Cooperative Finitely Excited Learning for Dynamical Games

Yongliang Yang, *Member, IEEE*, Hamidreza Modares, *Senior Member, IEEE*, Kyriakos G. Vamvoudakis, *Senior Member, IEEE*, and Frank L. Lewis, *Life Fellow, IEEE*

***Abstract*—In this article, we propose a way to enhance the learning framework for zero-sum games with dynamics evolving in continuous time. In contrast to the conventional centralized actor–critic learning, a novel cooperative finitely excited learning approach is developed to combine the online recorded data with instantaneous data for efficiency. By using an experience replay technique for each agent and distributed interaction amongst agents, we are able to replace the classical persistent excitation condition with an easy-to-check cooperative excitation condition. This approach also guarantees the consensus of the distributed actor–critic learning on the solution to the Hamilton–Jacobi–Isaacs (HJI) equation. It is shown that both the closed-loop stability of the equilibrium point and convergence to the Nash equilibrium can be guaranteed. Simulation results demonstrate the efficacy of this approach compared to previous methods.**

***Index Terms*—Cooperative finite excitation (cFE), distributed actor–critic, Hamilton–Jacobi–Isaacs (HJI) equation, Nash equilibrium, zero-sum game.**

I. Introduction

RECENT years have witnessed remarkable progress in the game-theoretical development and application advancement of distributed large-scale multiagent systems, such as pursuit-evasion games [1] and graphical games [2]. As a superior alternative to the fully connected one-to-all communication network in the centralized strategy, the distributed protocol design depends on local interaction through a sparse communication network consisting of node-to-node information propagation. Distributed synchronization protocols have been successfully designed for engineering systems, including wireless sensor networks [3] and vehicle networks [4]. As a computationally efficient alternative to *centralized* learning algorithms, cooperative *distributed* learning mechanisms benefit from the efficiency, flexibility, and scalability of distributed protocols [5].

A. Related Work

Game theory is a powerful framework for modeling decision-making problems that involve multiple strategic players. It allows us to analyze the collaborative behavior of all players and find strategies that benefit everyone [6]. A centerpiece in game theory is Nash equilibrium, which refers to the set of strategies where no player can improve their gain by unilaterally changing their strategy. H_∞ control problems with sensitivity reduction [7] and disturbance rejection [8] can be effectively addressed within the zero-sum game framework by considering the controller and the disturbance as minimizing and maximizing players, respectively, [9]. The Hamilton–Jacobi–Isaacs (HJI) equation for nonlinear systems [10], [11] and the game algebraic Riccati equation for linear quadratic games [12] play an important role in finding the Nash equilibrium [13]. Unfortunately, it is difficult to obtain an analytical solution to the HJI equation even for simple cases due to the intrinsic nonlinearity [14], [15]. Therefore, many efforts have been developed to approximate the Nash equilibrium for zero-sum games, such as Newton-like iterations [16] and Galerkin approximations [17].

Reinforcement learning (RL) and adaptive dynamic programming (ADP) bring together adaptive critic design [18] with dynamic programming [19] to assist the agent in optimizing a long-term reward through interaction with the environment [20]. Iterative ADP algorithms have been successfully employed to find the Nash equilibrium of zero-sum games for the H_∞ problems [21], [22], [23]. Online synchronous actor–critic learning algorithms [24], [25] have been successfully used to solve the HJI equations derived from zero-sum games [26]. Several recent developments can be found in [27] and [28] that consider finite-time optimal control problems as well as an output feedback design. Existing ADP/RL methods for zero-sum games are mainly based on *centralized* online synchronous learning and offline asynchronous learning, where the convergence to the Nash equilibrium can be ensured provided that a persistent excitation (PE) condition

Manuscript received 7 March 2023; revised 23 April 2023; accepted 7 May 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 61903028; in part by the Natural Science Foundation of Beijing Municipality under Grant J210005; and in part by NSF under Grant CAREER CPS-1851588, Grant CPS-2038589, and Grant CPS-2227185. This article was recommended by Associate Editor R. Tagliaferri. *(Corresponding author: Yongliang Yang.)*

Yongliang Yang is with the School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China (e-mail: yangyongliang@ieee.org).

Hamidreza Modares is with the Mechanical Engineering Department, Michigan State University, East Lansing, MI 48824 USA (e-mail: modaresh@msu.edu).

