ies.ustb.edu.cn)
²School of Metallurgical and Ecological Engineering, University of Science and Technology Beijing, Beijing 100083, P. R. China. (e-mail: chengsusen@metall.ustb.edu.cn)
³Department of Electrical and Computer Engineering, Missouri University of Science and Technology, Rolla, MO 65409, USA. (e-mail: wunsch@ieee.org)

Corresponding author: Yongliang Yang (e-mail: yangyongliang@ieee.org).

This work was supported in part by the National Natural Science Foundation of China under Grant Number 61333002, in part by the Fundamental Research Funds for the Central Universities under Grant Number FRF-TP-18-031A1, in part by the Mary K. Finley Missouri Endowment, in part by the Missouri S&T Intelligent Systems Center and in part by the Army Research Laboratory under Cooperative Agreement Number W911NF-18-2-0260. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

ABSTRACT This paper presents a distributed optimal model reference adaptive control (DO-MRAC) approach for solving containment control of heterogeneous multi-agent systems (MASs) with non-autonomous leaders. First, a fully distributed adaptive observer is designed to provide for each agent a desired reference trajectory by estimating the convex hull spanned by leaders. The distributed observer dynamics serves as a reference model for each follower to synchronize. The global communication graph information or the leader dynamics is not required to design the observer. In contrast to existing model reference adaptive controllers (MRAC) for single-agent systems and containment control solutions for MASs, the proposed MRAC approach imposes optimality and presents a distributed adaptive optimal solution to the containment control problem. To impose optimality, a performance function is defined based on the adaptive observers' states as well as the followers' local measurements. It is shown that considering non-autonomous leaders in this optimal control problem leads to solving inhomogeneous algebraic Riccati equations (AREs), instead of normal AREs in standard optimal control problems. To obviate the requirement of knowing the agents’ dynamics, an off-policy reinforcement learning approach implemented on an actor-critic structure is utilized for solving the inhomogeneous ARE. A simulation example is conducted to illustrate the effectiveness of the presented method.

INDEX TERMS containment control, heterogeneous multi-agent systems, non-autonomous leader, fully distributed observer, optimal model reference approach, model-free reinforcement learning.

I. INTRODUCTION

THE cooperative control problem of multi-agent systems (MASs) has drawn significant attention in recent years because of its broad applications in many fields, such as power systems, sensor networks, robotic swarms, unmanned aerial/underwater/ground vehicles etc [1]–[6]. In MASs, the interaction among agents is usually captured by a communication graph [7], [8]. Based on the type of communication topology and the control objective, different distributed cooperative control problems are considered in the literature, such as consensus control [9], flocking control [10], formation control [11], and containment control [12], which is the focus...
of this paper.

In the containment control problem, the MAS consists of two tuples of agents: leaders and followers. The leaders’ movements are independent of other agents in the network. The objective for the followers is to design a distributed control protocol based on their local information to fall into the convex hull spanned by multiple leaders [12]–[15]. As an application, for a group of unmanned aerial vehicles that navigates in an environment, the safety region can be represented by the convex hull spanned by all leaders, i.e., agents that are empowered with more capable sensing devices or leverage being closer to threats or obstacles. Most existing works in the containment control problem consider the leaders to be static [12] or dynamic, but autonomous [13], [16]. Autonomous leaders are assumed to generate their trajectories using an exosystem without a control input, and they are not capable of reacting to environmental or mission changes as there is no input to change their behavior. In many practical situations, the control command signal is required to change the leaders’ trajectories for seeking advantages and avoiding disadvantages in order to generate a safety region for all the followers. Therefore, containment control of MASs with dynamic and non-autonomous leaders is desired.

The above-mentioned results are presented for homogeneous MASs, where all agents have the same dynamics. In contrast, heterogeneous MASs have a broader range of applications where agents are allowed to have non-identical dynamics in order to achieve different tasks. Containment control of heterogeneous MASs with single- and second-order agent dynamics was considered in [17]. Output regulation theory is applied for heterogeneous linear MASs [18], [19] and heterogeneous nonlinear dynamics [20]. However, these works focused on containment control of heterogeneous MASs with autonomous leaders [17]–[20]. Containment control of MASs with non-autonomous leaders were considered in [21], but the result is still limited to homogeneous systems. Moreover, the global information about the communication topology of MASs is supposed to be known in most results (e.g. the minimum eigenvalue of the Laplacian matrix [2], [3], [8]). To obviate this requirement, a fully distributed protocol is developed for leaderless and leader-follower MASs [22]–[24]. To our knowledge, however, no fully distributed control solution exists for containment problem.

Existing solutions to the containment control problem require complete knowledge of the followers’ dynamics as well as the leaders’ dynamics, which might not be available in many applications. Model reference adaptive control (MRAC) method has been adopted to obtain adaptive cooperative control protocols in [25]–[29]. However, for the case of the containment control problem, the interior point inside the convex hull spanned by the leaders is not unique. Therefore, MRAC can not be directly used to handle the containment control problem. Moreover, adaptive solutions, however, are generally far from optimal. Optimality of the solution is desired to achieve a good transient response, in addition to stability. To our knowledge, optimality has also been ignored in existing solutions to the containment control problems. Reinforcement learning (RL) has been successfully used to design adaptive optimal solutions for both single-agent and MASs. However, to our knowledge there is no adaptive optimal solution to the containment control of MASs with non-autonomous leaders. Existence of non-autonomous leaders makes learning an optimal solution more challenging and extension of the existing approaches is not straightforward. This is because, as shown in this paper, solving optimal containment control problem with non-autonomous leaders results in solving inhomogeneous AREs for which there is no solution in the literature. The containment control problem of heterogeneous MASs with non-autonomous leaders is formulated within the MRAC framework in this paper.

This paper presents a fully distributed adaptive optimal solution to the containment control problem of MASs with non-autonomous leaders, as shown in Figure 1. A fully distributed optimal MRAC framework is presented to solve this problem. First, a distributed adaptive observer is designed to estimate the safety region for all the followers, i.e. the convex hull spanned by the non-autonomous leaders. The proposed observer is fully distributed in the sense that it does not need any knowledge of the Laplacian matrix. The observer dynamics for each follower is considered as a reference model to be tracked. This formulation results in a MRAC problem. Then, distributed adaptive optimal protocols are designed for the formulated MRAC problem based on their available local information and the state of the distributed observer. The presented optimal solution to the MRAC control for containment control problem not only stabilizes the local tracking error dynamics but also minimizes the transient performance. It has been shown that solving the optimal MRAC for containment control with non-autonomous leaders leads to solving inhomogeneous AREs. To obviate the requirement of complete knowledge about agent dynamics, an off-policy RL method is combined with an actor-critic structure to solve the inhomogeneous ARE in a model-free manner.

The rest of this paper is organized as follows: Section II
formulates the containment control problem for heterogeneous MASs with non-identical non-autonomous leaders and followers. A fully distributed observer that does not require the global graph topology information is developed in Section III to estimate the safety region and serve as a reference model. In Section IV, the containment control problem is formulated as a MRAC problem and converted to an optimal synchronization problem. Inhomogeneous AREs are derived to find the solution to the optimal model reference containment control. In Section V, the fully distributed observer designed in Section III and the optimal model reference containment control developed in Section IV are integrated. In Section VI, an off-policy RL approach is combined with an actor-critic algorithm to solve the inhomogeneous AREs in a model-free manner. Numerical simulation is carried out in Section VII to validate the effectiveness of the presented approach. Section VIII provides the conclusion for the entire paper.

**NOTATIONS**

Through this paper, the following notations of graph topology to represent the interaction among agents are required for following discussions.

- \( G \) \( \triangleq \) set of all agents (\( G = \{1, \ldots, M + N\} \))
- \( R \) \( \triangleq \) set of leaders (\( R = \{1, \ldots, M\} \))
- \( F \) \( \triangleq \) set of followers (\( F = \{M + 1, \ldots, M + N\} \))
- \( A \) \( \triangleq \) adjacency matrix of subgraph \( F \)
- \( a_{ij} \) \( \triangleq \) entry of adjacency matrix \( A \)
- \( d_i \) \( \triangleq \) in-degree of node \( i \)
- \( D \) \( \triangleq \) in-degree matrix of subgraph \( F \) (\( D = \text{diag} \{d_1, \ldots, d_N\} \))
- \( N_i \) \( \triangleq \) set of neighbors of node \( i \)
- \( L \) \( \triangleq \) Laplacian matrix of subgraph \( F \) (\( L = D - A \))
- \( \delta_k^i \) \( \triangleq \) pinning gain from \( k \)-th leader to \( i \)-th follower
- \( \Delta_k \) \( \triangleq \) pinning matrix of \( k \)-th leader to all followers (\( \Delta_k = \text{diag} \{\delta_k^1, \ldots, \delta_k^{M+N}\} \))
- \( \lambda_i(M) \) \( \triangleq \) \( i \)-th dimension of matrix \( M \)
- \( \mathbb{1}_N \) \( \triangleq \) \( N \)-dimensional column vector with all entries being 1.
- \( \|\cdot\| \) \( \triangleq \) \( L_2 \) norm

**Assumption 1.** The communication graph of the followers set \( F \) is undirected. For each follower, there exists at least one leader that has a directed path to that follower.

