Optimized Formation Control Using Simplified Reinforcement Learning for a Class of Multiagent Systems With Unknown Dynamics

Guoxing Wen, C. L. Philip Chen, *Fellow, IEEE*, and Bin Li

***Abstract*—The article proposes an optimized leader-follower formation control using a simplified reinforcement learning (RL) of identifier–critic–actor architecture for a class of nonlinear multiagent systems. In general, optimal control is expected to be obtained by solving Hamilton–Jacobi–Bellman (HJB) equation, but the equation associated with a nonlinear system is difficult to solve by analytical method. Although the difficulty can be effectively overcome by the RL strategy, the existing RL algorithms are very complex because their updating laws are obtained by carrying out gradient descent algorithm to square of the approximated HJB equation (Bellman residual error). For a multiagent system, due to the state coupling problem, these methods will become difficult implementation. In the proposed optimized scheme, the RL updating laws are derived from negative gradient of a simple positive function, which is the equivalence to HJB equation; therefore, the control algorithm is significantly simple. Furthermore, in order to solve the problem of unknown system dynamics, an adaptive identifier is integrated into the control. Finally, the theory and simulation demonstrate that the optimized formation scheme can guarantee the desired control performance.**

***Index Terms*—Identifier–critic–actor architecture, Lyapounov function, neural networks (NNs), optimized formation control, simplified reinforcement learning (RL).**

Manuscript received May 8, 2019; revised August 1, 2019; accepted September 30, 2019. Date of publication October 15, 2019; date of current version April 30, 2020. This work was supported in part by the Shandong Provincial Natural Science Foundation, China under Grant ZR2018MF015, in part by the National Natural Science Foundation of China under Grant 61751202, Grant 61751205, Grant 61572540, Grant U1813203, and Grant U1801262, in part by the Macau Science and Technology Development Fund (FDCT) under Grant 079/2017/A2, Grant 024/2015/AMJ, and Grant 0119/2018/A3, in part by the Shandong Province Key Research and Development Program, China under Grant 2018GGX103054. *(Corresponding author: Guoxing Wen.)*

G. Wen is with the College of Science, Binzhou University, Binzhou 256600, China (e-mail: wengx_sd@hotmail.com).

C. L. P. Chen is with the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510641, China, also with the Navigation College, Dalian Maritime University, Dalian 116026, China, and also with the Faculty of Science and Technology, University of Macau, Macau 99999, China (e-mail: philip.chen@ieee.org).

B. Li is with the School of Mathematics and Statistics, Qilu University of Technology (Shandong Academy of Sciences), Jinan 250353, China (e-mail: ribbenlee@126.com).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIE.2019.2946545

I. Introduction

OPTIMAL control refers to realize control objective by minimizing the performance index, which balances both performance and resource. Since energy saving and environmental protection have become the topic of times developing, it is very necessary to consider optimization as a basic principle in modern control design. As we all know, optimal control is expected to be derived from solving Hamilton–Jacobi–Bellman (HJB) equation. However, analytical solution of the equation is obtained difficultly due to its strong nonlinearity and intractability. For overcoming the difficulty, reinforcement learning (RL) [1] can be considered as an effective way. In general, RL is performed by the critic–actor architecture, where critic evaluates control performance and returns the evaluation to actor, and the actor executes the control behavior. However, one of the disadvantages for most RL algorithms is that the complete dynamic knowledge is required [2]. In fact, the practical applications often desire that the control is able to deal with uncertain or unknown dynamic systems.

Since neural networks (NNs) and fuzzy logic systems (FLSs) were proven to have the excellent approximating and learning abilities, they are widely applied to the nonlinear control for solving the problem of unknown or uncertain dynamics, such as backstepping control, observer control, and dead-zone control [3]–[5]. Based on NNs or FLSs, many adaptive RL methods for optimizing control are reported in the literature [6]–[11]. Wen *et al.* [6] propose a new control technique named optimized backstepping (OB) to strict feedback system by melting optimization into backstepping. The basic design idea is that actual control and all virtual control are designed to be the optimized solutions of corresponding backstepping steps, such that the whole control can be optimized. But it requires complete knowledge of the system dynamics. In [7], the OB technique is applied to surface vessel control. In [8]–[11], for solving unknown dynamic problem in optimizing, the NN- or FLS-based adaptive identifier or observer is constructed to estimate the unknown dynamic.

Multiagent systems are composed of multiple interacting intelligent individuals, they can exceed the capability of multiple single agents, and are meeting the development requirement of modern industry, civilian, and military. In multiagent community, formation control is one of the most fundamental and important research topics due to their wide applications. Formation control aims to steer multiple intelligent agents arriving prescribed constraints on their states by only using the neighbors' information. In general, formation control is designed based on leader-follower [12], virtual structure [13], behavior-based [14], and potential function [15] strategies, where leader-follower has always been the most popular formation strategy because of the simplicity and scalability.

However, due to the state coupling problem, the multiagent control becomes much more complex and challenging than single-system control, whether it be control design or stability analysis. Noteworthy, several optimal formation approaches are developed recently [16]–[19]. Liu and Geng [16] propose an optimal formation control using finite-time strategy to the two agent case, and ref. [17] extends the formation approach to the multivehicle system. In [18], the optimal dynamic formation control is addressed for mobile leader-follower networks by maximizing a given objective function. Wen *et al.* [19] propose an optimized leader-follower formation approach by using FLS-based RL in identifier–critic–actor architecture.

Nevertheless, almost all optimal algorithms using RL are very complex in both critic and actor training. Moreover, the assumption of persistence of excitation is required. Therefore, these optimal control approaches are difficult to be applied in practical engineering. Motivated by the above discussion, this article proposes an NN-based optimized leader-follower formation control by using the simplified RL algorithm of identifier–critic–actor architecture for a class of nonlinear multiagent systems. The main contributions are listed in the following:

i) The proposed optimized approach is significantly simple because the training laws of RL algorithms are derived from the negative gradient of a simple positive function rather than the square of Bellman residual error.

ii) The proposed optimized approach can remove the assumption of persistence of excitation, which is required in most RL-based optimal control methods, and therefore it can be easily applied to the practical engineering.

iii) An identifier technique for estimating the unknown dynamic functions is proposed by integrating adaptive NN approximator into optimized control design. The technique can more effectively combine with optimal control than the existing identifier methods.

II. Preliminaries

A. Algebraic Graph Theory

In this research, the multiagent communication network is depicted by an undirected connected graph $G = (V, \Psi, A)$, where $V = \{\nu_1, \nu_2, \ldots, \nu_n\}$ is the node set, $\Psi \subseteq V \times V$ is the edge set, and $A = [a_{ij}]$ is the adjacency matrix. If there is an information flow from node ν_j to node ν_i, then the edge $\bar{\nu}_{ij} = (\nu_i, \nu_j) \in \Psi$, node ν_j is called as a neighbor of node ν_i, and the neighbor set is denoted by $N_i = \{\nu_j | (\nu_i, \nu_j) \in \Psi\}$. When $\bar{\nu}_{ij} \in \Psi$, the corresponding adjacency element $a_{ij} = 1$, and when $\bar{\nu}_{ij} \notin \Psi$, $a_{ij} = 0$. The graph G is said to be undirected if only if $a_{ij} = a_{ji}$. An undirected graph is called as connected if there is a path for any pair of distinct nodes, where the path is an edge sequence in the form $(\nu_i, \nu_{i_1}), (\nu_{i_1}, \nu_{i_2}), \ldots, (\nu_{i_l}, \nu_j)$.

With respect to the graph G, the Laplacian matrix is $L = D - A \in R^{n\times n}$, where $D = diag\{\sum_{j=1}^{n} a_{1j}, \ldots, \sum_{j=1}^{n} a_{nj}\}$. The communication matrix for agents and leader is $B = diag\{b_1, b_2, \ldots, b_n\}$, where b_i, $i = 1, \ldots, n$, is the communication weight between agent i and leader. It is assumed that at least one of the agents connects with leader, which implies $b_1 + b_2 + \cdots + b_n > 0$.

Lemma 1: [20] The Laplacian matrix with respect to an undirected connected graph is irreducible.

Lemma 2: [20] If the Laplacian matrix L is irreducible, then $L + B$ are a positive definite matrix.

