2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Optimal Containment Control of Nonlinear MASs: A Time-Aggregation-based Policy Iteration Algorithm

Xiongtao Shi, Yanjie Li, *Member, IEEE*, Chenglong Du, *Member, IEEE*

Abstract— **In this paper, the optimal containment control of a class of unknown nonlinear multi-agent systems (MASs) is studied via a time-aggregation (TA) based model-free reinforcement learning (RL) algorithm. By proposing TA-based event-state, event-control, and integration-reward, the model-free TA-based policy iteration (TA-PI) approach is synthesized such that the policy evaluation and policy improvement steps are only executed for finite event-state, and the optimal control protocol is obtained with fewer computational requirements. Besides, the control input is intermittently updating only when the event-set is visited, which greatly reduce the updating frequency of control. Therefore, the proposed learning algorithm helps to save computational resources in both learning process and control updating. Moreover, armed with a finite predefined event-set, the developed TA-PI algorithm without employing function approximator and state discretization, resulting a strict convergence analysis via the mathematical induction. Finally, simulation results are given to show the feasibility and effectiveness of the proposed algorithm.**

Index Terms— **Time-aggregation, policy iteration, model-free control, optimal containment control.**

I. Introduction

The containment control of multi-agent systems (MASs) has received great attention in the past two decades due to its wide applications [1]–[4], including the smart transportation, emergency rescue, and other scenarios. Note that most existing works only study the stability of containment control, which is the basic requirement in system design. To achieve the containment control in a better way, the optimal containment control is developed such that not only the followers enter the convex hull spanned by multiple leaders, but also a predefined performance index is minimized for a better control performance [5]–[8].

To achieve this optimal containment control, the model-free reinforcement learning (RL) algorithm, such as policy iteration (PI) and value iteration (VI), have been taken into consideration, and enable the synthesis of the control protocol in an optimal model-free manner [6]–[8]. Nevertheless, the majority of existing literature requires a lot of computational resources, because the learning process is executed for continuous and uncountable states rather than finite important event-state, and the learned optimal control protocol usually needs persistently updating [9], [10]. Moreover, the practical implementation of existing model-free RL-based optimal control protocol requires a function approximator or state discretization. When a function approximator is employed, the theoretical convergence of the algorithm can not be strictly guaranteed due to the appearance of approximation error [11], and on the contrary, using state discretization can lead to the curse of dimensionality if the discretization is particularly accurate [12].

To overcome these drawbacks, a novel time-aggregation (TA) technique [13]–[15] is introduced in this paper to synthesize TA-based policy iteration (TA-PI) algorithm, in which the continuous and uncountable state is replaced by the finite predefined event-set, and enable a tremendous reduction of the state space. The contributions of this paper are as follows.

1. The improved TA-based event-state, event-control, and integration-reward are developed. Thus, the learning processes, such as the policy evaluation and policy improvement, are only executed when the current state belongs to event-set, enabling a more computationally efficient way compared with conventional learning algorithms [9], [10]. Besides, the developed TA-based event-control effectively avoids persistent control updating, which can greatly reduce more computing consumption in the control updating than [16], [17].

2. By introducing a finite predefined event-set, the utilization of the function approximator and state discretization are avoided in the proposed TA-PI algorithm. Meanwhile, the convergences of the proposed TA-PI algorithm is proved based on the mathematical induction, and the monotonicity and boundedness property of the iterative value function are derived in detail.

Notations: $\mathbb{R}$ indicates the set of real numbers; $\mathbb{R}^n$ represents the set of real vectors with n elements; $\mathbb{R}^{n\times m}$ stands for the set of real matrices with n rows and m columns; I_l indicates the l dimensional identity matrix; $\otimes$ represents Kronecker product; $\mathrm{diag}\{d_1,...,d_n\}$ is a diagonal matrix whose diagonal entries are $d_1,...,d_n$ and all other entries are zero.

II. Preliminaries and problem formulation

A. Preliminaries

The considered communication graph $\mathcal{G}(\mathcal{V},\mathcal{E})$ is composed of nodes set $\mathcal{V}=\{1,2,...,n\}$ and edges set $\mathcal{E}\subseteq\mathcal{V}\times\mathcal{V}$. The corresponding adjacent matrix is represented by

This work was supported by the National Natural Science Foundation of China (61977019, U1813206) and the Shenzhen Fundamental Research Program (JCYJ20220818102415033, JCYJ20180507183837726, JSGG20201103093802006). *(Corresponding author: Yanjie Li)*

Xiongtao Shi and Yanjie Li are with the School of Mechanical Engineering and Automation, Harbin Institute of Technology (Shenzhen), Shenzhen, 518055, China (e-mail:xiongtaoshi@stu.hit.edu.cn; autolyj@hit.edu.cn). Chenglong Du is with the School of Automation, Central South University, Changsha, 410083, China (e-mail:chenglong_du@csu.edu.cn).