**Lemma 1.** [30] (Barbalat’s Lemma) If the differentiable function \( f(t) \) has a finite limit as \( t \to \infty \), and if \( f \) is uniformly continuous, then \( f \to 0 \) as \( t \to \infty \).

**II. PROBLEM STATEMENT**

This section formulates the containment control problem for MASs with non-autonomous leaders. A solution to this problem will be presented in the subsequent sections.

The leaders are assumed to be non-autonomous (i.e., leaders with non-zero control inputs) with unknown control input, and are considered as

\[
\dot{x}_k(t) = Ax_k + Bu_k, \quad k \in R, \quad (1)
\]

where \( x_k \in \mathbb{R}^n \) and \( u_k \in \mathbb{R}^m \) are the state and control input of the \( k \)-th leader, respectively.

**Assumption 2.** For the \( k \)-th leader, the unknown control input \( u_k \) is uniformly continuous and belongs to \( L_2 \) space.

**Assumption 3.** The pair \((A, B)\) is stabilizable, and \( A \) is neutrally stable, i.e., all the eigenvalues of matrix \( A \) lie on the imaginary axis and are distinct from one another.

The followers are described as

\[
\dot{x}_i(t) = A_ix_i + B_iu_i, \quad i \in F, \quad (2)
\]

where \( x_i \in \mathbb{R}^n \) and \( u_i \in \mathbb{R}^m \) are the state and control of the \( i \)-th follower, respectively.

**Assumption 4.** For \( \forall i \in F \), the pair \((A_i, B_i)\) is stabilizable but unknown.

**Remark 1.** Note that the heterogeneous MASs consists of M leaders (1) with identical dynamics \((A, B)\) and N followers with non-identical dynamics \((A_i, B_i)\). The objective of containment control problem is to develop distributed control protocol for each follower to be synchronized into the convex hull spanned by the leaders. The interior points in the convex hull spanned by the leaders share the same dynamics as the leaders have the same dynamics. Otherwise, the convex hull either expands or shrinks as time goes on. Therefore, in containment control of heterogeneous MASs, the leaders with identical dynamics is a standard assumption in existing literature [18], [19], [21], [31].

**Remark 2.** In contrast to existing results consider heterogeneous MASs with autonomous leaders [18], the leaders considered in this paper are formulated by non-autonomous differential equations (1), where the leaders’ trajectories \( x_k \) are influenced by a bounded control input \( u_k \). Therefore, the leaders dynamics (1) can represent a more class of desired trajectories in many applications, compared to autonomous leaders with no control input.

**Remark 3.** In [25] and [26], it is assumed that the leader’s control input \( u_k \) is known and the followers input matrix, i.e., \( B_i \), is available to design distributed cooperative control protocol. Moreover, in [26], it is assumed that there exist \( H_i^* \) and \( K_i^* \) such that \( A = A_i + B_iH_i^* \) and \( B = B_iK_i^* \). In contrast to the MRAC based cooperative control of MASs in [25], [26], this paper relaxes the requirement of availability of \( u_k \) and followers’ input matrix \( B_i \).

**Remark 4.** Based on Lemma 1, it can be inferred from Assumption 2 that there exists a positive constant \( \varphi_k \) such that

\[
\|u_k\| \leq \varphi_k, \quad u_k(t) \to 0. \quad (3)
\]
The compact form of (6) can be expressed as [18], [33],

\[
\xi = Hz - \sum_{k \in \mathcal{R}} H_k \bar{x}_k, \tag{7}
\]

where

\[
h_k = \frac{\mathcal{L}}{M} + \Delta_k, \quad H_k = h_k \otimes I_n, \quad h_k = \sum_{k \in \mathcal{R}} h_k = \mathcal{L} + \Delta,
\]

\[
H = \sum_{k \in \mathcal{R}} H_k = h \otimes I_n, \quad \bar{x}_k = I_N \otimes \bar{x}_k, \quad k \in \mathcal{R},
\]

\[
z = \left[ z_{M+1}^T, z_{M+2}^T, \ldots, z_{M+N}^T \right]^T,
\]

\[
\xi = \left[ \xi_{M+1}, \xi_{M+2}, \ldots, \xi_{M+N} \right]^T \tag{8}
\]

From (1), (3) and (7), the dynamic of \( \xi \) can be determined as

\[
\dot{\xi} = \bar{A} \xi + \left( \bar{h} \hat{D} \otimes BF \right) \xi + \left( \hat{D} \otimes B \right) G - \sum_{k \in \mathcal{R}} H_k \bar{B} \bar{u}_k, \tag{9}
\]

where

\[
\bar{A} = I_N \otimes A, \quad \bar{B} = I_N \otimes B, \quad \bar{u}_k = I_N \otimes u_k,
\]

\[
G = \left[ g^T(F \xi_{M+1}) \ldots g^T(F \xi_{M+N}) \right]^T
\]

\[
\hat{D} = \text{diag} \left( \left[ \hat{d}_{M+1} \ldots \hat{d}_{M+N} \right] \right), \tag{10}
\]

Therefore, the distributed observer design problem can be formulated as follows.

Problem 2. (Distributed Observer Design Problem) Design \( v_i \) in (4) to achieve \( \lim_{t \to \infty} \xi(t) = 0 \), which guarantees the convergence of the distributed observer state \( z_i \) in (3) to the convex hull spanned by the M leaders.

Lemma 2. [18] Suppose Assumption 1 holds. Then the following holds:

- \( H_k \) and \( h_k \) are positive definite;
- \( H \) and \( h \) are positive definite.

It is shown in [18] that \( \xi = 0 \) if and only if Problem 2 is solved. Therefore, \( \xi \) in (7) is referred to as a containment error and is to be regulated to zero. In addition, according to Lemma 2, \( H \) is nonsingular since it is positive definite. Let \( \varepsilon = H^{-1} \xi \), then, \( \xi = 0 \) if and only if \( \varepsilon = 0 \). In the following, a distributed adaptive observer is designed to drive \( \varepsilon \) approach to the origin asymptotically.

The parameters \( F, T \) of a fully distributed adaptive observer in (4) can be determined by the next theorem.

Theorem 1. (Distributed Adaptive Observer Design) Consider the distributed observer (3) with the control (4). If the design parameters satisfy

\[
F = -B^T \mathcal{P}^{-1}, \tag{11}
\]

\[
T = \mathcal{P}^{-1} B B^T \mathcal{P}^{-1}, \tag{12}
\]

where \( \mathcal{P} \succ 0 \) is a solution to the following LMI

\[
A \mathcal{P} + \mathcal{P} A^T - 2B B^T \mathcal{P} \prec 0. \tag{13}
\]

Then, the containment error \( \varepsilon \) converges to the origin asymptotically.
Proof: First, based on (7), we define the variable transformation

\[
\varepsilon = (h^{-1} \otimes I_n) \xi = H^{-1} \xi
\]

\[
= H^{-1} \left( H z - \sum_{k \in \mathcal{K}} H_k \tilde{x}_k \right) = z - \sum_{k \in \mathcal{K}} H^{-1} H_k \xi_k
\]

(14)

with

\[
f_k := H^{-1} H_k = (h^{-1} h_k) \otimes I_n.
\]

Multiplying \( H^{-1} = h^{-1} \otimes I_n \) on both sides of (9) yields

\[
(h^{-1} \otimes I_n) \dot{\xi} = \left( (I_N \otimes A) \xi + (h^{-1} \otimes I_n) \left( h \dot{\mathcal{D}} \otimes BF \right) \xi + (h^{-1} \otimes I_n) \left( \dot{\mathcal{D}} \otimes B \right) G - (h^{-1} \otimes I_n) \sum_{k \in \mathcal{K}} H_k \tilde{B} \xi_k \right),
\]

(15)

From (14), the left hand side of (15) satisfies

\[
(h^{-1} \otimes I_n) \dot{\xi} = \dot{\varepsilon}.
\]

(16)

Using the properties of Kronecker matrix product [34], the terms in the right hand side of (15) can be further rewritten as

\[
(h^{-1} \otimes I_n) \left( I_N \otimes A \right) \xi = (h^{-1} \otimes A) \xi
\]

\[
= (I_N \otimes A) \left( h^{-1} \otimes I_n \right) \xi = \bar{A} \varepsilon,
\]

(17)

\[
(h^{-1} \otimes I_n) \left( h \dot{\mathcal{D}} \otimes BF \right) \xi = (h^{-1} h \dot{\mathcal{D}} \otimes BF) \xi
\]

\[
= (\dot{\mathcal{D}} h^{-1} \otimes BF) \xi = (\dot{\mathcal{D}} h \otimes BF) \left( h^{-1} \otimes I_n \right) \xi
\]

\[
= (\dot{\mathcal{D}} h \otimes BF) \varepsilon,
\]