B. Neural Networks (NNs)

NNs had been proven to have excellent approximation ability, they can approximate a continuous function $f(x) : R^n \to R^m$ by the following form over a compact set $\Omega_x \subset R^n$:

$$f_{NN}(x) = W^T S(x) \tag{1}$$

where $W \in R^{p\times m}$ is the NN weight with neuron number p; $S(x) = [s_1(x), \ldots, s_p(x)]^T \in R^p$, where $s_i(x) = \exp(-(x - \tau_i)^T (x - \tau_i)/2\mu_i^2) \in R$ with the centers $\tau_i = [\tau_{i1}, \tau_{i2}, \ldots, \tau_{in}]^T \in R^n$ and width $\mu_i \in R$.

Define the ideal weight matrix W^* as

$$W^* := \arg \min_{W\in R^{p\times m}} \left\{ \sup_{x\in\Omega_x} \left\| f(x) - W^T S(x) \right\| \right\} \tag{2}$$

which is an "artificial" quantity only for analyzing. Then, the function $f(x)$ can be reexpressed as

$$f(x) = W^{*T} S(x) + \varepsilon(x) \tag{3}$$

where $\varepsilon(x) \in R^m$ is the approximation error to satisfy $\|\varepsilon(x)\| \leq \delta$, δ is a constant.

III. Main Results

A. Problem Formulation

The nonlinear multiagent system is described in the following:

$$\dot{x}_i(t) = f_i(x_i(t)) + u_i, i = 1, 2, \ldots, n \tag{4}$$

where $x_i(t) = [x_{i1}(t), \ldots, x_{im}(t)]^T \in R^m$, $u_i = [u_{i1}, \ldots, u_{im}]^T \in R^m$ are the system state and control, respectively; $f_i(\cdot) \in R^m$ is the unknown continuous nonlinear dynamic function. These terms $f_i(\cdot) + u_i$ are assumed Lipschitz continuous so that (4) has unique solution for bounded initial states. The multiagent system (4) is stabilizable, i.e., there exist the continuous control functions $u_i, i = 1, \ldots, n$, such that the system is asymptotically stable [2].

The desired trajectory of formation movement is denoted by a time variable $x_d(t) \in R^m$, where it and its derivative $\dot{x}_d(t)$ are assumed bounded. The tracking errors are defined as

$$z_i(t) = x_i(t) - x_d(t) - \eta_i, i = 1, 2, \ldots, n \tag{5}$$

where $\eta_i = [\eta_{i1}, \eta_{i2}, \ldots, \eta_{im}]^T \in R^m$ is the relative position between agent i and leader, which depicts the formation pattern.

Definition 1: [12] The multiagent formation control is said to be achieved if the following equations are satisfied:

$$\lim_{t\to\infty} \|z_i(t)\| = \lim_{t\to\infty} \|x_i(t) - x_d(t) - \eta_i\| = 0$$
$$i = 1, \ldots, n.$$

From (4), the following error dynamics can be generated:

$$\dot{z}_i(t) = f_i(x_i) + u_i - \dot{x}_d(t), i = 1, \ldots, n. \tag{6}$$

Define the formation errors as

$$\begin{aligned} e_i(t) = &\sum_{j\in N_i} a_{ij}\left(x_i(t) - \eta_i - x_j(t) + \eta_j\right) \\ &+ b_i\left(x_i(t) - x_d(t) - \eta_i\right), i = 1, \ldots, n. \end{aligned} \tag{7}$$

Using (5), the formation error (7) can be rewritten as

$$e_i(t) = \sum_{j\in N_i} a_{ij}(z_i - z_j) + b_i z_i, i = 1, \ldots, n. \tag{8}$$

Remark 1: According to Lemma 2, the matrix $\tilde{L} = L + B$ is a positive definite matrix. Let $\chi_1, \chi_2, \ldots, \chi_n$ be its eigenvectors associated with the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$, then the following inequality holds [20]:

$$\lambda_{\min}\|e(t)\|^2 \leq z^T(t)\left(\tilde{L}\otimes I_m\right) z(t) \leq \lambda_{\max}\|e(t)\|^2 \tag{9}$$

where $z(t) = [z_1^T(t), \ldots, z_n^T(t)]^T \in R^{nm}$, $e(t) = [e_1^T(t), \ldots, e_n^T(t)]^T \in R^{nm}$, $\lambda_{\min}$ and $\lambda_{\max}$ is the minimal and maximal eigenvalues of matrix $M = (Q^T\Lambda^{-1}Q)\otimes I_m$ with $Q = [\chi_1, \chi_2, \ldots, \chi_n] \in R^{n\times n}$ and $\Lambda = diag\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$

The time derivative of formation error $e_i(t)$ along (6) is

$$\begin{aligned} \dot{e}_i(t) = &c_i f_i(x_i) + c_i u_i - b_i\dot{x}_d(t) - \sum_{j\in N_i} a_{ij} f_j(x_j) \\ &- \sum_{j\in N_i} a_{ij} u_j, i = 1, \ldots, n \end{aligned} \tag{10}$$

where $c_i = \sum_{j\in N_i} a_{ij} + b_i$.

Define the performance index for overall multiagent (4) as

$$J(e(0)) = \int_0^\infty r\left(e(\tau), u(e)\right) d\tau \tag{11}$$

where $r(e, u) = e^T e + u^T u \in R$ with $e = [e_1^T, \ldots, e_n^T]^T \in R^{nm}$ and $u = [u_1^T, \ldots, u_n^T]^T \in R^{nm}$.

Definition 2: [21] The multiagent control protocols u_i, $i = 1, \ldots, n$, for (4) is said to be admissible on the set Ω denoted by $u_i \in \Psi(\Omega)$, if u_i is continuous with $u_i(0) = 0$, u_i stabilizes (4) on Ω, and $J(e(0))$ is finite.

The optimal formation problem is to find the control protocols $u_i \in \Psi(\Omega), i = 1, \ldots, n$, for the multiagent system (4) such that the performance index (11) is minimized.

The Control Objective: Design the optimized formation control on the strength of a simplified RL algorithm for the nonlinear multiagent system (4), such that i) all error signals are semi-globally uniformly ultimately bounded (SGUUB); ii) the tracking errors $z_i(t)$, $i = 1, \ldots, n$, convergence to desired accuracy.

In order to achieve the control objective, the performance index function is defined on the basis of (11) as

$$J(e) = \int_t^\infty r\left(e(\tau), u(e)\right) d\tau \tag{12}$$

then the distributed performance function can be generated as

$$J_i(e_i) = \int_t^\infty r_i\left(e_i(\tau), u_i(e_i)\right) d\tau, i = 1, \ldots, n. \tag{13}$$

where $r_i(e_i, u_i) = e_i^T e_i + u_i^T u_i \in R$. Obviously, there is the following relationship for both index functions (12) and (13):

$$J(e) = \sum_{i=1}^n J_i(e_i). \tag{14}$$

Let $u^* = [u_1^{*T}, \ldots, u_n^{*T}]^T \in R^{nm}$ be the optimal formation control; substituting u^* into (12), the optimal performance index function $J^*(e) \in R$ is yielded as

$$J^*(e) = \min_{u_{i=1,\ldots,n}\in\Psi(\Omega)}\left\{\int_t^\infty r(e, u) d\tau\right\} = \int_t^\infty r\left(e, u^*\right) d\tau. \tag{15}$$

The optimal distributed performance index functions are

$$\begin{aligned} J_i^*(e_i) = &\min_{u_i\in\Psi(\Omega)}\left\{\int_t^\infty r_i\left(e_i, u_i\right) d\tau\right\} = \int_t^\infty r_i\left(e_i, u_i^*\right) d\tau, \\ &i = 1, \ldots, n \end{aligned} \tag{16}$$

where Ω is a compact set containing origin.

According to the principle of optimality, calculating time derivative on both sides of (16), the following distributed HJB equation can be yielded:

$$\begin{aligned} &H_i\left(e_i, u_i^*, \frac{dJ_i^*}{de_i}\right) \\ &= r_i\left(e_i, u_i^*\right) + \frac{dJ_i^*(e_i)}{dt} = \|e_i\|^2 + \|u_i^*\|^2 + \frac{dJ_i^*(e_i)}{de_i^T} \\ &\quad\times\left(c_i f_i(x_i) + c_i u_i^* - b_i\dot{x}_d(t) - \sum_{j\in N_i} a_{ij} f_j(x_j) - \sum_{j\in N_i} a_{ij} u_j^*\right) \\ &= 0, \\ &i = 1, \ldots, n \end{aligned} \tag{17}$$

where $\frac{dJ_i^*(e_i)}{de_i} \in R^m$ is the gradient of J_i^* with respect to e_i.