979-8-3503-0123-6/23/$31.00 ©2023 IEEE

$\mathcal{A} = [a_{ij}] \in \mathbb{R}^{n\times n}$, where $a_{ij} = 0$ if $(v_j, v_i) \notin \mathcal{E}$ and $a_{ij} = 1, (i \neq j)$ otherwise. Define neighbors set of node v_i as $\mathcal{N}_i = \{j : v_j \in \mathcal{V}, (v_j, v_i) \in \mathcal{E}\}$. The agent with the empty neighbor set is a leader, and it is a follower otherwise. A (directed) path from agent 1 to agent l is a sequence of edges as the form $(1, 2), (2, 3), ..., (l-1, l)$ with distinct nodes. A directed spanning forest is a graph $\mathcal{G}$ in which there exists at least one available leader for any follower [18]. The Laplacian matrix $\mathcal{L}$ is calculated as $\mathcal{L} = \mathcal{D} - \mathcal{A}$, where $\mathcal{D} = \text{diag}\{d_1, d_2, ..., d_n\}$, $d_i = \sum_{j\in\mathcal{N}_i} a_{ij}$.

B. Problem formulation

Consider an unknown nonlinear MAS consisting of m leaders and $n-m$ $(n > m)$ followers with the following dynamics:

$$\begin{cases} \dot{x}_i(t) = 0, & i \in L, \\ \dot{x}_i(t) = f_i(x_i(t), u_i(t)), & i \in F, \end{cases} \tag{1}$$

where $x_i(t) \in \mathbb{R}$ and $u_i(t) \in \mathbb{R}$ represent the state and control input, respectively; $L = \{1, ..., m\}$ and $F = \{m+1, ..., n\}$ represent the leader set and follower set, respectively; $f_i(\cdot)$ is an unknown nonlinear function. To achieve optimal containment control, the performance index is defined as:

$$J_i(e_i(t), u_i(t)) = \textstyle\int_t^\infty r_i(e_i(\tau), u_i(\tau))d\tau, \tag{2}$$

where

$$r_i(e_i(t), u_i(t)) = e_i(t)Q_ie_i(t) + u_i(t)P_iu_i(t), \tag{3}$$

$e_i(t) = \sum_{j\in\mathcal{N}_i} a_{ij}(x_j(t) - x_i(t))$ is a network-induced error; $Q_i > 0$ and $P_i > 0$.

With m leaders and $n-m$ followers, the Laplacian matrix $\mathcal{L}$ can be spitted as

$$\mathcal{L} = \left[\begin{array}{c|c} 0_{m\times m} & 0_{m\times(n-m)} \\ \hline \mathcal{L}_1 & \mathcal{L}_2 \end{array}\right], \tag{4}$$

where $\mathcal{L}_1 \in \mathbb{R}^{(n-m)\times m}$ and $\mathcal{L}_2 \in \mathbb{R}^{(n-m)\times(n-m)}$.

To proceed, we give following assumption, lemma, and definition in advance.

Assumption 2.1: There exists a directed spanning forest in $\mathcal{G}$.

Lemma 2.1 ([19]): Under Assumption 2.1, all the real parts of eigenvalues of $\mathcal{L}_2$ are positive. In addition, each element of $-\mathcal{L}_2^{-1}\mathcal{L}_1$ is non-negative and all row sums of $-\mathcal{L}_2^{-1}\mathcal{L}_1$ equal to one.

Definition 2.2 ([18]): If $(1-\gamma)x + \gamma y \in K$ for $\gamma \in [0, 1]$ and $\forall x, y \in K$, then $K \in \mathbb{R}^l$ is said to be convex. $Co\{x_1, ..., x_m\}$ represents the minimal convex hull spanned by a finite set of points $x_1, ..., x_m \in \mathbb{R}^l$. More specifically, $Co\{x_1, ..., x_m\} = \{\sum_{i=1}^m \alpha_i x_i | \alpha_i \geq 0, \alpha_i \in \mathbb{R}, \sum_{i=1}^m \alpha_i = 1\}$.

Let $x_L(t) = [x_1(t), ..., x_m(t)]^T$ and $x_F(t) = [x_{m+1}(t), ..., x_n(t)]^T$. Define the containment error as

$$\delta(t) = x_F(t) + \mathcal{L}_2^{-1}\mathcal{L}_1 x_L(t), \tag{5}$$

where $\delta(t) = [\delta_1(t), ..., \delta_{n-m}(t)]^T$. For each follower $i \in F$, we have

$$\delta_{i-m}(t) = x_i(t) + \sum_{j=1}^m h_{(i-m)j}x_j(t), \tag{6}$$

where $h_{(i-m)j}$ is $((i-m), j)$th element of $\mathcal{L}_2^{-1}\mathcal{L}_1$ satisfying

$$-h_{(i-m)j} \geq 0, \; -\sum_{j=1}^m h_{(i-m)j} = 1 \tag{7}$$

from Lemma 2.1. With Definition 2.2, if $\lim_{t\to\infty} \delta_i(t) = 0$, one has $\lim_{t\to\infty} x_i(t) = -\lim_{t\to\infty} \sum_{j=1}^m h_{(i-m)j}x_j(t)$, which implies follower i reach the convex hull spanned by the leaders with coefficients $-h_{(i-m)j}$. Thus, we refer to $\delta(t)$ as the containment error.

Remark 2.3: Note that the containment control problem of high-dimensional MASs can be solved by simply expanding the dimension via the Kronecker product. For example, assume each agent with $l > 1$ dimensions, (5) can be extended as $\delta(t) = x_F(t) + (\mathcal{L}_2^{-1}\mathcal{L}_1 \otimes I_l)x_L(t)$, where $\delta(t) \in \mathbb{R}^{(n-m)l}$, $x_L(t) = [x_1^T(t), ..., x_m^T(t)]^T \in \mathbb{R}^{ml}$, $x_F(t) = [x_{m+1}^T(t), ..., x_n^T(t)]^T \in \mathbb{R}^{(n-m)l}$. Thus, without loss of generality, in this paper, the one-dimensional MASs is considered.