(18)

\[
(h^{-1} \otimes I_n) \left( \dot{\mathcal{D}} \otimes B \right) G = (\dot{\mathcal{D}} h^{-1} \otimes B) G,
\]

(19)

\[
- \left( h^{-1} \otimes I_n \right) \sum_{k \in \mathcal{K}} H_k \tilde{B} \xi_k
\]

\[
= - \left( h^{-1} \otimes I_n \right) \sum_{k \in \mathcal{K}} (h_k \otimes I_n) \tilde{B} \xi_k
\]

\[
= - \sum_{k \in \mathcal{K}} f_k \tilde{B} \xi_k.
\]

(20)

In (19), we use the fact \( \dot{\mathcal{D}} h^{-1} = h^{-1} \dot{\mathcal{D}} \), which results from that \( \dot{\mathcal{D}} \) is a diagonal matrix. Taking (16) – (20) back into (15) yields

\[
\dot{\varepsilon} = \left[ \bar{A} + (\dot{\mathcal{D}} h \otimes BF) \right] \varepsilon + (\dot{\mathcal{D}} h^{-1} \otimes B) G - \sum_{k \in \mathcal{K}} f_k \tilde{B} \xi_k.
\]

(21)

Based on Lemma 2, the matrix \( H \) is nonsingular. Therefore, \( \xi = 0 \) if and only if \( \varepsilon = 0 \). Next, we analyze the stability of system (21). Consider the following Lyapunov candidate

\[
V = \frac{\rho}{2} \varepsilon^T \left( h \otimes P^{-1} \right) \varepsilon + \sum_{i \in \mathcal{F}, j \in \mathcal{J}} \frac{1}{2 \tau_i} \left( \hat{d}_i - \beta \right)^2,
\]

where \( \beta \) is a constant satisfies

\[
\beta > \max \left\{ \frac{\rho}{\lambda}, \varphi \right\}
\]

(23)

\[
\varphi = \max_{k \in \mathcal{F}} \nu_k, \quad \lambda = \min_{i \in \mathcal{I}} \lambda_i(h)
\]

\[
\rho > \max \left\{ \lambda, 1 \right\}
\]

(24)

\[
\rho \in \mathbb{R}.
\]

where \( \lambda = \max_{i \in \mathcal{I}} \lambda_i(h) \). According to Assumption 1 and Lemma 2, it is known that \( V \) is positive semidefinite. Along the evolution of leaders in (1), distributed observer in (3) with controller (4), and \( \varepsilon \) in (15), the time derivative of \( V \) can be expressed as in (25).

Considering \( \dot{F} = -B^T P^{-1} \) and \( T = P^{-1} B B^T P^{-1} \), then \( V_2 \) in (25) can be equivalently rewritten as

\[
V_2 = \rho \varepsilon^T (h \dot{\mathcal{D}} \otimes P^{-1} BF) \varepsilon
\]

\[
= \rho \varepsilon^T \left( h \otimes I_n \right) \left( \dot{\mathcal{D}} \otimes P^{-1} BF \right) \left( h \otimes I_n \right) \varepsilon
\]

\[
= \rho \varepsilon^T \left( \dot{\mathcal{D}} \otimes P^{-1} BF \right) \varepsilon
\]

\[
= -\rho \sum_{k \in \mathcal{F}} \hat{d}_k \xi_k^T P^{-1} BB^T P^{-1} \xi_k
\]

\[
= -\rho V_5.
\]

(26)

Hence,

\[
V_2 + V_5 = (1 - \rho) V_5 \leq 0
\]

(27)

can be guaranteed if \( \rho > 1 \).

Similarly, \( V_3 \) can be rewritten as

\[
V_3 = \rho \varepsilon^T (h^{-1} \dot{\mathcal{D}} \otimes P^{-1} B) G
\]

\[
= \rho \varepsilon^T \left( h^{-1} \dot{\mathcal{D}} \otimes P^{-1} B \right) G
\]

\[
\leq -\frac{\rho}{\lambda} \sum_{k \in \mathcal{F}} \hat{d}_k \left\| B P^{-1} \xi_k \right\| = -V_7.
\]

(28)

Then,

\[
V_3 + V_7 = \left( 1 - \frac{\rho}{\lambda} \right) V_7 \leq 0
\]

(29)

as long as \( \rho > \lambda \).

Also, note that

\[
V_1 + V_6 = \rho \varepsilon^T (h \otimes P^{-1} A) \varepsilon - \beta \varepsilon^T \left( h^2 \otimes P^{-1} B B^T P^{-1} \right) \varepsilon
\]

\[
= \frac{1}{2} \varepsilon^T \chi \varepsilon,
\]

(30)

where

\[
\chi = \rho \left[ h \otimes \left( P^{-1} A + A^T P^{-1} \right) \right] - 2\beta \left( h^2 \otimes P^{-1} B B^T P^{-1} \right).
\]

(31)

On the other hand, one has

\[
\varepsilon^T \left[ h \otimes I_n \right] = \left[ \sum_{y \in \mathcal{F}} h_{1y} \varepsilon_{1y}^T \cdots \sum_{y \in \mathcal{F}} h_{Ny} \varepsilon_{Ny}^T \right],
\]

(32)
\[
\dot{V} = \rho \varepsilon^T \left( h \otimes P^{-1} A \right) \varepsilon + \rho \varepsilon^T \left( h \dot{D} h \otimes P^{-1} B F \right) \varepsilon + \rho \varepsilon^T \left( \dot{D} \otimes P^{-1} B \right) G - \rho \varepsilon^T \left( h \otimes P^{-1} \right) \sum_{k \in R} f_k \bar{B} u_k \\
+ \sum_{k \in R} \tilde{d}_k \xi^T_k \xi_k - \sum_{k \in R} \beta \xi^T_k \xi_k + \sum_{k \in R} \tilde{d}_k \| F \| \xi_k \| - \sum_{k \in R} \beta \| F \| \xi_k \|
\]

\[
V_4 = -\rho \left[ \sum_{j \in F} h_{1j} \varepsilon_j \cdots \sum_{j \in F} h_{Nj} \varepsilon_j^T \right] \left[ \begin{array}{c} \sum_{k \in R} f_{k1} P^{-1} B u_k \\ \vdots \\ \sum_{k \in R} f_{kN} P^{-1} B u_k \end{array} \right] = -\rho \sum_{i \in F} \sum_{j \in F} h_{ij} \varepsilon_j^T P^{-1} B \sum_{k \in R} f_{kj} u_k \\
\leq \rho \sum_{i \in F} \left\| B^T P^{-1} \sum_{j \in F} h_{ij} \varepsilon_j \right\| \sum_{k \in R} \| f_{kj} \| u_k \| \leq \rho \sum_{i \in F} \left\| B^T P^{-1} \sum_{j \in F} h_{ij} \varepsilon_j \right\| \max_{k \in R} \varphi_k = \rho \bar{\varphi} \sum_{i \in F} \left\| B^T P^{-1} \sum_{j \in F} \xi_j \right\| \quad (33)
\]

where \( h_{ij} \) denotes the entry of matrix \( h \). Then (33) holds.

Based on (33), one can obtain

\[
V_4 + V_8 \leq (\rho \bar{\varphi} - \beta) \sum_{i \in F} \left\| B^T P^{-1} \sum_{j \in F} \xi_j \right\|. \quad (34)
\]

Collecting (27), (29), (30) and (34), then the time derivative of Lyapunov function \( V \) in (22) satisfies

\[
\dot{V} \leq \frac{1}{2} \varepsilon^T \chi \varepsilon + (\rho \bar{\varphi} - \beta) \sum_{i \in F} \left\| B^T P^{-1} \sum_{j \in F} \xi_j \right\|. \quad (35)
\]

By selecting \( \beta \) sufficiently large such that \( \beta \geq \rho \bar{\varphi} \), then (35) leads to

\[
\dot{V} \leq \frac{1}{2} \varepsilon^T \chi \varepsilon. \quad (36)
\]

By considering the LMI in (13) and \( \chi \) defined in (31), multiplying the positive definite matrix \( h^{-\frac{1}{2}} \otimes P \) to both sides of \( \frac{1}{2} \chi \), one can obtain

\[
\left( h^{-\frac{1}{2}} \otimes P \right) \frac{1}{2} \chi \left( h^{-\frac{1}{2}} \otimes P \right) = \frac{\rho}{2} I_N \otimes \left( AP + PA^T \right) - \beta h \otimes BB^T \\
\leq \frac{\rho}{2} I_N \otimes \left( AP + PA^T - \frac{2\beta \lambda_{\text{min}}(h)}{\rho} BB^T \right) \quad (37)
\]

If \( \beta \) is also large enough such that \( \beta > \frac{\rho}{\lambda_{\text{min}}(h)} \) for \( \forall i \), then, combined with the LMI (13) yields

\[
\left( h^{-\frac{1}{2}} \otimes P \right) \frac{1}{2} \chi \left( h^{-\frac{1}{2}} \otimes P \right) < 0 \quad (38)
\]

Furthermore, considering that \( \left( h^{-\frac{1}{2}} \otimes P \right) \) in (37) is positive definite. Then, one can obtain

\[
\chi < 0, \quad (39)
\]

and

\[
\dot{V} \leq 0 \quad (40)
\]

can be guaranteed.