Because the right-hand side of (16) is existent and unique, which implies that u_i^* is the unique solution of HJB equation (17), the optimal distributed control u_i^* can be obtained by solving $\partial H_i(e_i, u_i^*, \frac{dJ_i^*}{de_i})/\partial u_i^* = 0$ [22]

$$u_i^* = -\frac{c_i}{2}\frac{dJ_i^*(e_i)}{de_i}, i = 1, \ldots, n. \tag{18}$$

In the optimal distributed control (18), the term $\frac{dJ_i^*(e_i)}{de_i}$ is required. Substituting (18) into (17) yields

$$\begin{aligned} &\|e_i\|^2 + \frac{dJ_i^*(e_i)}{de_i^T} \\ &\quad\times\left(c_i f_i(x_i) - b_i\dot{x}_d - \sum_{j\in N_i} a_{ij} f_j(x_j) - \sum_{j\in N_i} a_{ij} u_j^*\right) \\ &\quad - \frac{c_i^2}{4}\frac{dJ_i^*(e_i)}{de_i^T}\frac{dJ_i^*(e_i)}{de_i} = 0 \\ &i = 1, \ldots, n. \end{aligned} \tag{19}$$

The term $\frac{dJ_i^*(e_i)}{de_i}$ is expected to be obtained by solving (19). However, analytical solution of the equation is obtained difficultly due to the strong nonlinearities. Therefore, the RL-based approximation strategy is usually considered for achieving the optimization.

B. Simplified Reinforcement Learning Design

In order to realize the optimized leader-follower formation control, the gradient $\frac{dJ_i^*(e_i)}{de_i}$ is segmented as

$$\frac{dJ_i^*(e_i)}{de_i} = 2\frac{\gamma_i}{c_i}e_i(t) + \frac{2}{c_i}f_i(x_i) + \frac{1}{c_i}J_i^o(x_i,e_i),$$
$$i = 1,\ldots,n \tag{20}$$

where $\gamma_i > 0$ is a designed constant, $J_i^o(x_i,e_i) = -2\gamma_i e_i(t) - 2f_i(x_i) + c_i\frac{dJ_i^*(e_i)}{de_i}$. Substituting (20) into (18) has

$$u_i^* = -\gamma_i e_i(t) - f_i(x_i) - \frac{1}{2}J_i^o(x_i,e_i), i = 1,\ldots,n. \tag{21}$$

Since the unknown terms $f_i(x_i)$ and $J_i^o(x_i,e_i)$ are continuous, they can be approximated by NNs based on the compact set Ω in the following form:

$$f_i(x_i) = W_{fi}^{*T}S_{fi}(x_i) + \varepsilon_{fi}(x_i) \tag{22}$$

$$J_i^o(x_i,e_i) = W_i^{*T}S_i(x_i,e_i) + \varepsilon_i(x_i,e_i)$$
$$i = 1,\ldots,n \tag{23}$$

where $W_{fi}^* \in R^{p_1\times m}$, $W_i^* \in R^{p_2\times m}$ are the ideal NN weight matrices; $S_{fi}(x_i) \in R^{p_1}$, $S_i(x_i,e_i) \in R^{p_2}$ are the basis function vectors; p_1, p_2 are the neuron numbers; $\varepsilon_{fi} \in R^{p_1}$ and $\varepsilon_i \in R^{p_2}$ are the approximation errors, which are bounded by constants δ_{fi} and δ_i, respectively, i.e., $\|\varepsilon_{fi}\| \leq \delta_{fi}$, $\|\varepsilon_i\| \leq \delta_i$.

Inserting (22) and (23) into (20), (21), the following results are obtained:

$$\frac{dJ_i^*(e_i)}{de_i} = 2\frac{\gamma_i}{c_i}e_i(t) + \frac{2}{c_i}W_{fi}^{*T}S_{fi}(x_i) + \frac{1}{c_i}W_i^{*T}S_i(x_i,e_i)$$
$$+ \frac{2\varepsilon_{fi}}{c_i} + \frac{\varepsilon_i}{c_i} \tag{24}$$

$$u_i^* = -\gamma_i e_i(t) - W_{fi}^{*T}S_{fi}(x_i) - \frac{1}{2}W_i^{*T}S_i(x_i,e_i)$$
$$-\varepsilon_{fi} - \frac{1}{2}\varepsilon_i, i = 1,\ldots,n. \tag{25}$$

However, the optimal control (25) is unavailable because the ideal weight W_{fi}^* and W_i^* are unknown. For obtaining the available control, the identifier, critic, and actor NNs are constructed based on (22), (24), and (25).

The following identifier is used to identify the unknown dynamic function:

$$\hat{f}_i(x_i) = \hat{W}_{fi}^T(t)S_{fi}(x_i), i = 1,\ldots,n \tag{26}$$

where $\hat{W}_{fi}(t) \in R^{p_1\times m}$ is the identifier NN weight, and $\hat{f}(x_i)$ is the output. The adaptive updating law for (26) is designed as

$$\dot{\hat{W}}_{fi}(t) = \Gamma_i\left(S_{fi}(x_i)e_i^T - \sigma_i\hat{W}_{fi}(t)\right)$$
$$i = 1,\ldots,n \tag{27}$$

where $\Gamma_i \in R^{p_1\times p_1}$ is a positive definite matrix, and σ_i is a positive design constant.

The following critic is used to evaluate the control performance:

$$\frac{d\hat{J}_i^*(e_i)}{de_i} = 2\frac{\gamma_i}{c_i}e_i(t) + \frac{2}{c_i}\hat{W}_{fi}^T(t)S_{fi}(x_i) + \frac{1}{c_i}\hat{W}_{ci}^T(t)S_i(x_i,e_i)$$
$$i = 1,\ldots,n \tag{28}$$

where $\frac{d\hat{J}_i^*(e_i)}{de_i}$ is the output, $\hat{W}_{ci}(t) \in R^{p_2\times m}$ is the critic NN weight. The critic weight updating law is designed as

$$\dot{\hat{W}}_{ci}(t) = -\kappa_{ci}S_i(x_i,e_i)S_i^T(x_i,e_i)\hat{W}_{ci}(t)$$
$$i = 1,\ldots,n \tag{29}$$

where $\kappa_{ci} > 0$ is the critic design parameter.

The following actor is used to implement the control behavior:

$$u_i = -\gamma_i e_i(t) - \hat{W}_{fi}^T(t)S_{fi}(x_i) - \frac{1}{2}\hat{W}_{ai}^T(t)S_i(x_i,e_i)$$
$$i = 1,\ldots,n \tag{30}$$

where $\hat{W}_{ai}(t) \in R^{p_2\times m}$ is the actor NN weight. The actor updating law is designed as

$$\dot{\hat{W}}_{ai}(t) = -S_i(x_i,e_i)S_i^T(x_i,e_i)\left(\kappa_{ai}\left(\hat{W}_{ai}(t) - \hat{W}_{ci}(t)\right)\right.$$
$$\left. + \kappa_{ci}\hat{W}_{ci}(t)\right), i = 1,\ldots,n \tag{31}$$

where $\kappa_{ai} > 0$ is the actor design parameter.

Remark 2: Substituting (28), and (30) into (17), the approximated HJB equation is yielded as

$$\begin{aligned}
&H_i\left(e_i,u_i,\frac{d\hat{J}_i^*}{de_i}\right)\\
&= \|e_i\|^2 + \left\|\gamma_i e_i(t) + \hat{W}_{fi}^T(t)S_{fi}(x_i) + \frac{1}{2}\right.\\
&\left.\times\hat{W}_{ai}^T(t)S_i(x_i,e_i)\right\|^2 - \left(2\frac{\gamma_i}{c_i}e_i(t) + \frac{2}{c_i}\hat{W}_{fi}^T(t)S_{fi}(x_i)\right.\\
&\left.+\frac{1}{c_i}\hat{W}_{ci}^T(t)S_i(x_i,e_i)\right)^T\left(\gamma_i c_i e_i(t) + \frac{c_i}{2}\hat{W}_{ai}^T(t)S_i(x_i,e_i)\right.\\
&+ c_i\hat{W}_{fi}^T(t)S_{fi}(x_i) - c_i f_i(x_i)\\
&\left.+ b_i\dot{x}_d(t) + \sum_{j\in N_i}a_{ij}f_j(x_j) + \sum_{j\in N_i}a_{ij}u_j^*\right)\\
&i = 1,\ldots,n.
\end{aligned} \tag{32}$$

From the previous analysis in Sections III-A and III-B, the formation control (30) is expected to satisfy $H_i(e_i,u_i,\frac{d\hat{J}_i^*}{de_i}) \to 0$. If $H_i(e_i,u_i,\frac{d\hat{J}_i^*}{de_i}) = 0$ holds, since HJB equation has the unique solution [22], then there is the fact

$$\frac{\partial H_i(e_i, u_i, \frac{d\hat{J}_i^*}{de_i})}{\partial \hat{W}_{ai}(t)} = \frac{1}{2} S_i(x_i, e_i) S_i^T(x_i, e_i) \left(\hat{W}_{ai}(t) - \hat{W}_{ci}(t)\right) = 0_{p_2 \times m} \in R^{p_2 \times m}. \quad (33)$$

Define a positive function as $\Psi(t) = Tr((\hat{W}_{ai}(t) - \hat{W}_{ci}(t))^T(\hat{W}_{ai}(t) - \hat{W}_{ci}(t)))$; obviously, (33) is equivalent to $\Psi(t) = 0$. For achieving the condition (33), then the updating laws (29) and (31) are derived from the negative gradient of the positive function $\Psi(t)$. The mathematical analysis is given in the following.