Kyriakos G. Vamvoudakis is with the Daniel Guggenheim School of Aerospace Engineering, Georgia Tech, Atlanta, GA 30332 USA (e-mail: kyriakos@gatech.edu).

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

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCYB.2023.3274908.

Digital Object Identifier 10.1109/TCYB.2023.3274908

is satisfied [26], [29]. However, the satisfaction of such a PE condition is a stringent requirement and challenging to be verifiable online due to dependence on future time instants. In addition, to enforce the PE condition in the entire time horizon, probing signals are commonly added and preserved in the control input to enrich the excitation level [30]. Although such a strategy contributes to the satisfaction of the PE condition, the inclusion of a probing signal with frequency-rich information might also degrade the smooth operation during transient.

There have been many efforts to relax the PE condition in adaptive learning and control. In the concurrent learning technique, a data memory matrix with a rank-oriented criterion is introduced to update the historical data record for online learning [31]. The concurrent learning is combined with an experience-replay-based RL for zero-sum games [32], non-zero-sum games [33], and optimal tracking/regulation problems [34]. However, the update of the singular value for the recorded data [35] and the state derivative information [36] increase the computational cost. In contrast to the data memory matrix in concurrent learning [35], composite learning [37] utilizes a moving window integral to record the online historical data to avoid the singular value calculation and state derivative differentiation. For efficient implementation purposes, it is desired to further avoid the requirement of a moving window integral, as indicated by the online filter design with a finite excitation (FE) condition [38], [39].

Contributions: Motivated by the above limitations, the contributions of the present work are as follows. In contrast to actor–critic learning for zero-sum game problems utilizing only current data [26], a novel composite learning error is defined to take the effect of both current and online historical data into account. Based on the composite critic learning error, the cooperative adaptive critic design can be implemented with online filters instead of a data matrix stack or moving window integral. To relax the requirement of the PE condition for the performance optimization convergence [12], [33], a cooperative *distributed* adaptive critic learning mechanism is designed with a novel cooperative FE (cFE) condition. It is shown that the cFE condition on the online historical data is sufficient to guarantee the stability and boundedness of signals of the closed-loop system, and the convergence of cooperative adaptive actor–critic learning to the optimal policy.

Notation: Denote $1_N = [1, \ldots, 1]^T \in \mathbb{R}^N$. Let $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ be the minimum and maximum eigenvalue of a matrix, respectively. $I_p \in \mathbb{R}^{p\times p}$ stands for the identity matrix. $\|.\|$ denotes the Euclidean norm for a vector and the Frobenius norm for a matrix. diag[·] denotes a diagonal matrix. The Kronecker product is represented by $\otimes$. The minimum and maximum operators are denoted as min(·) and max(·), respectively. $\mathbb{R}$, $\mathbb{R}^+$, $\mathbb{R}^n$, and $\mathbb{R}^{n\times m}$ denote the spaces of real numbers, positive real numbers, real n-vectors, and real $n\times m$-matrices, respectively, where n and m are positive integers. L_2 and L_∞ denote the spaces of square-integrable and bounded signals, respectively. $\text{vec}(\cdot) : \mathbb{R}^{m\times n} \to \mathbb{R}^{mn\times 1}$ is a mapping by stacking the columns of a $m \times n$ matrix, and $\text{vec}^{-1}_{m,n}(\cdot)$ is the corresponding inverse mapping.

II. Preliminaries

A. Zero-Sum Game With Finite L_2-Gain

Consider the nonlinear time-invariant nonlinear dynamics evolving in continuous time as follows:

$$\dot{x}(t) = f(x(t)) + g(x(t))u(t) + h(x(t))v(t) \tag{1}$$

where $x \in \mathbb{R}^n$ denotes the state vector, $u \in \mathbb{R}^m$ denotes the control input, $v \in \mathbb{R}^q$ denotes the disturbance input, and the initial condition $x(0) = x_0$ is given. In addition, $f(\cdot) : \mathbb{R}^n \to \mathbb{R}^n$, $g(\cdot) : \mathbb{R}^n \to \mathbb{R}^{n\times m}$, and $h(\cdot) : \mathbb{R}^n \to \mathbb{R}^{n\times q}$ are system dynamics. On a compact set $\Omega \subseteq \mathbb{R}^n$, the functions $f(\cdot)$, $g(\cdot)$, and $h(\cdot)$ are locally Lipschitz functions and satisfy $\|f(\cdot)\| \le \eta_f\|\cdot\|$, $\|g(\cdot)\| \le \eta_g$, $\|h(\cdot)\| \le \eta_h$ with positive constants η_f, η_g, and η_h. Moreover, it is assumed that $f(0) = 0$, which implies that the origin is an equilibrium of the system.