In the following, we recall the definition of optimal containment control of a class of unknown nonlinear MASs

Definition 2.4: Consider a nonlinear MAS (1) with unknown dynamics over directed graph $\mathcal{G}$ satisfying Assumption 2.1. For any bounded $x_i(0)$, the optimal containment control is achieved if the containment error satisfies $\lim_{t\to\infty} \delta(t) = 0$, and the performance index (2) is minimized meanwhile.

The goal of this paper is to achieve optimal containment control while avoiding great computing requirement in both learning process and control updating, and ensuring theoretical convergence strictly.

III. TA-BASED EVENT-SET

In this section, the TA-based event-set is formulated for the preparation of TA-PI algorithm. Based on this event-set, the event-triggered control protocol and integration-reward are developed in the sequel. Different from conventional definition, the TA-based event-set help to reduce state space greatly; the event-triggered control protocol makes the control updating be intermittent; and the integration-reward is obtained with time-varying integration length.

First, with Lebesgue sampling [13], a finite subset of the full state space is picked out as a finite predefined event-set:

$$\mathbb{S}_i^{\mathbb{D}_i} = \{s_i^d : d \in \mathbb{D}_i\}, \tag{8}$$

where s_i^d is the important state for agent i, $\mathbb{D}_i = \{1, ..., D_i\}$, and D_i is the size of predefined event-set for agent i.

Follow from (8), the corresponding event-triggered mechanism is formulated as

$$t_{k+1}^i = \min\{t : t > t_k^i, \; e_i(t) \in \mathbb{S}_i^{\mathbb{D}_i}, \; e_i(t^-) \notin \mathbb{S}_i^{\mathbb{D}_i}\}, \tag{9}$$

where t^- is the left limit of t. Obviously, the event is occurred only when network-induced error $e_i(t)$ belongs to the predefined event-set. Thus, the corresponding event-triggered control protocol is designed as:

$$u_i(t) = u_i(t_k^i),\ t_k^i \leq t < t_{k+1}^i,\ i \in F. \tag{10}$$

With this event-triggered control protocol, the integration-reward is given as

$$R_i(e_i(t_k^i), u_i(t_k^i)) = \textstyle\int_{t_k^i}^{t_{k+1}^i} r_i(e_i(\tau), u_i(\tau))d\tau. \tag{11}$$

To minimize (2), a value function is defined as

$$V_i(e_i(t_k^i)) = \textstyle\int_{t_k^i}^{\infty} r_i(e_i(\tau), u_i(\tau))d\tau. \tag{12}$$

And the optimal value function can be written as

$$V_i^*(e_i(t_k^i)) = \min_{u_i}\{\textstyle\int_{t_k^i}^{\infty} r_i(e_i(\tau), u_i(\tau))d\tau\}. \tag{13}$$

Armed with the (12) and the Bellman optimal principle, the optimal value function can be rewritten as an iterative form:

$$\begin{aligned} & V_i^*(e_i(t_k^i)) \\ &= \min_{u_i}\{\int_{t_k^i}^{t_{k+1}^i} r_i(e_i(\tau), u_i(\tau))d\tau + V_i^*(e_i(t_{k+1}^i))\} \\ &= \min_{u_i}\{R_i(e_i(t_k^i), u_i(t_k^i)) + V_i^*(e_i(t_{k+1}^i))\}. \end{aligned} \tag{14}$$

The following definition and assumptions are needed for further analysis.

Definition 3.1 ([20]): The control protocol $u_i(t)$ are said to be admissible if $\lim_{t\to\infty} e_i(t) = 0$ and $J_i(e_i(t), u_i(t))$ is finite.

Assumption 3.2: The initial control protocol is admissible.

Assumption 3.3: By properly picking $V_i^0 = V_c > 0, 0 < \alpha_i \leq 1 \leq \beta_i < \infty$, the inequalities, $0 \leq \alpha_i V_i^* \leq V_i^0 \leq \beta_i V_i^*$, are hold.

Assumption 3.4: The optimal value function satisfies $0 \leq V_i^*(e_i(t_{k+1}^i)) \leq \theta_i R_i(e_i(t_k^i), u_i(t_k^i))$, where $0 < \theta_i < \infty$.

IV. Model-free TA-PI algorithm

In this section, based on the designed TA-based event-set, event-triggered control, and integration-reward, a TA-PI algorithm is developed to obtain the optimal control protocol in a model-free manner, and its convergence analysis is given later.

The developed TA-PI algorithm given in Algorithm 1.