In accordance with the fact $\frac{\partial \Psi(t)}{\partial \hat{W}_{ai}(t)} = -\frac{\partial \Psi(t)}{\partial \hat{W}_{ci}(t)} = 2(\hat{W}_{ai}(t) - \hat{W}_{ci}(t))$, calculating the time derivative of $\Psi(t)$ along (29) and (31) has

$$\begin{aligned}\dot{\Psi}(t) &= Tr\left(\frac{\partial \Psi(t)}{\partial \hat{W}_{ai}^T(t)} \dot{\hat{W}}_{ai}(t) + \frac{\partial \Psi(t)}{\partial \hat{W}_{ci}^T(t)} \dot{\hat{W}}_{ci}(t)\right) \\ &= Tr\left(-\frac{\partial \Psi(t)}{\partial \hat{W}_{ai}^T(t)} S_i(x_i, e_i) S_i^T(x_i, e_i)\right. \\ &\qquad \times \left(\kappa_{ai}\left(\hat{W}_{ai}(t) - \hat{W}_{ci}(t)\right) + \kappa_{ci}\hat{W}_{ci}(t)\right) \\ &\qquad \left. + \kappa_{ci} \frac{\partial \Psi(t)}{\partial \hat{W}_{ai}^T(t)} S_i(x_i, e_i) S_i^T(x_i, e_i) \hat{W}_{ci}(t)\right) \\ &= -\frac{\kappa_{ai}}{2} Tr\left(\frac{\partial \Psi(t)}{\partial \hat{W}_{ai}^T(t)} S_i(x_i, e_i) S_i^T(x_i, e_i) \frac{\partial \Psi(t)}{\partial \hat{W}_{ai}(t)}\right) \\ &\le 0. \end{aligned} \quad (34)$$

The above inequality (34) implies that the RL updating laws (29) and (31) can guarantee that (33) is held finally so that the approximated HJB can converge to zero. The main advantages for the designed idea are that 1) RL is significantly simple in comparison with existing optimal methods, such as [2], [6]–[9], [19], [21]; 2) the requirement of persistence excitation is removed. □

C. Stability Analysis

Lemma 3: [23] Give a continuous function $G(t) \in R$ satisfying $\dot{G}(t) \le -\alpha G(t) + \beta$, where $\alpha, \beta > 0$ are two constants, then there is the following one:

$$G(t) \le e^{-\alpha t} G(0) + \frac{\beta}{\alpha}\left(1 - e^{-\alpha t}\right). \quad (35)$$

Theorem 1: Consider the multiagent system (4) with bounded initial conditions. If the optimized formation control is performed by the simplified RL algorithm of identifier–critic–actor architecture, where identifier, critic, and actor are given by (26), (28), and (30) with the updating laws (27), (29) and (31), respectively. The design parameters γ_i, κ_{ai}, and κ_{ci} are chosen satisfying the following conditions:

$$\gamma_i > 1, \kappa_{ai} > \frac{1}{2}, \kappa_{ai} > \kappa_{ci} > \frac{1}{2}\kappa_{ai}. \quad (36)$$

Then, the following control objective can be achieved by choosing appropriate design parameters:

1) All error signals are SGUUB.

2) The tracking error $z_i(t)$ convergence to desired accuracy.

Proof: 1) Choose the following Lyapunov function candidate:

$$\begin{aligned} V(t) = {} & \frac{1}{2} z^T(t)\left(\tilde{L} \otimes I_m\right) z(t) \\ & + \frac{1}{2}\sum_{i=1}^{n}\left(Tr\left\{\tilde{W}_{fi}^T(t)\Gamma_i^{-1}\tilde{W}_{fi}(t)\right\}\right. \\ & + \frac{1}{2}\sum_{i=1}^{n} Tr\left\{\tilde{W}_{ci}^T(t)\tilde{W}_{ci}(t)\right\} \\ & + \frac{1}{2}\sum_{i=1}^{n} Tr\left\{\tilde{W}_{ai}^T(t)\tilde{W}_{ai}(t)\right\} \end{aligned} \quad (37)$$

where $z = [z_1^T, \ldots, z_n^T]^T$, $\tilde{W}_{fi}(t) = \hat{W}_{fi}(t) - W_{fi}^*$, $\tilde{W}_{ai}(t) = \hat{W}_{ai}(t) - W_i^*$, $\tilde{W}_{ci}(t) = \hat{W}_{ci}(t) - W_i^*$.

Calculating time derivative along (6), (27), (29), and (31) has

$$\begin{aligned} \dot{V}(t) = {} & \sum_{i=1}^{n} e_i^T(t)\left(f_i(x_i) - \dot{x}_d(t) + u_i\right) \\ & + \sum_{i=1}^{n} Tr\left\{\tilde{W}_{fi}^T(t)\left(S_{fi}(x_i)e_i - \sigma_i\hat{W}_{fi}(t)\right)\right\} \\ & - \sum_{i=1}^{n}\kappa_{ci} Tr\left\{\tilde{W}_{ci}^T(t)S_i(x_i, e_i)S_i^T(x_i, e_i)\hat{W}_{ci}(t)\right\} \\ & - \sum_{i=1}^{n} Tr\left\{\tilde{W}_{ai}^T(t)S_i(x_i, e_i^T)S_i^T(x_i, e_i)\right. \\ & \qquad \left. \times\left(\kappa_{ai}\left(\hat{W}_{ai}(t) - \hat{W}_{ci}(t)\right) + \kappa_{ci}\hat{W}_{ci}(t)\right)\right\}. \end{aligned} \quad (38)$$

Substituting (22) and (30) into (38) yields

$$\begin{aligned} \dot{V}(t) = {} & \sum_{i=1}^{n} e_i^T(t) \\ & \times\left(-\gamma_i e_i(t) - \tilde{W}_{fi}^T(t)S_{fi}(x_i)\right. \\ & \left. - \frac{1}{2}\hat{W}_{ai}^T(t)S_i(x_i, e_i) - \dot{x}_d(t) + \varepsilon_{fi}\right) \\ & + \sum_{i=1}^{n} Tr\left\{\tilde{W}_{fi}^T(t)\ \left(S_{fi}(x_i)e_i^T\ -\ \sigma_i\hat{W}_{fi}(t)\right)\right\} \\ & - \sum_{i=1}^{n}\kappa_{ci} Tr\left\{\tilde{W}_{ci}^T(t)S_i(x_i, e_i)S_i^T(x_i, e_i)\hat{W}_{ci}(t)\right\} \\ & - \sum_{i=1}^{n} Tr\left\{\tilde{W}_{ai}^T(t)\ S_i(x_i, e_i)S_i^T(x_i, e_i)\right. \\ & \left. \times\left(\kappa_{ai}\left(\hat{W}_{ai}(t) - \hat{W}_{ci}(t)\right)\ + \kappa_{ci}\hat{W}_{ci}(t)\right)\right\}. \end{aligned} \quad (39)$$