The infinite-horizon cost functional with a minimizing player u and a maximizing player v is defined as follows:

$$J(x_0, u, v) = \int_0^\infty r(x(\tau), u(\tau), v(\tau))\mathrm{d}\tau \tag{2}$$

with an instantaneous reward function given by $r(x, u, v) = Q(x) + u^T Ru - \gamma\|v\|^2$, where $Q(\cdot) \succeq 0$, $R = R^{\mathrm{T}} \succ 0$, and $\gamma > 0$ is the disturbance attenuation level.

Assumption 1: The pair $\{f(x), g(x)\}$ is stabilizable and $\{f(x), Q(x)\}$ is zero-state detectable (ZSD) [40] with $v(t) = 0$ for all $t \ge 0$.

The system (1) has a finite L_2-gain if

$$\int_t^\infty \big(Q(x(\tau)) + u^{\mathrm{T}}(\tau)Ru(\tau)\big)\mathrm{d}\tau \le \gamma^2 \int_t^\infty \|v(\tau)\|^2\mathrm{d}\tau \tag{3}$$

for all $v \in L_2[0, +\infty)$ [9]. In addition, it is assumed that γ satisfies $\gamma \ge \gamma^\star \ge 0$ with $\gamma^\star$ being the smallest disturbance attenuation level for which the system is stabilized [10].

Given feedback policy $\mu(\cdot) : \mathbb{R}^n \to \mathbb{R}^m$ and disturbance policy $\nu(\cdot) : \mathbb{R}^n \to \mathbb{R}^q$, that is, $u = \mu(x)$ and $v = \nu(x)$, the value function for system (1) is defined as follows:

$$V(x(t)) = \int_t^\infty r(x(\tau), \mu(x(\tau)), \nu(x(\tau)))\mathrm{d}\tau \ \ \forall x \in \Omega. \tag{4}$$

Note that the value function mapping is determined based on the policy pair $\{\mu(\cdot), \nu(\cdot)\}$ given the dynamical system (1). Given that the value function is finite on Ω, the value function can be obtained by solving the following equation:

$$0 = H(x, \nabla V(x), \mu(x), \nu(x)), V(0) = 0 \ \forall x \in \Omega \tag{5}$$

where $\nabla V(x) \equiv \partial V(x)/\partial x \in \mathbb{R}^n$ is the value gradient, and the Hamiltonian is defined as follows:

$$\begin{aligned} H(x, \nabla V(x), \mu(x), \nu(x)) &\\ = [\nabla V(x)]^T\big[f(x) + g(x)\mu(x) + h(x)\nu(x)\big] &\\ + r(x, \mu(x), \nu(x)). & \end{aligned} \tag{6}$$

The zero-sum differential game is defined as follows:

$$V^\star(x_0) = \min_u \max_v J(x_0, u, v). \tag{7}$$

Assumption 2: The zero-sum differential game has a unique solution, that is, the saddle point exists and satisfies

$$\min_u \max_v J(x_0, u, v) = \max_v \min_u J(x_0, u, v).$$

By applying the stationarity conditions [12]

$$\begin{aligned} 0 &= \frac{\partial}{\partial \mu(x)} H\big(x, \nabla V^\star(x), \mu(x), \nu(x)\big) \\ &= \frac{\partial}{\partial \nu(x)} H\big(x, \nabla V^\star(x), u(x), \nu(x)\big) \end{aligned}$$

one can obtain the Nash equilibrium policies $\{\mu^\star(\cdot), \nu^\star(\cdot)\}$ as [13] and [41]

$$\begin{aligned} \mu^\star(x) &= -\frac{1}{2} R^{-1} g^{\mathrm{T}}(x) \nabla V^\star(x) \\ \nu^\star(x) &= \frac{1}{2\gamma^2} h^{\mathrm{T}}(x) \nabla V^\star(x) \ \forall x \in \Omega. \end{aligned} \tag{8}$$