Based on the fact that \( V \) in (22) is nonnegative and \( \dot{V} \leq 0 \), then the boundedness of \( \varepsilon \) and \( \dot{d}_i \) is guaranteed. Following Assumption 2, then \( \dot{\varepsilon} \) in (21) and \( \dot{\xi} \) in (9) are also bounded. Since \( V \) is non-increasing (see (40)) and bounded from below by 0 (see (22)), the Lyapunov function \( V \) has a finite limit as \( t \to \infty \), denoted by \( V(\infty) \). By integrating both sides of (36) over \([0, \infty)\) one can obtain

\[
- \int_0^\infty \frac{1}{2} \varepsilon^T \chi \varepsilon d\tau \leq V(0) - V(\infty). \quad (41)
\]

Therefore, \( \int_0^\infty \frac{1}{2} \varepsilon^T \chi \varepsilon d\tau \) exists and is finite. As mentioned above, both \( \varepsilon \) and \( \dot{\varepsilon} \) are bounded, so \( \varepsilon^T \chi \varepsilon \) is also bounded. That is, both \( V \) and \( \dot{V} \) are bounded, which guarantees that \( V \) is uniformly continuous. Based on Lemma 1, the following can be guaranteed

\[
V \to 0, \ t \to \infty, \quad (42)
\]

or equivalently,

\[
\varepsilon \to 0, \ t \to \infty. \quad (43)
\]

Note that \( T \geq 0 \) and \( \tau_i > 0 \) in (4), thus \( d_i \) increases monotonically and converges to some finite constant. \( \blacksquare \)

**Remark 5.** The conventional distributed cooperative control takes the form of [2], [3]

\[
v_i = c F \xi_i, \quad (44)
\]

where \( c \) is the coupling gain and is lower bounded by some function of the eigenvalues of Laplacian \( L \) [8], [23], [35]. Therefore, the distributed cooperative control in (44) is based on the information about eigenvalues of Laplacian \( L \), which is global information of the overall communication topology.
In contrast, the adaptive observer (4) designed by Theorem 1, can be regarded as an alternative of the coupling gain c in (44). In this way, the requirement for knowledge about the Laplacian eigenvalues can be relaxed.

Remark 6. For leader-follower consensus control of MASs with a single leader, the leader represents the reference model for all the followers [25]–[27], [29]. In contrast, in the case of containment control of MASs with multiple leaders, the interior point inside the convex hull spanned by the leaders is not unique, and each leader is only connected to its neighbors. Therefore, MRAC cannot be directly used to solve containment control of MASs with multiple leaders. The distributed adaptive observer (3) with (4) designed by Theorem 1 can be viewed as a novel reference model for all the followers. In this way, the containment control problem can be formulated as a MRAC problem, as described in the next section.

As shown in (3), the leaders’ dynamics, \((A, B)\) in (1), are required for the distributed adaptive observer design. However, in the following, we will develop distributed containment control protocol for each follower without the requirement of followers’ dynamics, the pair \((A_i, B_i)\) in (2).

IV. DISTRIBUTED OPTIMAL MODEL REFERENCE CONTAINMENT CONTROL WITH NON-AUTONOMOUS LEADERS

Theorem 1 provides the estimation of the safety region, i.e., an interior point in the convex hull spanned by the leaders. Therefore, Problem 1 is solved only if each follower state \(x_i\) is synchronized to the distributed observer state \(z_i\) in (3).

Based on the distributed adaptive observer, the containment control problem can be formulated as a distributed MRAC problem with distributed observer (3) as reference model. Moreover, in order to impose optimality to the distributed containment control protocol, inhomogeneous AREs are derived to obtain the distributed optimal model reference control in this section.

A. DISTRIBUTED OPTIMAL MODEL REFERENCE CONTAINMENT CONTROL PROBLEM

Note that the adaptive distributed observer design in (4) along with Theorem 1 generate the estimation of the desired trajectory, i.e., the convex hull spanned by the leaders for each follower. By considering these results as a reference model, Problem 1 can be equivalently formulated as the MRAC problem such that the distributed control law \(u_i\) is designed for follower \(i, i \in F\) to assure each follower’s state is synchronized to the distributed adaptive observer’s state, i.e., \(x_i(t) \rightarrow z_i(t)\), as \(t \rightarrow \infty\).

In the MRAC formulation of cooperative control of MASs in [25]–[27], [29], only stability and/or boundedness of the containment error, \(\xi\) in (7) or \(\epsilon\) in (14), is taken into account. Contrary to these results, in this paper, optimality is explicitly imposed to the transient performance of the containment error. Therefore, we present the following distributed optimal control problem to solve Problem 1 in an optimal fashion.

Problem 3. (Distributed Optimal Containment Control Problem) For \(\forall i \in F\), design the distributed control protocol \(u_i(t)\) in (2) for follower \(i\) so that its state tracks the distributed observer output \(z_i(t)\) in (3) by minimizing the following discounted performance function

\[
V (x_i(t), z_i(t)) = \frac{1}{2} \int_t^\infty e^{-\gamma_i (\tau-t)} \left[ x_i(\tau) - z_i(\tau) \right]^T Q_i \left[ x_i(\tau) - z_i(\tau) \right] + u_i(\tau)^T R_i u_i(\tau) \, d\tau,
\]

where \(\gamma_i > 0\) is the discount factor, and \(Q_i\) is a symmetric positive definite matrix \(R_i\) is a positive definite matrix with positive elements on the diagonal, i.e., \(R_i = \text{diag}\left( r_1^2, \ldots, r_m^2 \right) \), \(r_j^2 > 0, j = 1, \ldots, m\).

B. INHOMOGENEOUS BELLMAN EQUATION

An augmented state composed of the follower state \(x_i\) in (2) and observer state \(z_i\) in (3) can be represented as

\[
\zeta_i(t) = \begin{bmatrix} z_i^T(t) & z_i^T(t) \end{bmatrix}^T, \quad i \in F,
\]

where \(x_i\) and \(z_i\) are the \(i\)-th distributed state of follower and observer in (2) and (3), respectively. Then, the augmented system dynamics is given as

\[
\dot{\zeta}_i = \begin{bmatrix} A_i & 0 \\ 0 & A \end{bmatrix} \zeta_i + \begin{bmatrix} B_i \\ 0 \end{bmatrix} u_i + \begin{bmatrix} 0 \\ B \end{bmatrix} v_i = A_i \zeta_i + B_i u_i + D_i v_i.
\]

The performance function in (45) can be equivalently rewritten as

\[
V (\zeta_i(t)) = \int_t^\infty e^{-\gamma_i (\tau-t)} r_i (\zeta_i(\tau), u_i(\tau)) \, d\tau,
\]

where \(r_i (\zeta_i, u_i) = \frac{1}{2} (\zeta_i^T C_i^T Q_i C_i \zeta_i + u_i^T R_i u_i) + C_i = \begin{bmatrix} I & -I \end{bmatrix}\). Therefore, Problem 3 can be viewed as optimal regulation of augmented system (47) with respect to the performance (48).

Remark 7. It is necessary to stress the difference between the classical optimal tracking problem and the optimal regulation of the augmented system (47), with respect to the performance (48). The classical tracking control problem with autonomous reference can be reformulated as a regulation problem of augmented systems by incorporating the reference dynamics with plant dynamics [36]–[38]. The augmented dynamics only contain one inhomogeneous item which is the plant dynamics. However, in the optimal regulation of system (47), the augmented system is obtained by incorporating an inhomogeneous observer (3) with the agent dynamics (2). The observer control \(v_i\) also exists in the augmented system (47), which makes the augmented system different from classical cases. This difference results from the non-autonomous leaders, because the observer control input is required to make the observer \(z_i\) track the convex hull spanned by the non-autonomous leader. In other words, \(v_i\) in the observer...
helps to compensate the bounded and unknown control $u_i$ in the non-autonomous dynamics of the leaders. □

**Lemma 3.** [39] Consider the infinite horizon performance function in Problem 3. Suppose that the following admissible control law is applied to follower $i$ with dynamics given in (2)

$$u_i = k^1_i x_i + k^2_i z_i + c_i = K_i \zeta_i + c_i \quad i \in \mathcal{F},$$

(49)

where $k^1_i$ and $k^2_i$ are the state feedback and feed-forward gains and $c_i$ is the inhomogeneous item. Then, the value function (45) can be written as the following quadratic polynomial form

$$V_i(\zeta_i(t)) = \frac{1}{2} \left[ \zeta_i^T(t) P_i \zeta_i(t) + 2 \zeta_i^T(t) M_i(t) + N_i(t) \right]$$

(50)

Differentiating the value function (50) along the trajectory of the augmented system (47) and control input (49) yields,

$$0 = 2 \left( \left[ \frac{\partial V(\zeta)}{\partial \zeta_i}, \dot{\zeta}_i \right] + r(\zeta_i, u_i) - \gamma_i V(\zeta) \right)$$

$$= (A_i \zeta_i + B_i u_i + \bar{D}_i v_i)^T (P_i \zeta_i + M_i)$$

$$+ (P_i \zeta_i + M_i)^T (A_i \zeta_i + B_i u_i + \bar{D}_i v_i)$$

$$- \gamma_i (\zeta_i^T P_i \zeta_i + 2 \zeta_i^T M_i + N_i)$$

$$+ \zeta_i^T C_i Q_i C_i \zeta_i + u_i^T R_i u_i.$$ (51)

In the following, (51) is referred to as an inhomogeneous Bellman equation in contrast to the homogeneous Bellman equation in classical optimal control theory [40].