According to the property of trace operator $Tr\{ab^T\} = a^T b = b^T a$, where $a, b \in R^n$, equation (39) can become the following one:

$$\begin{aligned}\dot{V}(t) = \sum_{i=1}^{n} \Big(& -\gamma_i \left\| e_i(t) \right\|^2 - \frac{1}{2} e_i^T(t) \hat{W}_{ai}^T(t) S_i(x_i, e_i) \\ & - e_i^T(t) \dot{x}_d(t) + e_i^T(t) \varepsilon_{fi} \Big) \\ & - \sum_{i=1}^{n} Tr\left\{ \sigma_i \tilde{W}_{fi}^T(t) \hat{W}_{fi}(t) \right\} \\ & - \sum_{i=1}^{n} \kappa_{ci} Tr\left\{ \tilde{W}_{ci}^T(t) S_i(x_i, e_i) S_i^T(x_i, e_i) \hat{W}_{ci}(t) \right\} \\ & - \sum_{i=1}^{n} \kappa_{ai} Tr\left\{ \tilde{W}_{ai}^T(t) S_i(x_i, e_i) S_i^T(x_i, e_i) \hat{W}_{ai}(t) \right\} \\ & + \sum_{i=1}^{n} (\kappa_{ai} - \kappa_{ci})\, Tr\Big\{ \tilde{W}_{ai}^T(t) S_i(x_i, e_i) \\ & \times S_i^T(x_i, e_i) \hat{W}_{ci}(t) \Big\}. \end{aligned} \tag{40}$$

According to Cauchy–Schwartz inequality, $(x^T y)^2 \leq \|x\|^2 \|y\|^2$, where $x, y \in R^n$, and Young's inequality, $ab \leq \frac{1}{2} a^2 + \frac{1}{2} b^2$, where $a, b \in R$, there are the following facts:

$$\begin{aligned} -e_i^T(t) \dot{x}_d(t) & \leq \frac{1}{2} \left\| e_i(t) \right\|^2 + \frac{1}{2} \left\| \dot{x}_d \right\|^2 \\ e_i^T(t) \varepsilon_{fi} & \leq \frac{1}{2} \left\| e_i(t) \right\|^2 + \frac{1}{2} \delta_{fi}^2 \\ -\frac{1}{2} e_i^T(t) \hat{W}_{ai}^T(t) S_i(x_i, e_i) & \leq \frac{1}{4} \left\| e_i(t) \right\|^2 \\ & + \frac{1}{4} Tr\left\{ \hat{W}_{ai}^T(t) S_i(x_i, e_i) S_i^T(x_i, e_i) \hat{W}_{ai}(t) \right\}. \end{aligned} \tag{41}$$

Inserting inequaly (41) into (40) has

$$\begin{aligned} & \dot{V}(t) \\ & \leq -\sum_{i=1}^{n} (\gamma_i - 2) \left\| e_i \right\|^2 - \sum_{i=1}^{n} Tr\left\{ \sigma_i \tilde{W}_{fi}^T(t) \hat{W}_{fi}(t) \right\} \\ & - \sum_{i=1}^{n} \kappa_{ci} Tr\left\{ \tilde{W}_{ci}^T(t) S_i(x_i, e_i) S_i^T(x_i, e_i) \hat{W}_{ci}(t) \right\} \\ & - \sum_{i=1}^{n} \kappa_{ai} Tr\left\{ \tilde{W}_{ai}^T(t) S_i(x_i, e_i) S_i^T(x_i, e_i) \hat{W}_{ai}(t) \right\} \\ & + \sum_{i=1}^{n} (\kappa_{ai} - \kappa_{ci})\, Tr\left\{ \tilde{W}_{ai}^T(t) S_i(x_i, e_i) S_i^T(x_i, e_i) \hat{W}_{ci}^T(t) \right\} \\ & + \sum_{i=1}^{n} \frac{1}{4} Tr\left\{ \hat{W}_{ai}^T(t) S_i(x_i, e_i) S_i^T(x_i, e_i) \hat{W}_{ai}(t) \right\} \\ & + \frac{1}{2} \sum_{i=1}^{n} \delta_{fi}^2 + \frac{n}{2} \left\| \dot{x}_d \right\|^2. \end{aligned} \tag{42}$$

Based on the facts $\tilde{W}_{fi,ai,ci}(t) = \hat{W}_{fi,ai,ci}(t) - W_{fi,i,i}^*$, the following equations can be held:

$$\begin{aligned} Tr\left\{ \tilde{W}_{fi}^T(t) \hat{W}_{fi}(t) \right\} & = \frac{1}{2} Tr\left\{ \tilde{W}_{fi}^T(t) \tilde{W}_{fi}(t) \right\} \\ & + \frac{1}{2} Tr\left\{ \hat{W}_{fi}^T(t) \hat{W}_{fi}(t) \right\} - \frac{1}{2} Tr\left\{ W_{fi}^{*T} W_{fi}^* \right\} \end{aligned} \tag{43}$$

$$\begin{aligned} & Tr\left\{ \tilde{W}_{ci}^T(t) S_i(x_i, e_i) S_i^T(x_i, e_i) \hat{W}_{ci}(t) \right\} \\ & = \frac{1}{2} Tr\left\{ \tilde{W}_{ci}^T(t) S_i(x_i, e_i) S_i^T(x_i, e_i) \tilde{W}_{ci}(t) \right\} \\ & + \frac{1}{2} Tr\left\{ \hat{W}_{ci}^T(t) S_i(x_i, e_i) S_i^T(x_i, e_i) \hat{W}_{ci}(t) \right\} \\ & - \frac{1}{2} Tr\left\{ W_i^{*T} S_i(x_i, e_i) S_i^T(x_i, e_i) W_i^* \right\} \end{aligned} \tag{44}$$

$$\begin{aligned} & Tr\left\{ \tilde{W}_{ai}^T(t) S_i(x_i, e_i) S_i^T(x_i, e_i) \hat{W}_{ai}(t) \right\} \\ & = \frac{1}{2} Tr\left\{ \tilde{W}_{ai}^T(t) S_i(x_i, e_i) S_i^T(x_i, e_i) \tilde{W}_{ai}(t) \right\} \\ & + \frac{1}{2} Tr\left\{ \hat{W}_{ai}^T(t) S_i(x_i, e_i) S_i^T(x_i, e_i) \hat{W}_{ai}(t) \right\} \\ & - \frac{1}{2} Tr\left\{ W_i^{*T} S_i(x_i, e_i) S_i^T(x_i, e_i) W_i^* \right\} \end{aligned} \tag{45}$$

and, based on condition (36), the following fact can be directly obtained:

$$\begin{aligned} & (\kappa_{ai} - \kappa_{ci}) Tr\left\{ \tilde{W}_{ai}^T(t) S_i(x_i, e_i) S_i^T(x_i, e_i) \hat{W}_{ci}(t) \right\} \\ & \leq \frac{\kappa_{ai} - \kappa_{ci}}{2} Tr\left\{ \tilde{W}_{ai}^T(t) S_i(x_i, e_i) S_i^T(x_i, e_i) \tilde{W}_{ai}(t) \right\} \\ & + \frac{\kappa_{ai} - \kappa_{ci}}{2} Tr\left\{ \hat{W}_{ci}^T(t) S_i(x_i, e_i) S_i^T(x_i, e_i) \hat{W}_{ci}(t) \right\}. \end{aligned} \tag{46}$$

Substituting (43)–(46) into (42) yields

$$\begin{aligned} & \dot{V}(t) \\ & \leq -\sum_{i=1}^{n} (\gamma_i - 2) \|e_i\|^2 - \sum_{i=1}^{n} Tr\left\{ \frac{\sigma_i}{2\lambda_{\max}^{\Gamma_i^{-1}}} \tilde{W}_{fi}^T(t) \Gamma_i^{-1} \tilde{W}_{fi}(t) \right\} \\ & - \sum_{i=1}^{n} \frac{\kappa_{ci}}{2} Tr\left\{ \tilde{W}_{ci}^T(t) S_i(x_i, e_i) S_i^T(x_i, e_i) \tilde{W}_{ci}(t) \right\} \\ & - \sum_{i=1}^{n} \frac{\kappa_{ci}}{2} Tr\left\{ \tilde{W}_{ai}^T(t) S_i(x_i, e_i) S_i^T(x_i, e_i) \tilde{W}_{ai}(t) \right\} \\ & - \sum_{i=1}^{n} \left(\kappa_{ci} - \frac{\kappa_{ai}}{2} \right) Tr\left\{ \hat{W}_{ci}^T(t) S_i(x_i, e_i) S_i^T(x_i, e_i) \hat{W}_{ci}(t) \right\} \\ & - \sum_{i=1}^{n} \left(\frac{\kappa_{ai}}{2} - \frac{1}{4} \right) Tr\left\{ \hat{W}_{ai}^T(t) S_i(x_i, e_i) S_i^T(x_i, e_i) \hat{W}_{ai}(t) \right\} + \tau(t). \end{aligned} \tag{47}$$

where $\lambda_{\max}^{\Gamma_i^{-1}}$ is the maximal eigenvalue of Γ_i^{-1}, $\tau(t) = \frac{n}{2} \| \dot{x}_d(t) \|^2 + \frac{1}{2} \sum_{i=1}^{n} \delta_{fi}^2 + \frac{1}{2} \sum_{i=1}^{n} \sigma_i Tr\{W_{fi}^{*T} W_{fi}^*\} + \sum_{i=1}^{n} (\frac{\kappa_{ai}}{2} + \frac{\kappa_{ci}}{2}) Tr\{W_i^* S_i(x_i, e_i) S_i^T(x_i, e_i) W_i^*\}$, which can be bounded by a constant β, i.e., $\|\tau(t)\| \leq \beta$, because all of its terms are bounded.