Algorithm 1 Model-free TA-PI algorithm.

```
Initialization: Set the initial value function to a constant value $V_c$. Given an initial admissible control policy $u_i^0$. Select a small threshold $\varepsilon > 0$. Then, for each $e_i(t_k^i) \in \mathbb{S}_i^{\mathbb{D}_i}$ perform the following iteration for index $l$.
while $|V_i^{0,l+1}(e_i(t_k^i)) - V_i^{0,l}(e_i(t_k^i))| \leq \varepsilon$ do
    $V_i^{0,l+1}(e_i(t_k^i)) = R_i(e_i(t_k^i), u_i^0(t_k^i)) + V_i^{0,l}(e_i(t_{k+1}^i))$.
end while
while $|V_i^{s,0}(e_i(t_k^i)) - V_i^{s-1,0}(e_i(t_k^i))| \leq \varepsilon$ do
    Policy improvement: for each $e_i(t_k^i) \in \mathbb{S}_i^{\mathbb{D}_i}$, optimize $u_i^s$ under $V_i^{s,0}$, i.e., perform $u_i^s(e_i(t_k^i)) = \operatorname*{argmin}_{u_i}\{R_i(e_i(t_k^i), u_i(t_k^i)) + V_i^{s,0}(e_i(t_{k+1}^i))\}$.
    while $|V_i^{s,l+1}(e_i(t_k^i)) - V_i^{s,l}(e_i(t_k^i))| \leq \varepsilon$ do
        Policy evaluation: for each $e_i(t_k^i) \in \mathbb{S}_i^{\mathbb{D}_i}$, optimize $V_i^{s,l+1}$ under $u_i^s$, i.e., perform $V_i^{s,l+1}(e_i(t_k^i)) = R_i(e_i(t_k^i), u_i^s(t_k^i)) + V_i^{s,l}(e_i(t_{k+1}^i))$.
    end while
end while
Return $u_i^s(e_i(t_k^i))$.
```

Remark 4.1: From the algorithm 1, the proposed TA-PI algorithm has finite iterations because the calculation is only for the finite event-state $e_i(t_k^i) \in \mathbb{S}_i^{\mathbb{D}_i}$, thus saving a lot of computing resources. In addition, the designed event-control protocol (10) can greatly reduce update frequency, and the calculation consumption can be further decreased.

Theorem 4.2: Consider the unknown nonlinear MAS (1), satisfying Assumption 2.1, the proposed event-control protocol is designed as (10) which will learn via algorithm 1, then the optimal containment control of MAS (1) is achieved. Moreover, the following theoretical properties can be guaranteed:

(1): $\infty \geq V_i^{s,l+1}(e_i(t_k^i)) \geq V_i^{s,l}(e_i(t_k^i)),\ s = 0,\ l \geq 0$,

(2): $V_i^{s,l+1}(e_i(t_k^i)) \leq V_i^{s,l}(e_i(t_k^i)),\ s \geq 1,\ l \geq 0$,

(3): $V_i^{s+1,0}(e_i(t_k^i)) \leq V_i^{s,0}(e_i(t_k^i)),\ s \geq 1$,

(4): $V_i^{s,1}(e_i(t_k^i)) \leq (1 + \frac{\beta_i - 1}{(1+\theta_i^{-1})^s})V_i^*(e_i(t_k^i)),\ s \geq 1$,

(5): $\lim_{s\to\infty} V_i^{s,l}(e_i(t_k^i)) = V_i^*(e_i(t_k^i)),\ l \geq 0$,

(6): $\lim_{s\to\infty} u_i^s(e_i(t_k^i)) = u_i^*(e_i(t_k^i))$.

Proof: In the following, we first prove the six theoretical properties in Theorem 4.2, and further analyzes the convergence of optimal containment control.

Property (1): With the fact that $V_i^{0,0}(e_i(t_k^i)) = V_c$, one has

$$\begin{aligned} & V_i^{0,1}(e_i(t_k^i)) \\ &= R_i(e_i(t_k^i), u_i^0(t_k^i)) + V_i^{0,0}(e_i(t_{k+1}^i)) \\ &= R_i(e_i(t_k^i), u_i^0(t_k^i)) + V_i^{0,0}(e_i(t_k^i)) \\ &\geq V_i^{0,0}(e_i(t_k^i)). \end{aligned} \tag{15}$$

Assuming (15) holds for iterative index l, i.e., $V_i^{0,l}(e_i(t_k^i)) \geq V_i^{0,l-1}(e_i(t_k^i))$ for each $e_i(t_k^i) \in \mathbb{S}_i^{\mathbb{D}_i}$, we have

$$\begin{aligned} & V_i^{0,l+1}(e_i(t_k^i)) \\ &= R_i(e_i(t_k^i), u_i^0(t_k^i)) + V_i^{0,l}(e_i(t_{k+1}^i)) \\ &\geq R_i(e_i(t_k^i), u_i^0(t_k^i)) + V_i^{0,l-1}(e_i(t_{k+1}^i)) \\ &= V_i^{0,l}(e_i(t_k^i)). \end{aligned} \tag{16}$$

Thus, we can obtain that $V_i^{0,l+1}(e_i(t_k^i)) \geq V_i^{0,l}(e_i(t_k^i))$. Furthermore, based on Assumption 3.2, one has $\infty \geq V_i^{1,0}(e_i(t_k^i)) = V_i^{0,\infty}(e_i(t_k^i)) \geq V_i^{0,l+1}(e_i(t_k^i))$.