**C. INHOMOGENEOUS ARE AND OPTIMALITY DISCUSSION**

The inhomogeneous Bellman equation described the relationship between the inhomogeneous feedback control policy (49) and its corresponding value function (50). For the optimal case, the inhomogeneous Bellman equation will become an inhomogeneous ARE, in contrast to the homogeneous ARE in the classical optimal control theory.

The following theorem gives the formulation of an inhomogeneous ARE and the corresponding optimality discussions.

**Theorem 2.** (Inhomogeneous ARE) The optimal solution for the discounted infinite horizon optimal tracking problem is

$$u^*_i = -R_i^{-1} \bar{B}^T_i (P^*_i \zeta_i + M^*_i), \quad i \in \mathcal{F},$$

(52)

i.e., \[ \begin{bmatrix} k^{1*} & k^{2*} \end{bmatrix} = -R_i^{-1} \bar{B}^T_i P^*_i \] and $c^*_i = -R_i^{-1} \bar{B}^T_i M^*_i,$ where $P^*_i,$ $M^*_i$ and $N^*_i$ satisfies the inhomogeneous ARE

$$0 = \bar{A}_i^T P^*_i + P^*_i \bar{A}_i + \gamma_i P^*_i - P^*_i \bar{B}_i R_i^{-1} \bar{B}^T_i P^*_i + C_i^T Q_i C_i,$$

(53)

$$\bar{M}_i^* = - (\bar{A}_i^T - P^*_i \bar{B}_i R_i^{-1} \bar{B}^T_i) M_i^* - P^*_i \bar{B}_i R_i^{-1} \bar{B}^T_i P^*_i + C_i^T Q_i C_i,$$

$$\bar{N}_i^* = -2(M_i^*)^T \bar{D}_i v_i + (M_i^*)^T \bar{B}_i R_i^{-1} \bar{B}^T_i M_i^* + \gamma_i N_i^*,$$

(54)

$$\bar{M}_i^* (\infty) = N_i^* (\infty) = 0.$$ (55)

with $M^*_i (\infty) = N^*_i (\infty) = 0$. □

**Proof:** Define the Hamiltonian as

$$H(\zeta_i, u_i, V(\zeta_i)) = \frac{dv(\zeta_i)}{dt} - \gamma_i V(\zeta_i) + \zeta_i^T C_i^T Q_i C_i \zeta_i + u_i^T R_i u_i.$$ (56)

The optimal control is obtained by [40]

$$u^*_i = \arg\min_{u_i} H(\zeta_i, u_i; V(\zeta_i)), \forall \zeta_i \in \Omega.$$ (57)

By employing the stationarity condition $\partial H / \partial u_i = 0$ and considering the augmented system dynamics (47) and the value function (50), one has

$$\frac{\partial H}{\partial u_i} = 2 \bar{B}^T_i (P_i^* \zeta_i + M_i^*) + 2 R_i u_i^* = 0.$$ (58)

Substituting the value function (50) and the optimal control (58) into the Hamiltonian (56) yields

$$0 = \zeta_i^T \left( \bar{A}_i^T P_i^* + P_i^* \bar{A}_i - P_i^* \bar{B}_i R_i^{-1} \bar{B}^T_i P_i^* \right) \zeta_i$$

$$+ 2 \zeta_i^T \left( M_i^* + \bar{A}_i^T M_i^* - P_i^* \bar{B}_i R_i^{-1} \bar{B}^T_i M_i^* \right)$$

$$+ P_i^* \bar{D}_i v_i - \gamma_i M_i^*$$

$$+ \bar{N}_i^* - (M_i^*)^T \bar{B}_i R_i^{-1} \bar{B}^T_i M_i^* - \gamma_i N_i^* + 2(M_i^*)^T \bar{D}_i v_i.$$ (59)

Note that (59) holds for $\forall \zeta_i \in \Omega$. Therefore, (53), (54) and (55) hold. This completes the proof.

In this paper, (59) is referred to as an inhomogeneous ARE and it is used to obtain the distributed optimal containment control protocol (52).

**Remark 8.** Note that Theorem 2 gives the necessary condition of optimal distributed containment control protocols that can solve Problem 3. The inhomogeneous ARE (59) is broken down into the conventional homogeneous ARE (53), together with inhomogeneous items $M_i^*$ in (54) and $N_i^*$ in (55). From (55), one can know that $N_i^*$ depends on $M_i^*$. Furthermore, from (54), one can know that $M_i^*$ depends on $P_i^*$. Therefore, solving the inhomogeneous ARE (51) depends on solving the augmented ARE (53). □

**D. STABILITY ANALYSIS OF DISTRIBUTED OPTIMAL MODEL REFERENCE CONTAINMENT CONTROL BY SOLVING INHOMOGENEOUS ARE**

Theorem 2 only gives the optimality discussions of the distributed containment control (52) obtained by solving the inhomogeneous ARE. In this section, the stability of the observer-tracking error $\phi_i = x_i - z_i, \forall i \in \mathcal{F}$ is analyzed as following.

**Assumption 5.** The discount factor $\gamma_i > 0$ is selected such that $A - \frac{\gamma_i}{2} I$ is stable and satisfies $\gamma_i \leq 2 \left\| (B_i R_i^{-1} B_i^T Q_i)^{1/2} \right\|.$ □

**Theorem 3.** Suppose the optimal control law (52) is applied to the augmented system (47), and Assumptions 4, 5 are satisfied. Then, the following conclusions hold:

a) For $\forall i \in \mathcal{F},$ the matrix $A_i + B_i k^{1*}$ is Hurwitz.
b) The homogeneous ARE (53) has a unique positive semidefinite solution.

c) The error $\phi_i = x_i - z_i$ goes to zero asymptotically. □

Proof: a) For the proof of part a), refer to [41].
b) For the proof of part b), readers are referred to [36].
c) Multiplying both sides of homogeneous ARE (53) by $\zeta_i^T$ on the left and $\zeta_i$ on the right gives

$$2\zeta_i^T A_i^T P_i^* \zeta_i - \zeta_i^T C_i^T Q_i C_i \zeta_i - (P_i^* \zeta_i)^T B_i R_i^{-1} B_i^T P_i^* \zeta_i = 0.$$  

(60)

It can be seen from (60) that $P_i^* \zeta_i = 0$ guarantees $C_i \zeta_i = 0$, which indicates $\phi_i = x_i - z_i = 0$. Therefore, the null space of $P_i^*$ is a subspace where the observer-tracking error is zero. Select the optimal value function $V(\zeta_i)$ (50) as the Lyapunov function. Since $V(\zeta_i)$ is the integral of the utility function $r(\zeta_i, u_i) = \frac{1}{2}(\zeta_i^T C_i^T Q_i C_i \zeta_i + u_i^T R_i u_i)$ where $Q_i$ and $R_i$ are positive definite, then $V(\zeta_i)$ is a positive definite function, i.e.,

$$V(\zeta_i) = \frac{1}{2} \big( \zeta_i^T P_i^* \zeta_i + 2M_i^T \zeta_i + Ni \big) > 0.$$  

(61)

From (54), the following is true

$$(\gamma_i I - \bar{A}_i + P_i^* B_i R_i^{-1} B_i^T) M_i^* = P_i^* D_i v_i.$$  

(62)

As $t \to \infty$, $v_i \to 0$. Thus, $M_i^*(\infty) = 0$. Similarly, one can have $N_i(\infty) = 0$. Then as $t \to \infty$, $V(\zeta_i) \to \frac{1}{2} \zeta_i^T P_i^* \zeta_i$. Also, as $t \to \infty$, $\bar{V}(\zeta_i)$ approaches to

$$\bar{V}(\zeta_i) \to \frac{1}{2} \big[ \gamma_i \zeta_i^T P_i^* \zeta_i - \zeta_i^T C_i^T Q_i C_i \zeta_i - (P_i^* \zeta_i)^T B_i R_i^{-1} B_i^T P_i^* \zeta_i \big],$$

(63)