Substituting the Nash policies (8) into the Bellman (5) yields the HJI equation as follows [41]:

$$\begin{aligned} 0 = {} & \frac{1}{4\gamma^2} \big[\nabla V^\star(x)\big]^T h(x) h^T(x) \nabla V^\star(x) \\ & - \frac{1}{4} \big[\nabla V^\star(x)\big]^T g(x) R^{-1} g^T(x) \nabla V^\star(x) \\ & + Q(x) + \nabla V^\star(x) f(x) \quad \forall x \end{aligned} \tag{9}$$

with $V^\star(0) = 0$. The zero-sum game Nash equilibrium policies (8) for system (1) can be derived by the HJI (9). However, the HJI (9) is a nonlinear partial differential equation, which is difficult to be solved analytically.

B. Cooperative Finite Excitation

Existing RL and ADP methods aim to approximate the solution to nonlinear HJI equation using *centralized* learning [26], where the PE condition, as in Definition 1, plays an important role in the Nash equilibrium learning convergence.

Definition 1 (Persistent Excitation [42]): A vector signal $y(t) \in \mathbb{R}^p$ satisfies the PE condition if

$$\int_t^{t+T} y(\tau) y^{\mathrm{T}}(\tau) \mathrm{d}\tau \geq \alpha I_p \tag{10}$$

for all $t > 0$ with constants $T > 0$ and $\alpha > 0$.

To relax the PE condition, two types of relaxed condition are developed as follows.

Definition 2 (Finite Excitation [38]): A bounded signal $y(t) \in \mathbb{R}^p$ satisfies the FE condition over $[t_0, t_s]$ if there exists a constant $\alpha > 0$ such that $\int_{t_0}^{t_s} y(\tau) y^{\mathrm{T}}(\tau) \mathrm{d}\tau \geq \alpha I_p$.

Definition 3 (Cooperative PE [5]): Consider a group of signals $\{f_i(t)\}_{i=1}^N$ with $f_i : [t_0, \infty) \to \mathbb{R}^p$. Then, $\{f_i(t)\}_{i=1}^N$ is said to satisfy the cPE condition if there exists $\alpha > 0$ and $T > 0$ such that $\int_t^{t+T} \sum_{i=1}^N f_i(\tau) f_i^{\mathrm{T}}(\tau) \mathrm{d}\tau \geq \alpha I_p \ \ \forall t \geq t_0$.

Even the cPE condition is weaker than the PE condition, but it is still defined over the infinite horizon and not verifiable online. In [43], the cPE concept is further relaxed as cFE condition as follows.

Definition 4 (Type-I cFE): Consider a group of signals $\{f_i(t)\}_{i=1}^N$ with $f_i(t) \in \mathbb{R}^p$. Then, $\{f_i(t)\}_{i=1}^N$ satisfy the cFE condition over the interval $[0, t_s]$ with degree $\alpha > 0$ if $\int_0^{t_s} \sum_{i=1}^N f_i(\tau) f_i^{\mathrm{T}}(\tau) \mathrm{d}\tau \geq \alpha I_p$.

In this article, we extend the cFE condition as follows.

Definition 5 (Type-II cFE): For $\tau \leq t$, consider a group of signals $\{f_i(\tau, t)\}_{i=1}^N$ with $f_i(\tau, t) \in \mathbb{R}^p$. Then, $\{f_i(\tau, t)\}_{i=1}^N$ satisfy the cFE condition over $[t_0, t_s]$ with degree $\alpha > 0$ if there exist $t_s \geq 0$ such that $\mathcal{F}(t) \geq \alpha I_p \ \ \forall t \geq t_s$, where $\mathcal{F}(t) = \int_{t_0}^t \sum_{i=1}^N f_i(\tau, t) f_i^{\mathrm{T}}(\tau, t) \mathrm{d}\tau$.

The property of cFE-type conditions can be summarized as follows.

Lemma 1: For the Type-I and Type-II cFE conditions, the following hold.