Property (2): From the **policy improvement** step of TA-PI algorithm, one has that $R_i(e_i(t_k^i), u_i(t_k^i)) + V_i^{s,0}(e_i(t_{k+1}^i))$ will be minimized under policy $u_i^s(t_k^i)$. Therefore, it yields that

$$\begin{aligned} & V_i^{s,1}(e_i(t_k^i)) \\ &= R_i(e_i(t_k^i), u_i^s(t_k^i)) + V_i^{s,0}(e_i(t_{k+1}^i)) \end{aligned}$$

$$
\begin{aligned}
&= \min_{u_i}\{R_i(e_i(t_k^i), u_i(t_k^i)) + V_i^{s,0}(e_i(t_{k+1}^i))\}\\
&= \min_{u_i}\{R_i(e_i(t_k^i), u_i(t_k^i)) + V_i^{s-1,\infty}(e_i(t_{k+1}^i))\}\\
&\le R_i(e_i(t_k^i), u_i^{s-1}(t_k^i)) + V_i^{s-1,\infty}(e_i(t_{k+1}^i))\\
&= V_i^{s-1,\infty}(e_i(t_k^i))\\
&= V_i^{s,0}(e_i(t_k^i)).
\end{aligned} \tag{17}
$$

Assuming (17) holds for iterative index l, i.e., $V_i^{s,l}(e_i(t_k^i)) \le V_i^{s,l-1}(e_i(t_k^i))$ for each $e_i(t_k^i) \in \mathbb{S}_i^{\mathbb{D}_i}$, it follows that

$$
\begin{aligned}
&V_i^{s,l+1}(e_i(t_k^i))\\
&= R_i(e_i(t_k^i), u_i^s(t_k^i)) + V_i^{s,l}(e_i(t_{k+1}^i))\\
&\le R_i(e_i(t_k^i), u_i^s(t_k^i)) + V_i^{s,l-1}(e_i(t_{k+1}^i))\\
&= V_i^{s,l}(e_i(t_k^i)).
\end{aligned} \tag{18}
$$

Therefore, one can derive that for any iteration index $s \ge 1$ and iterative index $l \ge 0$, the inequality $V_i^{s,l}(e_i(t_k^i)) \le V_i^{s,l-1}(e_i(t_k^i))$ holds.

Property (3): According to $V_i^{s+1,0}(e_i(t_k^i)) = V_i^{s,\infty}(e_i(t_k^i))$ and $V_i^{s,\infty}(e_i(t_k^i)) \le V_i^{s,0}(e_i(t_k^i))$, it is concluded that $V_i^{s+1,0}(e_i(t_k^i)) \le V_i^{s,0}(e_i(t_k^i))$.

Property (4): The initial condition is need to be proven when the iterative index $s = 1, l = 0$. By the **policy improvement** step, the following value $R_i(e_i(t_k^i), u_i^1(t_k^i)) + V_i^{1,0}(e_i(t_{k+1}^i))$ will be minimized under policy $u_i^1(e_i(t_k^i))$. Therefore, we have

$$
\begin{aligned}
&V_i^{1,1}(e_i(t_k^i))\\
&= R_i(e_i(t_k^i), u_i^1(t_k^i)) + V_i^{1,0}(e_i(t_{k+1}^i))\\
&= \min_{u_i}\{R_i(e_i(t_k^i), u_i(t_k^i)) + V_i^{1,0}(e_i(t_{k+1}^i))\}.
\end{aligned} \tag{19}
$$

Due to $\infty \ge V_i^{1,0}(e_i(t_k^i)) = V_i^{0,\infty}(e_i(t_k^i)) \ge V_i^{0,l+1}(e_i(t_k^i))$, one can find a large enough parameter β_i such that $V_i^{1,0} \le \beta_i V_i^*$. Furthermore, from Assumptions 3.3 and 3.4, one can obtain

$$
\begin{aligned}
&V_i^{1,1}(e_i(t_k^i))\\
&= R_i(e_i(t_k^i), u_i^1(t_k^i)) + V_i^{1,0}(e_i(t_{k+1}^i))\\
&\le R_i(e_i(t_k^i), u_i^1(t_k^i)) + \beta_i V_i^*(e_i(t_{k+1}^i))\\
&\le \min_{u_i}\{R_i(e_i(t_k^i), u_i(t_k^i)) + \beta_i V_i^*(e_i(t_{k+1}^i))\\
&\quad + \frac{\beta_i - 1}{1+\theta_i}[\theta_i R_i(e_i(t_k^i), u_i(t_k^i)) - V_i^*(e_i(t_{k+1}^i))]\}\\
&= \min_{u_i}\{(1 + \frac{\beta_i - 1}{1+\theta_i^{-1}})R_i(e_i(t_k^i), u_i(t_k^i))\\
&\quad + (1 + \frac{\beta_i - 1}{1+\theta_i^{-1}})V_i^*(e_i(t_{k+1}^i))\}\\
&= (1 + \frac{\beta_i - 1}{1+\theta_i^{-1}})\min_{u_i}\{R_i(e_i(t_k^i), u_i(t_k^i)) + V_i^*(e_i(t_{k+1}^i))\}\\
&= (1 + \frac{\beta_i - 1}{1+\theta_i^{-1}})V_i^*(e_i(t_k^i)).
\end{aligned} \tag{20}
$$