Let $\gamma = \min_{i=1,\ldots,n}\{2(\gamma_i - 2)\}$, $\sigma = \min_{i=1,\ldots,n}\{\frac{\sigma_i}{\lambda_{max}^{\Gamma_i^{-1}}}\}$, $\kappa = \min_{i=1,\ldots,n}\{\kappa_{ci}\lambda_{min}^{s_i}\}$, where $\lambda_{\max}^{\Gamma_i^{-1}}$ is the maximal eigenvalue of matrix Γ_i^{-1}, $\lambda_{min}^{s_i}$ is the minimal eigenvalue of $S_i(x_i, e_i)S_i^T(x_i, e_i)$; based on condition (36), inequality (47) can be rewritten as

$$\begin{aligned}\dot{V}(t) \leq &-\frac{\gamma}{2}\sum_{i=1}^{n}\|e_i(t)\|^2 - \frac{\sigma}{2}\sum_{i=1}^{n}Tr\left\{\tilde{W}_{fi}^T(t)\Gamma_i^{-1}\tilde{W}_{fi}(t)\right\}\\ &-\frac{\kappa}{2}\sum_{i=1}^{n}Tr\left\{\tilde{W}_{ai}^T(t)\tilde{W}_{ai}(t)\right\}\\ &-\frac{\kappa}{2}\sum_{i=1}^{n}Tr\left\{\tilde{W}_{ci}^T(t)\tilde{W}_{ci}(t)\right\} + \beta. \qquad (48)\end{aligned}$$

According to (9) (in Remark 1), (48) becomes the following one:

$$\begin{aligned}\dot{V}(t) \leq &-\frac{\gamma}{2\lambda_{\max}}z^T(t)\left(\tilde{L}\otimes I_m\right)z(t)\\ &-\frac{\sigma}{2}\sum_{i=1}^{n}Tr\left\{\tilde{W}_{fi}^T(t)\Gamma_i^{-1}\tilde{W}_{fi}(t)\right\}\\ &-\frac{\kappa}{2}\sum_{i=1}^{n}Tr\left\{\tilde{W}_{ci}^T(t)\tilde{W}_{ci}(t)\right\}\\ &-\frac{\kappa}{2}\sum_{i=1}^{n}Tr\left\{\tilde{W}_{ai}^T(t)\tilde{W}_{ai}(t)\right\} + \beta. \qquad (49)\end{aligned}$$

Further, let $\alpha = \min\{\frac{\gamma}{\lambda_{\max}}, \sigma, \kappa\}$, (49) becomes the following one:

$$\dot{V}(t) \leq -\alpha V(t) + \beta. \qquad (50)$$

Applying Lemma 3, the following inequality can be held:

$$V(t) \leq e^{-\alpha t}V(0) + \frac{\beta}{\alpha}\left(1 - e^{-\alpha t}\right). \qquad (51)$$

The above inequality means that all error signals $z_i(t)$, $\tilde{W}_{ci}(t)$, $\tilde{W}_{ai}(t)$ $i = 1, \ldots, n$ are SGUUB, and implies that the tracking errors z_i, $i = 1, \ldots, n$, can converge to the desired accuracy by making γ_i, $i = 1, \ldots, n$, large enough. □

IV. Simulation Example

A numerical multiagent formation is carried out by using the proposed optimized control. In this example, the multiagent system is consisted of four agents moving on the two-dimensional (2-D) plane. The multiagent dynamic is described in the following:

$$\begin{aligned}\dot{x}_i(t) = -\alpha_i x_i(t) - &\begin{bmatrix}0.5x_{i1}\cos^2(\beta_i x_{i1})\\ x_{i2} - \sin^2(\beta_i x_{i2})\end{bmatrix} + u_i\\ i = 1, 2, 3, 4 & \qquad (52)\end{aligned}$$

where $\alpha_{i=1,2,3,4} = -0.7, 0.1, -0.5, 0.1$, and $\beta_{i=1,2,3,4} = 0.5, 0.4, -5.5, -11.5$, respectively. The initial positions are $x_{i=1,2,3,4}(0) = [5, 5]^T, [-5, 4]^T, [5, -3]^T, [-4, -5]^T$, respectively.

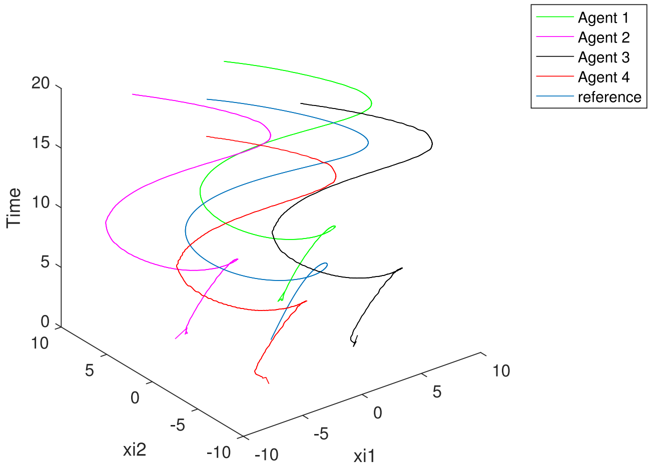

Fig. 1. Multiagent formation performance.

The desired formation trajectory is given as $x_d(t) = [3\cos(0.5t), 2\cos(0.7t)]^T$, and its initial positions are $x_d(0) = [0, 0]^T$. The formation patterns are depicted by $\eta_{i=1,2,3,4} = [4; 4], [-4; 4], [4; -4], [-4; -4]$.

The adjacency matrix describing the communication among agents is

$$A = \begin{bmatrix}0 & 1 & 0 & 0\\ 1 & 0 & 1 & 0\\ 0 & 1 & 0 & 1\\ 0 & 0 & 1 & 0\end{bmatrix}.$$

The communication weights for agents and leader is $B = diag\{1, 0, 0, 0\}$.

For identifier (26) with the initial position $\hat{f}_1(0) = [4, 3]^T$, $\hat{f}_2(0) = [-4, 7]^T$, $\hat{f}_3(0) = [5, -4]^T$, $\hat{f}_4(0) = [-5, -4]^T$, the NN is designed to contain 24 nodes with centers μ_i evenly spaced in the range $[-6, 6]$ with the width $\mu_i = 1$. Then, the design parameters for the updating laws (27) are chosen as $\Gamma_{i=1,2,3,4} = diag\{\underbrace{0.4, \ldots, 0.4}_{24}\}$, $\sigma_{i=1,2,3,4} = 0.6$, and the initial values are $\hat{W}_{f1}(0) = \hat{W}_{f2}(0) = \hat{W}_{f3}(0) = \hat{W}_{f4}(0) = [0.1]_{24\times 2}$.

In accordance with the control conditions (36), the design parameters for the optimized formation control are chosen as $\gamma_{i=1,2,3,4} = 40$. The NNs for critic and actor are designed to have 12 nodes, and the centers μ_i are also evenly spaced in the range $[-6, 6]$. For critic updating laws (29), the design parameters are chosen as $\kappa_{c1,c2,c3,c4} = 4$, and the initial values are $\hat{W}_{c1}(0) = [0.92]_{12\times 2}$, $\hat{W}_{c2}(0) = [0.94]_{12\times 2}$, $\hat{W}_{c3}(0) = [0.95]_{12\times 2}$, $\hat{W}_{c4}(0) = [0.96]_{12\times 2}$. For actor updating laws (31), the design parameters are chosen as $\kappa_{a1,a2,a3,a4} = 6$, and the initial values are $\hat{W}_{a1}(0) = [0.90]_{12\times 2}$, $\hat{W}_{a2}(0) = [0.91]_{12\times 2}$, $\hat{W}_{a3}(0) = [0.90]_{12\times 2}$, $\hat{W}_{a4}(0) = [0.91]_{12\times 2}$.