From the ARE (53), by rearranging some items, one can obtain

$$\gamma_i P_i^* - C_i^T Q_i C_i - P_i^* B_i R_i^{-1} B_i^T P_i^* = \bar{A}_i P_i^* + P_i^* \bar{A}_i - 2P_i^* B_i R_i^{-1} B_i^T P_i^*$$

$$= \tilde{A}_i P_i^* + P_i^* \tilde{A}_i,$$  

(64)

where $\tilde{A}_i = A_i - B_i R_i^{-1} B_i^T P_i^*$. Note $K_i^* = [k_i^1, k_i^2]^T$, therefore

$$\tilde{A}_i = \begin{bmatrix} A_i + B_i k_i^1 & B_i k_i^2 \\ 0 & A \end{bmatrix}.$$  

(65)

It is shown that $A_i + B_i k_i^1$ is Hurwitz and $A$ has all eigenvalues on the imaginary axis, therefore $\tilde{A}_i$ is marginally stable and there exists a matrix $W \geq 0$ such that

$$\bar{V}(\zeta_i) \to \frac{1}{2} \tilde{A}_i^T P_i^* + P_i^* \tilde{A}_i) \zeta_i = -\frac{1}{2} \tilde{A}_i^T W \zeta_i \leq 0.$$  

(66)

Based on LaSalle’s invariance principle, $\zeta_i$ converges to the largest invariance subspace where $\bar{V}(\zeta_i) = 0$. Based on (66), $\bar{V}(\zeta_i) = 0$ if $P_i \zeta_i = 0$. This completes the proof.

V. FULLY DISTRIBUTED OPTIMAL MODEL REFERENCE ADAPTIVE CONTAINMENT CONTROL

The adaptive distributed observer design in Section III can estimate the convex hull spanned by non-autonomous leaders asymptotically without requiring any global graph information. The distributed containment control design for each follower in Section IV considers not only the steady observer-tracking error asymptotic stability but also the optimality with respect to a performance function. To combine the advantages of the adaptive distributed observer in Section III and the inhomogeneous optimal tracking control design in Section IV, the separation principle is employed to solve Problem 1.

The following theorem shows that the adaptive distributed observer design in Theorem 1 and the distributed optimal inhomogeneous containment control protocol (52) will tackle the containment control of multi-agent heterogeneous systems with non-autonomous leaders.

Theorem 4. Consider the distributed observer (3) designed in Theorem 1 and the optimal tracking controller designed in Theorem 2. Then, the containment control problem is solved, provided that observer parameters $F$, $T$ and $P$ satisfy (11)-(13), and the discount factor $\gamma_i$ satisfies Assumption 5. □

Proof: Considering the distributed observer design in Theorem 1 and the optimal tracking controller design in Theorem 2, the augmented system (47) can be equivalently rewritten as

$$\begin{bmatrix} \dot{x}_i \\ \dot{z}_i \end{bmatrix} = \begin{bmatrix} A_i - B_i R_i^{-1} B_i^T P_{i1} & -B_i R_i^{-1} B_i^T P_{i2} \\ 0 & A \end{bmatrix} \begin{bmatrix} x_i \\ z_i \end{bmatrix}$$

$$+ \begin{bmatrix} -B_i R_i^{-1} B_i^T (M_i^1 + M_i^2) \\ 0 \end{bmatrix} V_i.$$  

(67)

Due to the block-triangular structure, the observers’ dynamics are independent of the agent state $x_i$, and thus, based on the separation principle, the observer and the tracking control can be designed independently of each other.

First, based on Theorem 1 that if $F$, $T$ and $P$ satisfy (11)-(13), then the containment error $\xi$ in (7) asymptotically converges to zero, which guarantees the convergence of the observer state $z_i$ to the convex hull spanned by the leaders, i.e., $z_i \to \text{Co} \left\{ \{x_k\}_{k \in \mathbb{R}} \right\}$.

Second, based on Theorem 3, if the discount factor satisfies Assumption 5, then $x_i \to z_i$ is ensured.

Therefore, when all the assumptions in Theorem 4 are satisfied, one can obtain $x_i \to z_i \to \text{Co} \left\{ \{w_k\} \right\}$. This completes the proof.

VI. FULLY DISTRIBUTED ADAPTIVE OPTIMAL CONTAINMENT CONTROL USING THE ACTOR-CRITIC OFF-POLICY RL APPROACH

A. MODEL-BASED ON-Policy RL FOR CONTAINMENT CONTROL OF HETEROGENEOUS MAS WITH NON-AUTONOMOUS LEADERS

In Theorem 4, it is proven that the combination of the distributed adaptive observer in Section III and the containment control protocol of Section IV will tackle the containment control of multi-agent heterogeneous systems with non-autonomous leaders. Therefore, the following theorem shows that the distributed adaptive control protocol (52) can be designed in a fully distributed manner.

Theorem 5. Consider the distributed adaptive observer (3) designed in Theorem 1 and the fully distributed optimal controller designed in Theorem 2. Then, the containment control problem is solved, provided that observer parameters $F$, $T$ and $P$ satisfy (11)-(13), and the discount factor $\gamma_i$ satisfies Assumption 5. □

Proof: Considering the distributed observer design in Theorem 1 and the optimal tracking controller design in Theorem 2, the augmented system (47) can be equivalently rewritten as

$$\begin{bmatrix} \dot{x}_i \\ \dot{z}_i \end{bmatrix} = \begin{bmatrix} A_i - B_i R_i^{-1} B_i^T P_{i1} & -B_i R_i^{-1} B_i^T P_{i2} \\ 0 & A \end{bmatrix} \begin{bmatrix} x_i \\ z_i \end{bmatrix}$$

$$+ \begin{bmatrix} -B_i R_i^{-1} B_i^T (M_i^1 + M_i^2) \\ 0 \end{bmatrix} V_i.$$  

(67)

Due to the block-triangular structure, the observers’ dynamics are independent of the agent state $x_i$, and thus, based on the separation principle, the observer and the tracking control can be designed independently of each other.

First, based on Theorem 1 that if $F$, $T$ and $P$ satisfy (11)-(13), then the containment error $\xi$ in (7) asymptotically converges to zero, which guarantees the convergence of the observer state $z_i$ to the convex hull spanned by the leaders, i.e., $z_i \to \text{Co} \left\{ \{x_k\}_{k \in \mathbb{R}} \right\}$.

Second, based on Theorem 3, if the discount factor satisfies Assumption 5, then $x_i \to z_i$ is ensured.

Therefore, when all the assumptions in Theorem 4 are satisfied, one can obtain $x_i \to z_i \to \text{Co} \left\{ \{w_k\} \right\}$. This completes the proof.
control in Section IV can solve Problem 1. In this section, to obviate the requirement of complete knowledge of agents dynamics, the off-policy RL method is combined with the actor-critic algorithm to solve the inhomogeneous ARE (59) in order to obtain the optimal distributed containment control for each follower.

Policy iteration, an RL method, has been employed to solve the optimal control problem for continuous-time systems [42], [43] and discrete-time systems [44] iteratively. The on-policy RL method to solve the inhomogeneous ARE (59) is listed in Algorithm 1, which consists of policy evaluation and policy improvement steps. Algorithm 1 is an extension of policy iteration based methods [42]–[44] to solving the inhomogeneous AREs (59). The convergence proof of the policy sequence \( \{ u^\kappa_i \} \) generated by Algorithm 1 to the optimal control \( u^*_i \) is similar to [42]–[44]. For the on-policy RL, there are two issues that need to be considered. First, the iterative policy, \( u^\kappa_i \) in the \( \kappa \)-th iteration, has to be applied to the system for policy evaluation to obtain the corresponding value function, \( V^\kappa_i(\cdot) \), which is further used for policy improvement. Therefore, the learning process during which the iterative policy is applied may generate additional uneconomic costs. On the other hand, it is can be seen from (68) and (69) that the complete knowledge of agent’s dynamics is required in both policy evaluation and policy improvement steps. However, in many real-world applications of heterogeneous MASs, it is usually impossible for each agent to obtain the exact system models. Therefore, it is necessary to develop an approach, which does not depend on complete system dynamics knowledge, for finding the solution of the inhomogeneous ARE (59).

Algorithm 1 Model-based On-Policy RL Algorithm

1: Initialize the agents with admissible control policies \( u^0_i(x_i(t)) \) for \( \forall i = 1, \ldots, N \) and set the iteration index to be \( \kappa = 0 \);
2: Policy Evaluation Step: Evaluate policies \( u^\kappa_i(x_i(t)) \) for \( \forall i = 1, \ldots, N \) by solving the homogeneous Bellman equation (68) for \( P^\kappa_i, M^\kappa_i \) and \( N^\kappa_i \) in \( V^\kappa_i(x_i(t)) \)
\[ H(\zeta_i, u^\kappa_i, V^\kappa_i(\zeta_i)) = 0, \quad \forall \zeta_i, i \in \mathcal{F} \] (68)
3: Policy Improvement Step: Update the distributed control \( u^\kappa_i((x_i(t))) \) as
\[ u^{\kappa+1}_i = -R_i^{-1}B_i^T(P^\kappa_i + M^\kappa_i), \quad \kappa = \kappa + 1; \]
4: If the criteria \( \| V^\kappa_i(x_i) - V^{\kappa+1}_i(x_i) \| \leq \varepsilon, \forall x_i \) is satisfied, then stop; otherwise, go to Step 1.