Assume (20) holds for iterative index s, i.e., $V_i^{s,1}(e_i(t_k^i)) \le (1 + \frac{\beta_i - 1}{(1+\theta_i^{-1})^s})V_i^*(e_i(t_k^i))$ for each $e_i(t_k^i) \in \mathbb{S}_i^{\mathbb{D}_i}$, then, with

$$
V_i^{s+1,0}(e_i(t_{k+1}^i)) = V_i^{s,\infty}(e_i(t_{k+1}^i)) \le V_i^{s,1}(e_i(t_{k+1}^i)), \tag{21}
$$

it is concluded that

$$
\begin{aligned}
&V_i^{s+1,1}(e_i(t_k^i))\\
&= R_i(e_i(t_k^i), u_i^{s+1}(t_k^i), t_{k+1}^i) + V_i^{s+1,0}(e_i(t_{k+1}^i))\\
&\le \min_{u_i}\{R_i(e_i(t_k^i), u_i(t_k^i)) + V_i^{s,1}(e_i(t_{k+1}^i))\}.
\end{aligned} \tag{22}
$$

Then, armed with Assumption 3.4, it yields that

$$
\begin{aligned}
&V_i^{s+1,1}(e_i(t_k^i))\\
&\le \min_{u_i}\{R_i(e_i(t_k^i), u_i(t_k^i))\\
&\quad + (1 + \frac{\beta_i - 1}{(1+\theta_i^{-1})^s})V_i^*(e_i(t_{k+1}^i)) + \frac{(\beta_i - 1)\theta_i^s}{(1+\theta_i)^{s+1}}\\
&\quad \times [\theta_i R_i(e_i(t_k^i), u_i(t_k^i)) - V_i^*(e_i(t_{k+1}^i))]\}\\
&= \min_{u_i}\{(1 + \frac{\beta_i - 1}{(1+\theta_i^{-1})^{s+1}})R_i(e_i(t_k^i), u_i(t_k^i))\\
&\quad + (1 + \frac{\beta_i - 1}{(1+\theta_i^{-1})^{s+1}})V_i^*(e_i(t_{k+1}^i))\}\\
&= (1 + \frac{\beta_i - 1}{(1+\theta_i^{-1})^{s+1}})\min_{u_i}\{R_i(e_i(t_k^i), u_i(t_k^i)) + V_i^*(e_i(t_{k+1}^i))\}\\
&= (1 + \frac{\beta_i - 1}{(1+\theta_i^{-1})^{s+1}})V_i^*(e_i(t_k^i)).
\end{aligned} \tag{23}
$$

Therefore, it is concluded that for any iterative index $s \ge 1$, one has $V_i^{s,1}(e_i(t_k^i)) \le (1 + \frac{\beta_i - 1}{(1+\theta_i^{-1})^s})V_i^*(e_i(t_k^i)), s \ge 1$.

Property (5): Since the optimal value is the minimum value, one has $V_i^*(e_i(t_k^i)) \le V_i^{s,l}(e_i(t_k^i))$. Furthermore, with the conditions $V_i^{s,1}(e_i(t_k^i)) \le (1 + \frac{\beta_i - 1}{(1+\theta_i^{-1})^s})V_i^*(e_i(t_k^i)), s \ge 1$ and $V_i^{s,l+1}(e_i(t_k^i)) \le V_i^{s,l}(e_i(t_k^i)), s \ge 1, l \ge 0$, we have $V_i^*(e_i(t_k^i)) \le V_i^{\infty,l}(e_i(t_k^i)) \le V_i^*(e_i(t_k^i))$.

Property (6): Follow from $\lim_{s\to\infty} V_i^{s,l}(e_i(t_k^i)) = V_i^*(e_i(t_k^i))$, one can get the optimized policy with the **policy improvement** step of TA-PI algorithm, which means that $\lim_{s\to\infty} u_i^s(e_i(t_k^i)) = u_i^*(e_i(t_k^i))$.

Finally, with the fact that $\lim_{s\to\infty} V_i^{s,l}(e_i(t_k^i)) = V_i^*(e_i(t_k^i))$ and (12), it is concluded that $\lim_{t\to\infty} r_i(e_i(t), u_i(t)) = 0$, i.e., $\lim_{t\to\infty} \delta_i(t) = 0$. As a result, the optimal containment control is achieved as expected. ■

Remark 4.3: It can be observed from proof of TA-PI algorithm, without using the function approximator or state discretization, the theoretical convergence is analyzed exactly via the mathematical induction, and the monotonicity and boundedness property of the iterative value function are derived.

V. Simulation

Consider following nonlinear continuous dynamic system

$$
\begin{cases}
\dot{x}_i(t) = 0, & i \in L,\\
\dot{x}_i(t) = -x_i^3(t) + x_i^2(t) + u_i(t), & i \in F.
\end{cases}
$$

It is noted that, the system model in (V) is only used to do simulation. The proposed TA-PI will not use the information of model structure and parameter. Thus, the proposed TA-PI is model-free algorithm. In other words, the proposed TA-PI algorithm can be adapted to any other controllable nonlinear continuous dynamic system without modifying the learning algorithm. This is a big advantage of the model-free learning algorithm.

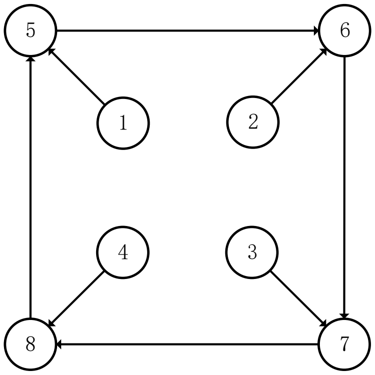

Fig. 1. Directed communication network $\mathcal{G}$.