1) For a group of signals $\{f_i(t)\}_{i=1}^N$ with $f_i(t) \in \mathbb{R}^{p \times 1}$. Suppose that there exist $\alpha > 0$ and $t_a > 0$ such that $\sum_{i=1}^N \int_0^t f_i(\tau) f_i^{\mathrm{T}}(\tau) \mathrm{d}\tau \geq \alpha I_p \ \ \forall t \geq t_a$. Then, there exists a positive constant $\bar{\alpha}$ such that $F(t) + (\mathcal{L} \otimes I_p) \geq \bar{\alpha} I_{Np}$ for all $t \geq t_a$, where $F(t) = \mathrm{diag}[\, F_1(t) \cdots F_N(t) \,]$ and $F_i(t) = \int_0^t f_i(\tau) f_i^T(\tau) \mathrm{d}\tau$.
2) For a group of signals $\{g_i(\tau, t)\}_{i=1}^N$ with $g_i(\cdot, \cdot) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^{p \times 1}$. Suppose that there exist $\beta > 0$ and $t_b > 0$ such that

$$\sum_{i=1}^N \int_0^t g_i(\tau, t) g_i^{\mathrm{T}}(\tau, t) \mathrm{d}\tau \geq \beta I_p \quad \forall t \geq t_b. \tag{11}$$

Then, there exists a positive constant $\bar{\beta}$ such that $G(t) + (\mathcal{L} \otimes I_p) \geq \bar{\beta} I_{Np}$ for all $t \geq t_b$, where $G(t) = diag[\, G_1(t) \cdots G_N(t) \,]$ and $G_i(t) = \int_0^t g_i(\tau, t) g_i^T(\tau, t) \mathrm{d}\tau$.

Proof: See Appendix A. ■

C. Value Function Approximation

1) Value Function Approximation for Bellman Equation: Using the function approximator, the value function associated with $\{u(\cdot), v(\cdot)\}$ can be denoted as $V(x) = W^T \phi(x) + \epsilon(x)$ and $\nabla V(x) = [\nabla \phi(x)]^T W + \nabla \epsilon(x)$ for all $x \in \Omega$, where $W = \arg\min_{w \in \mathbb{R}^p} \{\sup_{x \in \Omega} \| V(x) - w^{\mathrm{T}} \phi(x) \|\}$ is the ideal critic weight vector and $\phi(\cdot) : \mathbb{R}^n \to \mathbb{R}^p$ is the critic basis function. According to the universal approximation theorem [44], given the approximation level $\bar{\epsilon} > 0$, there exists an approximator with sufficient large basis function dimension p, such that the value function approximation residual $\epsilon(x) = V(x) - W^{\mathrm{T}} \phi(x)$ can be bounded on the compact set Ω as $\sup_{x \in \Omega} \| \epsilon(x) \| \leq \bar{\epsilon}$.

Given the value function approximation, the Bellman (5) for each agent can be expressed as follows:

$$\begin{aligned} - [\nabla \epsilon(x)]^T \big[f(x) + g(x) u(x) + h(x) v(x)\big] & \\ = W^T \nabla \phi(x) \big[f(x) + g(x) u(x) + h(x) v(x)\big] & \\ + r(x, u(x), v(x)). & \end{aligned} \tag{12}$$

Denote $\psi(x) = \nabla \phi(x)[f(x) + g(x)u(x) + h(x)v(x)]$ and $\sigma(x) = -[\nabla \epsilon(x)]^T [f(x) + g(x)u(x) + h(x)v(x)]$. Then, the residual for Bellman (5) resulting from the value function approximation can be written as follows:

$$\sigma(x) = r(x, u(x), v(x)) + W^T \psi(x) \ \forall x \in \Omega. \tag{13}$$

Assumption 3: Given the value function approximation, the following conditions hold.

1) The basis function and its gradient are bounded on Ω, that is, $\| \phi(\cdot) \| \leq \eta_\phi$, $\| \nabla \phi(\cdot) \| \leq \eta_{d\phi}$ with positive constants η_ϕ and $\eta_{d\phi}$.

2) The value function approximation residual is bounded on Ω, that is, $\sup_{x\in\Omega}\|\epsilon(x)\| \leq \bar{\epsilon}$ for all $x \in \Omega$ with constant $\bar{\epsilon} > 0$. In addition, the Bellman residual $\sigma(x)$ is bounded on Ω, that is, $\sup_{x\in\Omega}\|\sigma(x)\| \leq \bar{\sigma}$ for all $x \in \Omega$ with constant $\bar{\sigma} > 0$.
3) The ideal critic weight W is a bounded vector and uniquely exists.