B. MODEL-FREE OFF-POLICY RL FOR CONTAINMENT CONTROL OF HETEROGENEOUS MASS WITH NON-AUTONOMOUS LEADERS

As discussed in Section VI-A, the on-policy algorithm requires the full knowledge of both the leader and the followers’ dynamics. The off-policy algorithm, originally adopted for optimal control of continuous-time systems in [45], is able to solve the homogeneous ARE. Since then, off-policy RL algorithm was also applied to the optimal tracking problem [41], \( H_\infty \) control problem [46], robust stabilization problem [47] for single-agent systems, output synchronization problem [39] and containment problem [33] for multi-agent systems. In this section, the off-policy RL algorithm is combined with actor-critic structure [48] to the inhomogeneous ARE (59) with completely unknown dynamics.

Lemma 3 in Section IV-A shows that the value function is in the form of a quadratic polynomial. Therefore, the quadratic polynomial basis vector for the critic network of each follower is selected as
\[ \phi_c(\zeta_i) = \begin{bmatrix} 1 & \zeta_{i1} & \cdots & \zeta_{iN^\kappa_i} & \zeta_{i1}^2 & \zeta_{i1}\zeta_{i2} & \cdots & \zeta_{iN^\kappa_i}^2 \end{bmatrix}^T \]
where \( \zeta_i = [\zeta_{i1}, \cdots, \zeta_{iN^\kappa_i}]^T \) and \( N^\kappa_i = 2n \). Then the optimal value function \( V_i(\zeta_i) \) can be approximated perfectly by the optimal critic network
\[ V_i(\zeta_i) = (W_{ci})^T\phi_c(\zeta_i), \quad (70) \]
where \( M^c = \frac{n(n+1)}{2} + 1 \) and \( W_{ci} \in \mathbb{R}^{M^c} \) is the optimal critic network weight.

Similarly, the optimal distributed containment control policy in (52) can be approximated perfectly by an optimal actor network in the form of
\[ u_i(\zeta_i) = (W_{ai})^T\phi_a(\zeta_i), \quad (71) \]
with the basis
\[ \phi_a(\zeta_i) = [1, \zeta_{i1}, \cdots, \zeta_{iN}]^T, \quad (72) \]
and \( M^a = N + 1 \) and \( W_{ai} \in \mathbb{R}^{M^a \times m} \) is the optimal actor weight.

Suppose the value function corresponding to \( u_i \) can be written as
\[ V_i(\zeta_i) = \frac{1}{2} (\zeta_i^T P_i \zeta_i + 2 (M_i)^T \zeta_i + N_i). \quad (73) \]
Denote the estimation of \( W^c \) as \( \hat{W}^c \) in the \( \kappa \)-th iteration. Then, the value function approximation in iteration \( \kappa \) is
\[ \hat{V}^\kappa_i(\zeta_i) = (\hat{W}^c_{ci})^T\phi_c(\zeta_i), \quad (74) \]
and the value function gradient approximation is
\[ \nabla \hat{V}^\kappa_i(\zeta_i) = [\nabla \phi_c(\zeta_i)]^T \hat{W}^c. \quad (75) \]
Accordingly, in iteration \( \kappa \), the estimation of \( W^a \) is \( \hat{W}^a \) and the estimation of \( u_i \) is
\[ \hat{u}_i(\zeta_i) = (\hat{W}^a_{ai})^T\phi_a(\zeta_i). \quad (76) \]
The augmented system dynamics (47) can be rewritten as
\[ \dot{\zeta}_i = A_i \zeta_i + B_i \hat{u}_i + D_i v_i + B_i (u_i - \hat{u}_i), \quad (77) \]
where \( u_i \) is the admissible policy applied to the \( i \)-th follower and \( \hat{u}_i \in \mathbb{R}^{m \times 1} \) is the policy in iteration \( \kappa \) for learning.
Considering the control policy update law
\[
\hat{u}^\kappa_{t+1} = -R_i^{-1} \hat{B}_i^T \nabla \hat{V}^\kappa_i (\zeta_i) .
\]  
(78)

Taking the time derivative of \( V^\kappa (\zeta_i) \) along the system dynamics (77) and the time derivative of \( \hat{V}^\kappa (\zeta_i) \), we get
\[
\dot{\hat{V}}^\kappa_i = \left\langle P^\kappa_i \zeta + M^\kappa_i, \dot{\zeta}_i + \hat{B}_i \hat{u}^\kappa_i + \hat{D}_i v_i \right\rangle
+ \left\langle \hat{A}_i \zeta_i + \hat{B}_i \hat{u}^\kappa_i + \hat{D}_i v_i, P^\kappa_i \zeta_i + M^\kappa_i \right\rangle
+ 2(\hat{P}^\kappa_i \zeta_i + M^\kappa_i)^T \hat{B}_i (u_i - \bar{u}_i^\kappa)
= \gamma_i \left( \zeta^T_i \hat{P}^\kappa_i \zeta_i + (M^\kappa_i)^T \zeta_i + N^\kappa_i \right) - \zeta^T_i \hat{C}_i^T Q_i \zeta_i
- (\hat{u}^\kappa_i)^T R \hat{u}^\kappa_i - 2(\hat{u}^\kappa_i+1)^T R_i (u_i - \bar{u}_i^\kappa) .
\]  
(79)

Multiplying both sides of (79) \( e^{-\gamma_i (t - t)} \) and integrating both sides on the interval \([t, t + T]\) yield the following off-policy integral RL Bellman equation that the critic weight \( \hat{W}^\kappa_i \) and actor weight \( \hat{W}^\kappa_{a+1} \) should satisfy
\[
e^{-\gamma_i T} \hat{V}^\kappa_i (\zeta_i (t + T)) - \hat{V}^\kappa_i (\zeta_i (t))
= \int^t_{t+T} \frac{d}{dt} e^{-\gamma_i (t - \tau)} \hat{V}^\kappa_i (\zeta_i (\tau)) d\tau
= \int^t_{t+T} e^{-\gamma_i (t - \tau)} - \hat{c}^T \hat{C}_i \zeta_i - (\hat{u}^\kappa_i)^T R_i \hat{u}^\kappa_i d\tau
+ \int^t_{t+T} e^{-\gamma_i (t - \tau)} - 2(\hat{u}^\kappa_i+1)^T R_i (u_i - \bar{u}_i^\kappa) d\tau .
\]  
(80)

Let \( u^\kappa_i = u_i - \hat{u}_i^\kappa \), considering the critic network (74), actor network (76) and \( R_i = \text{diag} \left\{ r^{1_j}_i \ldots r^{m_j}_i \right\} \) with \( r^{1}_i > 0, j = 1, \cdots, m \), then (80) can be equivalently expressed as in (81) in terms of the critic weight \( \hat{W}^\kappa_i \) and actor weight \( \hat{W}^\kappa_{a+1} \), where
\[
\hat{W}^\kappa_{a+1} = \begin{bmatrix} \hat{W}^\kappa_{a+1}^1 & \cdots & \hat{W}^\kappa_{a+1}^{m} \end{bmatrix}^T ,
\hat{u}^\kappa_i = \begin{bmatrix} u^\kappa_i^1 & \cdots & u^\kappa_i^{m} \end{bmatrix}^T
de is the temporal difference error [49] to be minimized in order to obtain \( \hat{W}^\kappa_i \) and \( \hat{W}^\kappa_{a+1} \).  
\[
e_i^\kappa = \left( \hat{W}^\kappa_{a+1} \right)^T \left[ e^{-\gamma_i T} \phi_c (\zeta_i (t + T)) - \phi_c (\zeta_i (t)) \right]  
+ 2 \sum_{j=1}^{m} r^{1_j}_i \left( \hat{W}^\kappa_{a+1, j} \right)^T \int^t_{t+T} e^{-\gamma_i (t - \tau)} \phi_a (\zeta_i) \hat{v}^\kappa_{i, j} d\tau  
- \int^t_{t+T} e^{-\gamma_i (t - \tau)} - \zeta^T_i \hat{C}_i \zeta_i - 2 \phi_a^T (\zeta_i) \hat{W}^\kappa_{a+1} R_i \left( \hat{W}^\kappa_{a+1} \right)^T \phi_a (\zeta_i) d\tau .
\]  
(81)

By rearranging (81), the off-policy integral RL algorithm will be translated into solving the least square problem
\[
y^\kappa (t) = \left( \hat{W}^\kappa \right)^T \Phi (t) - e^\kappa (t) ,
\]  
(82)

where
\[
y^\kappa (t) = \int^t_{t+T} e^{-\gamma_i (t - \tau)} \left[ -\zeta_i^T C_i \zeta_i + \phi_a (\zeta_i) \right] d\tau ,
\]  
(83)
\[
\hat{W}^\kappa = \begin{bmatrix} \left( \hat{W}^\kappa_{a, 1} \right)^T \left( \hat{W}^\kappa_{a, 1} \right)^T \cdots \left( \hat{W}^\kappa_{a, m} \right)^T \end{bmatrix}^T ,
\]  
(84)
\[
e^\kappa \left[ e_1^\kappa \cdots e_\kappa^m \right]^T .
\]  
(85)

**Lemma 4.** The solution to the least square problem (82) is equivalent to the policy evaluation (68) and the policy improvement (69) in the Algorithm 1. Moreover, the convergence of \( u^\kappa_i \) as \( \kappa \to \infty \) to optimal policy \( u^* \) (52) can be guaranteed.

\[ \square \]

Proof: The proof is similar to [41] and is omitted.