The communication network is directed and shown in Fig. 1, where the agents $\{1, 2, 3, 4\}$ are leader agent, the agents $\{5, 6, 7, 8\}$ are follower agent.

Choosing the parameters in (3) $Q_i = R_i = 1$. And the finite predefined event-set $\mathbb{S}_i^{\mathbb{D}_i}$ in (8) is

$$\begin{aligned}\mathbb{S}_i^{\mathbb{D}_i} = \{&-1.0, -0.8, -0.6, -0.4, -0.2,\\ &0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}.\end{aligned}$$

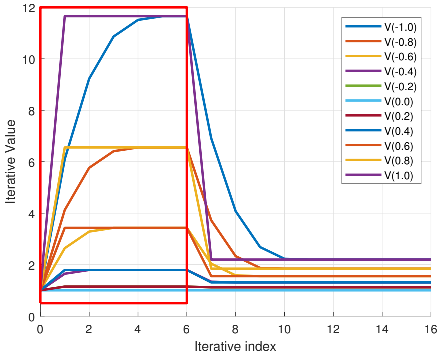

Fig. 2. The iterative value function of each state based on the TA-PI algorithm.

It is noted that, based on the TA approach, the continuous and uncountable state is divided into several segments, which means the state space is finite. Thus, the value function of each state in event-set can be stored in a finite table, which means that there is no need to use the function approximator and state discretization. Therefore, the value of each state in event-set can be calculated exactly. The iterative value function based TA-PI algorithm is given in Fig. 2. It is noted that each agent has the same Q_i and P_i. Thus, all the agents have the same iterative value function curve. Therefore, we only consider one of them. Furthermore, the red rectangle box in Fig. 2 represents the **Initialization** step, in which the value function is monotonically increasing. In the following iteration, the value function is monotonically decreasing. Thus, from Fig. 2, the correctness of Theorem 4.2 is validated.

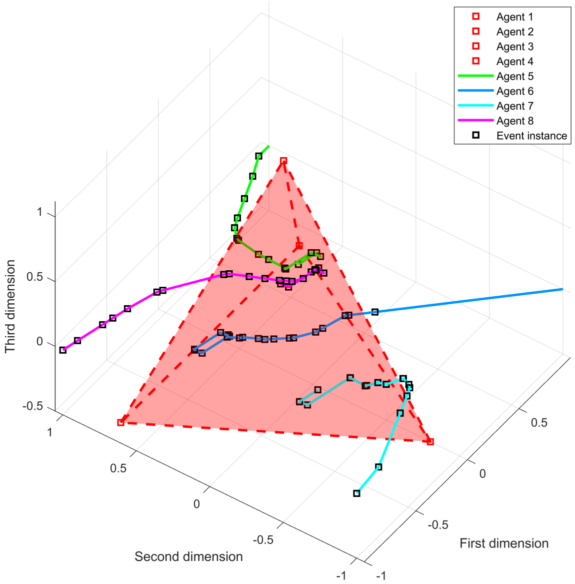

Fig. 3. The trajectory of each agent.

Later, to show that the learned policy via TA-PI could solve the optimal containment control, we would like to verify on three-dimensional space. More specifically, the dynamic is represented as

$$\begin{cases}\dot{x}_{i,k}(t) = 0, & i \in L,\\ \dot{x}_{i,k}(t) = -x_{i,k}^3(t) + x_{i,k}^2(t) + u_{i,k}(t), & i \in F,\end{cases}$$

with $k \in \{1, 2, 3\}$, where $x_{i,k}(t) \in \mathbb{R}$, $u_{i,k}(t) \in \mathbb{R}$.

Then, based on the proposed TA-PI algorithm, the learned event-control strategy is given as

$$u_{i,k}(t) = -\begin{cases}1.0 * sgn(e_{i,k}(t)), 1.0 < |e_{i,k}(t)|,\\ 1.0 * sgn(e_{i,k}(t)), |e_{i,k}(t)| = 1.0,\\ 1.0 * sgn(e_{i,k}(t)), |e_{i,k}(t)| = 0.8,\\ 0.9 * sgn(e_{i,k}(t)), |e_{i,k}(t)| = 0.6,\\ 0.5 * sgn(e_{i,k}(t)), |e_{i,k}(t)| = 0.4,\\ 0.2 * sgn(e_{i,k}(t)), |e_{i,k}(t)| = 0.2,\\ 0.0, 0.0 = |e_{i,k}(t)|,\end{cases}$$

where $e_{i,k}(t) = \sum_{j \in \mathcal{N}_i} a_{ij}(x_{j,k}(t) - x_{i,k}(t))$. The initial configuration of each agent is $x_1 = [0.87, 0.70, -0.50]^T$, $x_2 = [-0.87, 0.70, -0.50]^T$, $x_3 = [0.00, -0.80, -0.50]^T$, $x_4 = [0.00, 0.20, 1.12]^T$, $x_5 = [1, 1, 0]^T$, $x_6 = [1, -1, 0]^T$, $x_7 = [-1, -1, 0]^T$, $x_8 = [-1, 1, 0]^T$, where $x_i = [x_{i,1}, x_{i,2}, x_{i,3}]^T \in \mathbb{R}^3$.

The simulation results are given in Fig. 3. It is easily observed that the followers are driven into the three-dimensional convex hull spanned by the leaders by the learned control strategy without knowing any model information of each agent.