Figs. 1–8 show the simulation results. Fig. 1 shows that the proposed control approach can achieve the desired multiagent

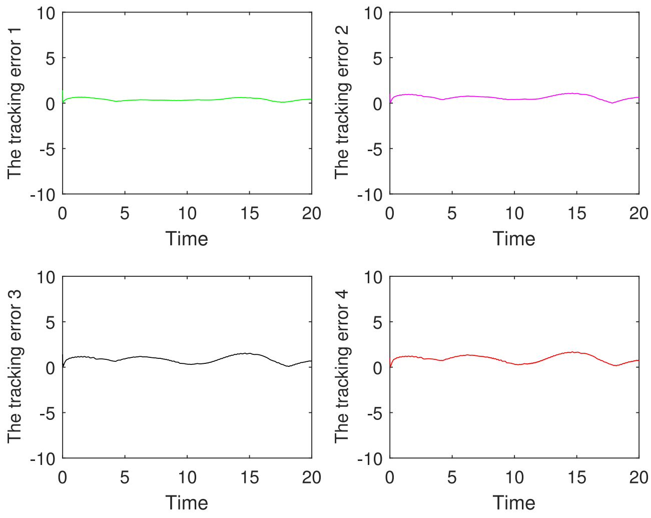

Fig. 2. Tracking error $\|z_i\|$, $i = 1, 2, 3, 4$.

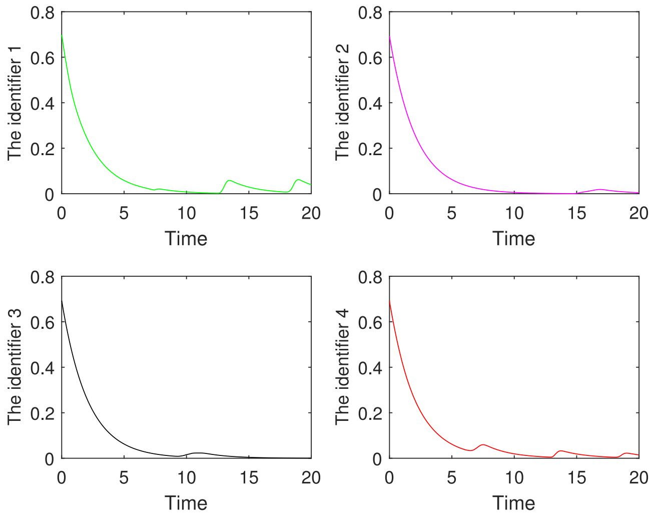

Fig. 3. Norm $\|W_{fi}\|$, $i = 1, 2, 3, 4$, of identifier NN weight.

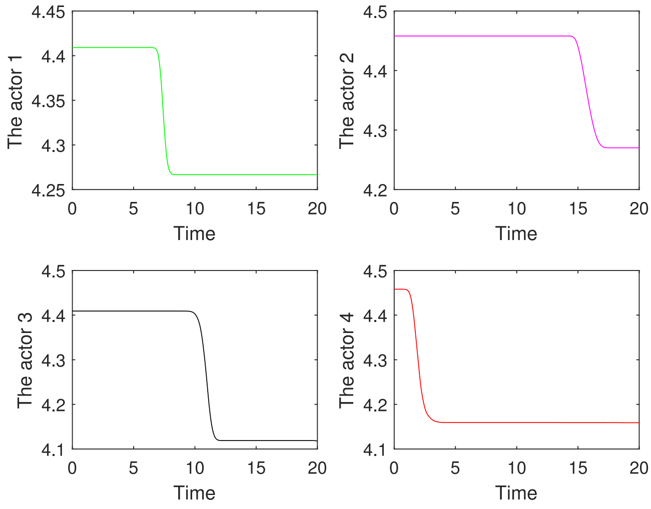

Fig. 4. Norm $\|W_{ai}\|$, $i = 1, 2, 3, 4$, of critic NN weight.

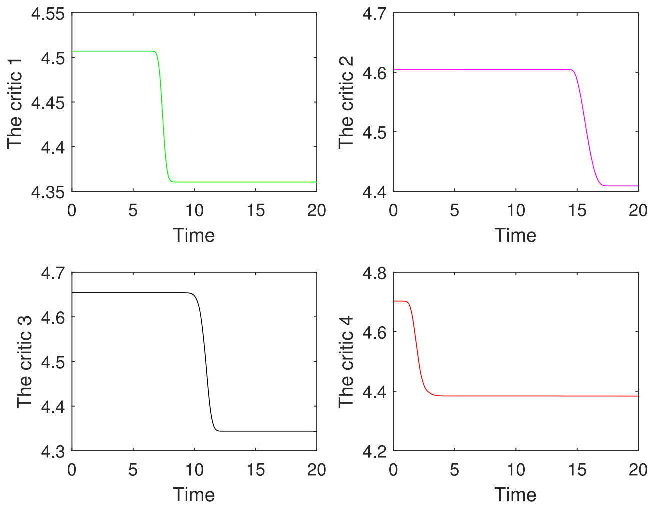

Fig. 5. Norm $\|W_{ci}\|$, $i = 1, 2, 3, 4$, of actor NN weight.

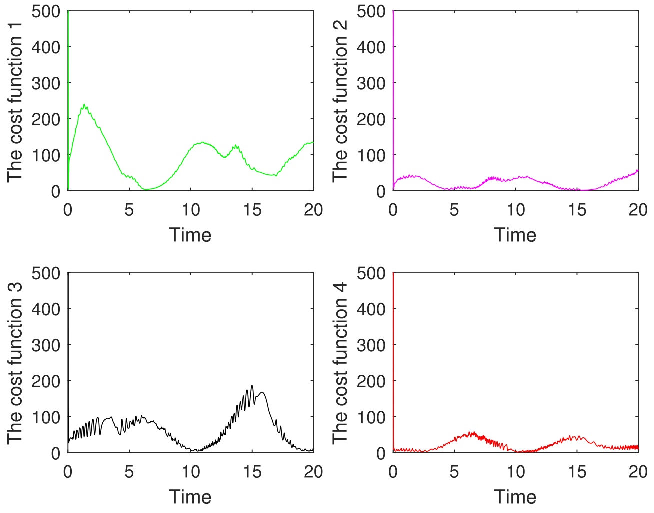

Fig. 6. Cost function $r_i = e_i^T e_i + u_i^T u_i$, $i = 1, 2, 3, 4$.

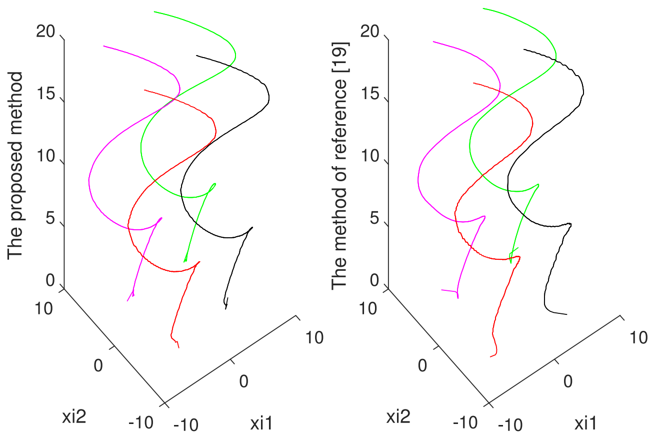

Fig. 7. Formation performances concerning the two methods of the article and ref. [19].

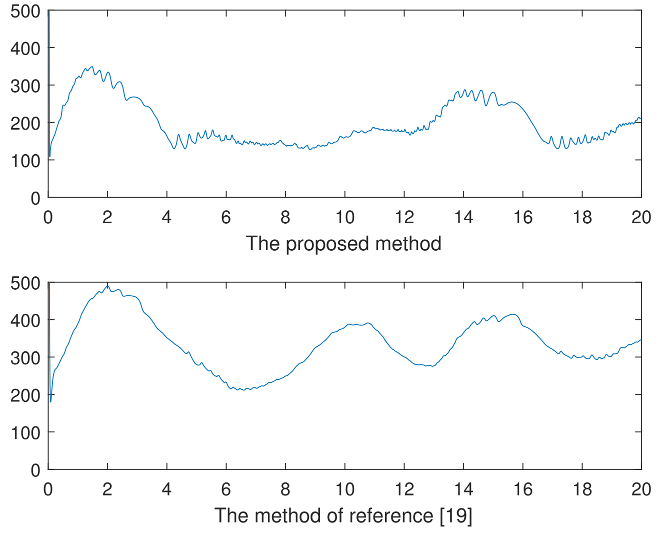

Fig. 8. Two total cost functions $r = r_1 + r_2 + r_3 + r_4$.

formation, and Fig. 2 shows the tracking error to be convergent. From Figs. 3 to 5, identifier, critic, and actor NN weights can be guaranteed to be bounded. Fig. 6 shows the cost functions. Figs. 7 and 8 show a comparison with the method of ref. [19]. Obviously, with the same control performance, the proposed method consumes less control resources than the method of ref. [19]. The simulation results further demonstrate that the proposed optimized formation scheme can realize the desired control objective.