2) Value Function Approximation for HJI Equation: Using the function approximator, the optimal value function can be denoted as $V^\star(x) = W_\star^{\mathrm{T}}\phi(x) + \epsilon^\star(x)$ and $\nabla V^\star(x) = [\nabla\phi(x)]^{\mathrm{T}}W_\star + \nabla\epsilon^\star(x)$ with the optimal critic weight as $W_\star = \arg\min_{W\in\mathbb{R}^p}\{\sup_{x\in\Omega}\|V_\star(x) - W^{\mathrm{T}}\phi(x)\|\}$, the critic basis function $\phi(\cdot) : \mathbb{R}^n \to \mathbb{R}^p$, and the value function approximation residual $\epsilon^\star\phi(\cdot) : \mathbb{R}^n \to \mathbb{R}$ defined as $\epsilon^\star(x) = V_\star(x) - W_\star^{\mathrm{T}}\phi(x)$.

The following assumption about the value function approximation for the optimal value function $V^\star(x)$ is required for the subsequent derivations.

Assumption 4: Given the value function approximation for $V^\star(x)$, the following conditions hold.

1) The value function approximation residual and its gradient are bounded on Ω, that is, $\|\epsilon^\star(\cdot)\| \leq \eta_\epsilon$ and $\|\nabla\epsilon^\star(\cdot)\| \leq \eta_{d\epsilon}$ with positive constants η_ϵ and $\eta_{d\epsilon}$
2) The optimal weight $W_\star$ is a bounded vector in the sense that $\|W_\star\| \leq \eta_W$ with a constant $\eta_W > 0$.

Accordingly, the HJI (9) can be expressed as follows:

$$\sigma^\star(x) = Q(x) + W_\star^{\mathrm{T}}\omega(x) - \frac{1}{4}W_\star^{\mathrm{T}}\Gamma(x)W_\star + \frac{1}{4\gamma^2}W_\star^{\mathrm{T}}\Xi(x)W_\star \tag{14}$$

with $\Gamma(x) = \nabla\phi(x)g(x)R^{-1}g^{\mathrm{T}}(x)[\nabla\phi(x)]^{\mathrm{T}}$, $\Xi(x) = \nabla\phi(x)h(x)h^T(x)[\nabla\phi(x)]^T$, and $\omega(x) = \nabla\phi(x)f(x)$. Denote $\psi^\star(x) = (1/2)\gamma^2\nabla\phi(x)h(x)h^T(x)[\nabla\phi(x)]^TW_\star - (1/2)\nabla\phi(x)g(x)R^{-1}g^T(x)[\nabla\phi(x)]^TW_\star + \nabla\phi(x)f(x)$ and then one has $W_\star^{\mathrm{T}}\psi^\star(x) = W_\star^T\omega(x) - (1/2)W_\star^T\Gamma(x)W_\star + (1/2)\gamma^2W_\star^{\mathrm{T}}\Xi(x)W_\star$. Inserting the fact into (14), one can obtain the residual for the HJI equation as $\sigma^\star(x) = W_\star^{\mathrm{T}}\psi^\star(x)+Q(x)+1/4W_\star^{\mathrm{T}}\Gamma(x)W_\star-1/4\gamma^2W_\star^{\mathrm{T}}\Xi(x)W_\star$. We will aim to solve the HJI equation (9) to find the Nash equilibrium policies $\{u^\star(\cdot), v^\star(\cdot)\}$ for the zero-sum game without the requirement of the PE/FE/cPE condition.

III. Problem Formulation

Consider a group of agents indexed as $\{1, \ldots, N\}$. The communication topology among all agents is captured by a graph. The adjacency matrix of the communication graph is denoted as $\mathcal{A}$, of which the entry is represented as a_{ij} with $i, j \in \{1, \ldots, N\}$. For the ith agent, the in-degree d_i is defined as $d_i = \sum_j a_{ij}$. The Laplacian matrix is defined as $\mathcal{L} = \mathcal{D} - \mathcal{A}$ with in-degree matrix $\mathcal{D} = \mathrm{diag}\{d_1, \ldots, d_N\}$. The group of agents have homogeneous nonlinear time-invariant dynamics given by

$$\dot{x}_i = f(x_i) + g(x_i)u_i + h(x_i)v_i, x_i(0) = x_{i0} \tag{15}$$

for $i = 1, \ldots, N$ and $t \geq 0$, where $x_i \in \mathbb{R}^n$ denotes the state, $u_i \in \mathbb{R}^m$ denotes the control input, $v_i \in \mathbb{R}^q$ denotes the disturbance input, and the initial condition x_{i0} can be distinct for each agent. The functions $f(\cdot)$, $g(\cdot)$, and $h(\cdot)$ have been defined in (1).