To this end, the off-policy RL algorithm, which solves the inhomogeneous ARE (59) in a model-free manner, is given in Algorithm 2.

**Algorithm 2** Model-Free Off-policy Integral RL Algorithm

1: Initialization: set \( \kappa = 0 \), and start with an admissible control policy \( u^\kappa_0 \);
2: Solve the least square problem (82) to obtain \( \hat{W}^\kappa_0 \) and \( \hat{W}^\kappa_{a+1} \) simultaneously;
3: Stop if convergence is achieved, otherwise set \( \kappa = \kappa + 1 \) and go to 2;
4: On convergence set \( u^* = \hat{u}^\kappa_i (\zeta_i) = \left( \hat{W}^\kappa_{a+1} \right)^T \phi_a (\zeta_i) \).

**Remark 9.** According to Lemma 4, Algorithm 2 is equivalent to Algorithm 1 in the sense that they generate the same sequence \( \{ u^\kappa_i (x_i) \}_{\kappa=0}^{\infty} \) and \( \{ V^\kappa_i (x_i) \}_{\kappa=0}^{\infty} \) when both algorithms start from the same initial admissible policy \( u^0_0 (x_i) \). Hence the convergence of policy sequence generated by Algorithm 2 to the optimal policy \( u^* \) (52) can be guaranteed. However, Algorithm 1 is essentially a model-based iterative methods because complete knowledge of agents’ dynamics, \( (A_i, B_i) \) for leaders and \( (A_j, B_j) \) for followers, are required in both policy evaluation (68) and policy improvement (69). In contrast, Algorithm 2, a partially model-free algorithm, is based on the data-driven least square problem (82), which does not require followers’ dynamics. It should be noted that the leaders’ dynamics is required for the distributed adaptive observer design to generate reference model for the followers.

**VII. SIMULATION**

In this section, a simulation example is provided to confirm the validity of the presented approach.
Consider the multi-agent system with the graph topology shown in Figure 2, where shaded nodes 1, 2 and 3 denote the leaders, and all the other nodes are followers. The interaction among the leaders is captured by an undirected graph, whereas the connection between the leaders and the followers is described by a directed graph. The weighted adjacency matrices $\mathbf{Q}$, $\mathbf{R}_i$ of the value function and the discounting factor $\gamma_i$ for the distributed optimal tracking problem in (45) are selected as

$$
\mathbf{Q}_4 = \begin{bmatrix} 70 & 0 \\ 0 & 100 \end{bmatrix}, \quad \mathbf{R}_4 = 0.05, \quad \gamma_4 = 0.037,
$$

$$
\mathbf{Q}_5 = \begin{bmatrix} 10 & 0 \\ 0 & 8 \end{bmatrix}, \quad \mathbf{R}_5 = 0.0055, \quad \gamma_5 = 0.026,
$$

$$
\mathbf{Q}_6 = \begin{bmatrix} 9.5 & 0 \\ 0 & 9 \end{bmatrix}, \quad \mathbf{R}_6 = 0.0065, \quad \gamma_6 = 0.03,
$$

$$
\mathbf{Q}_7 = \begin{bmatrix} 8 & 0 \\ 0 & 9 \end{bmatrix}, \quad \mathbf{R}_7 = 0.0085, \quad \gamma_7 = 0.035,
$$

$$
\mathbf{Q}_8 = \begin{bmatrix} 8.5 & 0 \\ 0 & 7.5 \end{bmatrix}, \quad \mathbf{R}_8 = 0.0075, \quad \gamma_8 = 0.03,
$$

$$
\mathbf{Q}_9 = \begin{bmatrix} 9 & 0 \\ 0 & 6 \end{bmatrix}, \quad \mathbf{R}_9 = 0.008, \quad \gamma_9 = 0.04.
$$

The design parameters of the distributed observer in (11) - (13) are selected as

$$
\mathbf{F} = \begin{bmatrix} 0.6326 & -0.5960 \\ 0.4369 & 1.1692 \end{bmatrix}, \quad \mathbf{T} = \begin{bmatrix} 0.4001 & -0.3770 \\ -0.3770 & 0.03552 \end{bmatrix},
$$

$$
\mathbf{P} = \begin{bmatrix} 0.6326 & -0.5960 \\ 0.4369 & 1.1692 \end{bmatrix}, \quad \tau_4 = 0.6, \quad \tau_5 = 0.8, \quad \tau_6 = 1.0, \quad \tau_7 = 1.2, \quad \tau_8 = 0.9, \quad \tau_9 = 1.5
$$

The results of the distributed adaptive observer design are shown in Figures 3 and 4. It can be seen from Figure 3 that all the distributed observers converge at the interior part in the convex hull spanned the leaders. The evolution of the distributed coupling gain $\tilde{d}_i$, $i \in \mathcal{F}$ is shown in Figure 4.

The heterogeneous followers are given by (2) for $i = 4 - 9$ and the corresponding dynamics are given by

$$
\mathbf{A}_4 = \begin{bmatrix} 0 & 1 \\ -19 & -1.5 \end{bmatrix}, \quad \mathbf{B}_4 = \begin{bmatrix} 0 \\ 3.9 \end{bmatrix},
$$

$$
\mathbf{A}_5 = \begin{bmatrix} 0 & 1 \\ -6.5 & -3.5 \end{bmatrix}, \quad \mathbf{B}_5 = \begin{bmatrix} 0 \\ 3.5 \end{bmatrix},
$$

$$
\mathbf{A}_6 = \begin{bmatrix} 0 & 1 \\ -9.5 & -0.85 \end{bmatrix}, \quad \mathbf{B}_6 = \begin{bmatrix} 0 \\ 6.2 \end{bmatrix},
$$

$$
\mathbf{A}_7 = \begin{bmatrix} 0 & 1 \\ -7.5 & -0.55 \end{bmatrix}, \quad \mathbf{B}_7 = \begin{bmatrix} 0 \\ 4 \end{bmatrix},
$$

$$
\mathbf{A}_8 = \begin{bmatrix} 0 & 1 \\ -7 & -0.4 \end{bmatrix}, \quad \mathbf{B}_8 = \begin{bmatrix} 0 \\ 3 \end{bmatrix},
$$

$$
\mathbf{A}_9 = \begin{bmatrix} 0 & 1 \\ -5 & -0.2 \end{bmatrix}, \quad \mathbf{B}_9 = \begin{bmatrix} 0 \\ 2 \end{bmatrix}.
$$
FIGURE 5: The evolution of the state trajectories for leaders and followers over time.

FIGURE 6: The evolution of the containment error for each follower over time.

optimal distributed feedback gain in (52) for $i \in \mathcal{F}$ can be obtained as


Finally, the results of the containment control problem solved by the model-free RL method are illustrated in Figure 5 – 7. These results show that over time the containment error $\xi_i(t)$ for all the followers approach to the origin asymptotically and the state trajectories $x_i(t)$ for all the followers converge at the interior part of the convex hull spanned by the leaders.

VIII. CONCLUSION

This paper investigated the containment control problem of MASs with linear general high-order dynamics and non-autonomous leaders. The desired trajectory for each follower, i.e., the convex hull spanned by all leaders is estimated by a fully distributed adaptive observer without any information from the global communication graph. The fully distributed adaptive observers serve as the distributed reference model for each follower. Then the distributed containment control of heterogeneous MASs with multiple non-autonomous leaders is formulated as a distributed MRAC problem. Contrary to existing results on distributed MRAC, this paper imposes optimality explicitly to the transient performance of the containment error and obtains distributed optimal model reference control. For each follower, the distributed optimal model reference control was obtained by solving an inhomogeneous ARE, which is solved in a model-free manner based on the combination of an off-policy RL and an actor-critic structure. A simulation is carried out to verify the presented approach in this paper.

ACKNOWLEDGMENT

The authors would like to thank Dr. Zhongkui Li at Peking University for his insightful discussions about some of the issues addressed in this paper and the anonymous reviewers for their valuable comments and suggestions to improve the quality of this paper.

REFERENCES