V. Conclusion

In this article, a simplified optimized control scheme was first proposed by performing leader-follower formation control to a class of nonlinear multiagent systems with unknown dynamics. For the control scheme, NN-based RL was constructed in identifier–critic–actor architecture, where identifier was used for approximating the unknown dynamic functions, critic was used for evaluating control performance and giving feedback of the evaluation to actor, and actor was used for carrying out control behaviors. For the sake of simplifying RL algorithm, the RL updating laws were derived from the negative gradient of a simple positive function. By using the simplified optimizing scheme, the assumption of persistence excitation required in most optimal schemes was released. Based on Lyapunov stability analysis, it is proven that the control objective can be realized and the desired control performance can be arrived. Simulation results further demonstrate the effectiveness of the proposed control approach.

References

[1] R. S. Sutton and A. G. Barto, "Reinforcement learning: An introduction," *IEEE Trans. Neural Netw.*, vol. 16, no. 1, pp. 285–286, 2005.

[2] K. G. Vamvoudakis and F. L. Lewis, "Online actor–critic algorithm to solve the continuous-time infinite horizon optimal control problem," *Automatica*, vol. 46, no. 5, pp. 878–888, 2010.

[3] J. Wu, W. Chen, D. Zhao, and J. Li, "Globally stable direct adaptive backstepping NN control for uncertain nonlinear strict-feedback systems," *Neurocomputing*, vol. 122, pp. 134–147, 2013.

[4] C. L. P. Chen, G. X. Wen, Y. J. Liu, and Z. Liu, "Observer-based adaptive backstepping consensus tracking control for high-order nonlinear semi-strict-feedback multiagent systems," *IEEE Trans. Cybern.*, vol. 46, no. 7, pp. 1591–1601, Jul. 2016.

[5] H. Li, L. Bai, L. Wang, Q. Zhou, and H. Wang, "Adaptive neural control of uncertain nonstrict-feedback stochastic nonlinear systems with output constraint and unknown dead zone," *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 47, no. 8, pp. 2048–2059, Aug. 2017.

[6] G. Wen, S. S. Ge, and F. Tu, "Optimized backstepping for tracking control of strict-feedback systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 8, pp. 3850–3862, Aug. 2018.

[7] G. Wen, S. S. Ge, C. L. P. Chen, F. Tu, and S. Wang, "Adaptive tracking control of surface vessel using optimized backstepping technique," *IEEE Trans. Cybern.*, vol. 49, no. 9, pp. 3420–3431, Sep. 2019.

[8] S. Bhasin, R. Kamalapurkar, M. Johnson, K. G. Vamvoudakis, F. L. Lewis, and W. E. Dixon, "A novel actor–critic–identifier architecture for approximate optimal control of uncertain nonlinear systems," *Automatica*, vol. 49, no. 1, pp. 82–92, 2013.

[9] D. Liu, Y. Huang, W. Ding, and Q. Wei, "Neural-network-observer-based optimal control for unknown nonlinear systems using adaptive dynamic programming," *Int. J. Control*, vol. 86, no. 9, pp. 1554–1566, 2013.

[10] S. Tong, K. Sun, and S. Sui, "Observer-based adaptive fuzzy decentralized optimal control design for strict-feedback nonlinear large-scale systems," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 2, pp. 569–584, Apr. 2018.

[11] Y. Li, K. Sun, and S. Tong, "Observer-based adaptive fuzzy fault-tolerant optimal control for SIS nonlinear systems," *IEEE Trans. Cybern.*, vol. 49, no. 2, pp. 649–661, Feb. 2019.

[12] J.-L. Wang and H.-N. Wu, "Leader-following formation control of multi-agent systems under fixed and switching topologies," *Int. J. Control*, vol. 85, no. 6, pp. 695–705, 2012.

[13] M. A. Lewis and K. H. Tan, "High precision formation control of mobile robots using virtual structures," *Auton. Robot.*, vol. 4, no. 4, pp. 387–403, 1997.

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

[15] R. Olfati-Saber and R. M. Murray, "Distributed cooperative control of multiple vehicle formations using structural potential functions," *IFAC Proc. Vol.*, vol. 35, no. 1, pp. 495–500, 2002.

[16] Y. Liu and Z. Geng, "Finite-time optimal formation control of multi-agent systems on the lie group SE(3)," *Int. J. Control*, vol. 86, no. 10, pp. 1675–1686, 2013.

[17] Y. Liu and Z. Geng, "Finite-time optimal formation tracking control of vehicles in horizontal plane," *Nonlinear Dyn.*, vol. 76, no. 1, pp. 481–495, 2014.

[18] K. Sun, S. Sui, and S. Tong, "Optimal adaptive fuzzy FTC design for strict-feedback nonlinear uncertain systems with actuator faults," *Fuzzy Sets Syst.*, vol. 316, pp. 20–34, 2016.

[19] G. Wen, C. L. P. Chen, J. Feng, and N. Zhou, "Optimized multi-agent formation control based on identifier–actor–critic reinforcement learning algorithm," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 5, pp. 2719–2731, Oct. 2018.

[20] G. Wen, C. L. P. Chen, Y. J. Liu, and L. Zhi, "Neural network-based adaptive leader-following consensus control for a class of nonlinear multiagent state-delay systems," *IEEE Trans. Cybern.*, vol. 47, no. 8, pp. 2151–2160, Aug. 2017.

[21] G. Wen, C. L. P. Chen, S. S. Ge, H. Yang, and X. Liu, "Optimized adaptive nonlinear tracking control using actor–critic reinforcement learning strategy," *IEEE Trans. Ind. Inform.*, vol. 15, no. 9, pp. 4969–4977, Sep. 2019.

[22] D. Vrabie and F. Lewis, "Neural network approach to continuous-time direct adaptive optimal control for partially unknown nonlinear systems," *Neural Netw.*, vol. 22, no. 3, pp. 237–246, 2009.

[23] G. Wen, C. C. L. Philip, D. Hui, Y. Hongli, and L. Chunfang, "Formation control with obstacle avoidance of second-order multi-agent systems under directed communication topology," *Sci. China Inf. Sci.*, vol. 62, no. 9, pp. 192205:1–192205:14, 2019.

Guoxing Wen received the M.S. degree in applied mathematics from the Liaoning University of Technology, Jinzhou, China, in 2011, and the Ph.D. degree in computer and information science from the Macau University, Macau, China, in 2014.

He was a Research Fellow with the Department of Electrical and Computer Engineering, Faculty of Engineering, National University of Singapore, Singapore, from 2015 to 2016.

He is currently an Associate Professor with the College of Science, Binzhou University, China. His research interests include adaptive nonlinear control, optimal control, multiagent control, neural networks, and fuzzy logic systems.

C. L. Philip Chen (S'88–M'88–SM'94–F'07) received the M.S. degree from the University of Michigan, Ann Arbor, MI, USA, in 1985 and the Ph.D. degree from Purdue University, West Lafayette, IN, USA, in 1988, both in electrical engineering.

He is currently the Chair Professor and Dean of the College of Computer Science and Engineering, South China University of Technology, Guangzhou, China.

Dr. Chen is a Fellow of IEEE, American Association for the Advancement of Science (AAAS), International Association of Pattern Recognition (IAPR), Chinese Association of Automation (CAA), and Hong Kong Institute of Engineers' (HKIE), European Academy of Sciences and Arts (EASA), and International Academy of Systems and Cybernetics Science (IASCYS). He received IEEE Norbert Wiener Award in 2018 for his contribution in systems and cybernetics, and machine learnings. He is a highly cited researcher by Clarivate Analytics in 2018 and 2019.

Bin Li received the bachelor's, master's, and Ph.D. degrees in control science, operational research and cybernetics, and pattern recognition and intelligent system from the Shandong University, Shandong, China, in 2002, 2005, and 2012, respectively.

He is currently an Associate Professor with the School of Science, Qilu University of Technology, Jinan, China, and holds a Postdoctoral position with Shandong University. His research interests include algorithms for neural networks, gait planning, and adaptive control of legged robots.