In the following, the cooperative actor–critic learning algorithm is designed based on the local interaction graph among the agents. Given the communication topology, the group of agents exchange the local actor–critic information to reach consensus on the Nash equilibrium given the existence of a spanning tree, that is, a tree consists of graph edges that connects all the nodes in the graph, which implies that the communication graph is connected [45]. For the subsequent discussions, the communication topology among all the agents under consideration in this article satisfies the following assumption.

Assumption 5: The communication topology amongst the agents is undirected and connected.

The problem of interest can be described as follows.

Problem 1: Consider the zero-sum game (7) with the Nash equilibrium (8) over the infinite-horizon. Design the cooperative learning algorithm with a group of agents (15) such that the following can be achieved without the requirement of the PE/FE/cPE condition.

1) Given the policy pair $\{\mu(\cdot), \nu(\cdot)\}$ for each agent, that is, $\{u_i = \mu(x_i), v_i = \nu(x_i)\}$ with $i = 1, \ldots, N$, design the distributed critic learning to learn the value function such that $\hat{W}_i(t) \to W$ as $t \to \infty$.
2) For the zero-sum game, design the distributed actor–critic learning algorithm for each agent in (15) with critic weight $\hat{W}_{i,c}$ and actor weights $\hat{W}_{i,u}$, $\hat{W}_{i,v}$ to learn the optimal value function and the Nash equilibrium such that $\hat{W}_{i,c}(t) \to W_\star$, $\hat{W}_{i,u}(t) \to W_\star$, and $\hat{W}_{i,v}(t) \to W_\star$ as $t \to \infty$ for $i = 1, \ldots, N$.

IV. Cooperative Critic Learning for Policy Evaluation

As discussed in Section II-A, the evaluation of given policy pair $\{(\mu(\cdot), \nu(\cdot))\}$ results in solving the Bellman (5). In this section, we consider the first subproblem in Problem 1 with given policy pair $\{(\mu(\cdot), \nu(\cdot))\}$ and $\{u_i = \mu(x_i), v_i = \nu(x_i)\}$ for the distributed systems (15). Based on the value function approximation in Section II-C, we employ the Type-I cFE condition in Definition 4 to construct the cooperative critic learning algorithm such that the distributed systems (15) could achieve consensus on the evaluation of the given admissible policy pair $\{(\mu(\cdot), \nu(\cdot))\}$.

A. Composite Critic Design

Define the following filtered signals:

$$\begin{aligned}\dot{r}_{i,f}(t) &= -\kappa \cdot r_{i,f}(t) + r(x_i(t), u(x_i(t)), v(x_i(t)))\\ \dot{\psi}_{i,f}(t) &= -\kappa \cdot \psi_{i,f}(t) + \psi(x_i(t))\\ \dot{\sigma}_{i,f}(t) &= -\kappa \cdot \sigma_{i,f}(t) + \sigma(x_i(t))\end{aligned} \tag{16}$$

with $\psi_{i,f}(0) = \sigma_{i,f}(0) = r_{i,f}(0) = 0$ and $\kappa > 0$ being the filter design parameter. The above signals $r(x_{i,f}(t))$, $\psi(x_{i,f}(t))$, and $\sigma(x_{i,f}(t))$ can be interpreted as frequency domain representation of the signals $r(x_i(t))$, $\psi(x_i(t))$, and $\sigma(x_i(t))$ passing through the filter $1/s + \kappa$.