VI. Conclusions

In this paper, the optimal containment control of unknown nonlinear MASs has been investigated via the developed model-free TA-PI algorithm. To reduce the computational burden of traditional PI algorithm, the TA technique is employed, in which the steps of policy improvement and policy evaluation is need to be executed for a finite event-state. Moreover, with the introduced event-set, the control updating can be reduced greatly, enabling a further computational resources saving. Furthermore, without employing the function approximator and state discretization, the convergence of the proposed TA-PI algorithm can be proved exactly. Finally, the feasibility and effectiveness of the proposed algorithm have been verified by numerical simulations.

References

[1] M. Mazouchi, F. Tatari, B. Kiumarsi, and H. Modares, "Fully heterogeneous containment control of a network of leader–follower systems," *IEEE Transactions on Automatic Control*, vol. 67, no. 11, pp. 6187–6194, Nov. 2021.

[2] C. Yuan, P. Stegagno, H. He, and W. Ren, "Cooperative adaptive containment control with parameter convergence via cooperative finite-time excitation," *IEEE Transactions on Automatic Control*, vol. 66, no. 11, pp. 5612–5618, Nov. 2021.

[3] X. Shi, Y. Li, Q. Liu, K. Lin, and S. Chen, "A fully distributed adaptive event-triggered control for output regulation of multi-agent systems with directed network," *Information Sciences*, vol. 626, pp. 60–74, May 2023.

[4] X. Shi, Y. Li, Y. Yang, B. Sun, and Y. Li, "Rotating consensus for double-integrator multi-agent systems with communication delay," *ISA Transactions*, vol. 128, pp. 207–216, Sept. 2022.

[5] F. Yan, X. Liu, and T. Feng, "Distributed minimum-energy containment control of continuous-time multi-agent systems by inverse optimal control," *IEEE/CAA Journal of Automatica Sinica*, 2022, doi:10.1109/JAS.2022.106067.

[6] Y. Yang, H. Modares, D. C. Wunsch, and Y. Yin, "Optimal containment control of unknown heterogeneous systems with active leaders," *IEEE Transactions on Control Systems Technology*, vol. 27, no. 3, pp. 1228–1236, Mar. 2018.

[7] Z. Peng, J. Hu, and B. K. Ghosh, "Data-driven containment control of discrete-time multi-agent systems via value iteration," *Science China Information Sciences*, vol. 63(189205), Apr. 2020.

[8] T. Li, W. Bai, Q. Liu, Y. Long, and C. P. Chen, "Distributed fault-tolerant containment control protocols for the discrete-time multiagent systems via reinforcement learning method," *IEEE Transactions on Neural Networks and Learning Systems*, 2021, doi:10.1109/TNNLS.2021.3121403.

[9] B. Luo, Y. Yang, H. Wu, and T. Huang, "Balancing value iteration and policy iteration for discrete-time control," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 50, no. 11, pp. 3948–3958, 2019.

[10] C. Li, J. Ding, F. L. Lewis, and T. Chai, "A novel adaptive dynamic programming based on tracking error for nonlinear discrete-time systems," *Automatica*, vol. 129(109687), 2021.

[11] J. Tsitsiklis and B. Van Roy, "An analysis of temporal-difference learning with function approximation," *IEEE Transactions on Automatic Control*, vol. 42, no. 5, pp. 674–690, May. 1997.

[12] Y. Li and X. Wu, "A unified approach to time-aggregated Markov decision processes," *Automatica*, vol. 67, pp. 77–84, May. 2016.

[13] Y. Xu and X. Cao, "Lebesgue-sampling-based optimal control problems with time aggregation," *IEEE Transactions on Automatic Control*, vol. 56, no. 5, pp. 1097–1109, May. 2010.

[14] Y. Wan and X. Cao, "The control of a two-level Markov decision process by time aggregation," *Automatica*, vol. 42, no. 3, pp. 393–403, 2006.

[15] X. Cao, Z. Ren, S. Bhatnagar, M. Fu, and S. Marcus, "A time aggregation approach to Markov decision processes," *Automatica*, vol. 38, no. 6, pp. 929–943, Jun. 2002.

[16] W. Jiang, G. Wen, Z. Peng, T. Huang, and A. Rahmani, "Fully distributed formation-containment control of heterogeneous linear multiagent systems," *IEEE Transactions on Automatic Control*, vol. 64, no. 9, pp. 3889–3896, Sept. 2018.

[17] J. Xu, P. Lin, L. Cheng, and H. Dong, "Containment control with input and velocity constraints," *Automatica*, vol. 142(110417), Aug. 2022.

[18] Y. Yang and W. Hu, "Containment control of double-integrator multi-agent systems with time-varying delays," *IEEE Transactions on Network Science and Engineering*, vol. 9, no. 2, pp. 457–466, Feb. 2021.

[19] J. Mei, W. Ren, and G. Ma, "Distributed containment control for Lagrangian networks with parametric uncertainties under a directed graph," *Automatica*, vol. 48, no. 4, pp. 653–659, Apr. 2012.

[20] N. Li, X. Li, J. Peng, and Z. Q. Xu, "Stochastic linear quadratic optimal control problem: A reinforcement learning method," *IEEE Transactions on Automatic Control*, vol. 67, no. 9, pp. 5009–5016, Sept. 2022.