Then, the fact in (13) implies that

$$\sigma_{i,f}(t) = r_{i,f}(t) + W^T \psi_{i,f}(t). \tag{17}$$

In [26], the critic error is considered based on the instantaneous data as $e_i(t) = r(x_i(t), u(x_i(t)), v(x_i(t))) + \hat{W}_i^T(t)\psi(x_i(t))$, where $\hat{W}_i(t)$ is the adaptive critic weight with learning error $\tilde{W}_i(t) = W - \hat{W}_i(t)$. In contrast, we propose the composite critic learning error as $e_{i,f}(\tau, t) = r_{i,f}(\tau) + \hat{W}_i^T(t)\psi_{i,f}(\tau)$, where $\hat{W}_i(t)$ represents the current critic weight estimation and $\{r_{i,f}(\tau), \psi_{i,f}(\tau)\}$ denote the stored system information with $0 < \tau < t$. Based on the composite critic error design, the critic learning objective for each agent is considered as follows:

$$J_{i,f}\left(\hat{W}_i(t)\right) = J_{i,1}^f\left(\hat{W}_i(t)\right) + J_{i,2}^f\left(\hat{W}_i(t)\right)$$

where $B_i(\tau) = 1 + [\psi_{i,f}(\tau)]^T\psi_{i,f}(\tau)$, $J_{i,1}^f(t) = 1/2\int_0^t \|e_{i,f}(\tau, t)\|^2/[B_i(\tau)]^2 \mathrm{d}\tau$, and $J_{i,2}^f(t) = (1/2)\sum_{j\in\mathcal{N}_i} a_{ij}\|\varpi_{i,j}(t)\|^2$, and $\varpi_{i,j}(t) = \hat{W}_j(t) - \hat{W}_i(t)$ is the discrepancy of the critic weights between agents i and j.

Applying the chain rule, the critic learning objective gradient for each agent can be obtained as follows:

$$\begin{aligned}\frac{\partial J_{i,f}\left(\hat{W}_i(t)\right)}{\partial \hat{W}_i(t)} &= \frac{\partial J_{i,1}^f\left(\hat{W}_i(t)\right)}{\partial \hat{W}_i(t)} + \frac{\partial J_{i,2}^f\left(\hat{W}_i(t)\right)}{\partial \hat{W}_i(t)} \\ &= \int_0^t \frac{\psi_{i,f}(\tau)\cdot r_{i,f}(\tau)}{[B_i(\tau)]^2}\mathrm{d}\tau \\ &\quad + \int_0^t \frac{\psi_{i,f}(\tau)\left[\psi_{i,f}(\tau)\right]^T}{[B_i(\tau)]^2}\mathrm{d}\tau \cdot \hat{W}_i(t)\end{aligned}$$

with $\partial J_{i,1}^f(\hat{W}_i(t))/\partial \hat{W}_i(t) = \int_0^t \psi_{i,f}(\tau)\cdot e_{i,f}(\tau, t)/[B_i(\tau)]^2\mathrm{d}\tau$ and $\partial J_{i,2}^f(\hat{W}_i(t))/\partial \hat{W}_i(t) = -\sum_{j\in\mathcal{N}_i} a_{ij}\varpi_{i,j}(t)$. To obtain an efficient online implementation of the objective gradient $\partial J_{i,1}^f(\hat{W}_i(t))/\partial \hat{W}_i(t)$, we further design the following filters as:

$$\begin{aligned}\delta_{i,f}(t) &= \int_0^t \frac{\psi_{i,f}(\tau)\sigma_{i,f}(\tau)}{\left\{1 + \left[\psi_{i,f}(\tau)\right]^{\mathrm{T}}\psi_{i,f}(\tau)\right\}^2}\mathrm{d}\tau \\ F_{i,f}(t) &= \int_0^t \frac{\psi_{i,f}(\tau)r_{i,f}(\tau)}{\left\{1 + \left[\psi_{i,f}(\tau)\right]^{\mathrm{T}}\psi_{i,f}(\tau)\right\}^2}\mathrm{d}\tau \\ E_{i,f}(t) &= \int_0^t \frac{\psi_{i,f}(\tau)\left[\psi_{i,f}(\tau)\right]^T}{\left\{1 + \left[\psi_{i,f}(\tau)\right]^{T}\psi_{i,f}(\tau)\right\}^2}\mathrm{d}\tau.\end{aligned}$$

Then, (17) further implies that

$$\delta_{i,f}(t) = F_{i,f}(t) + E_{i,f}(t)\cdot W. \tag{18}$$

Then, the gradient $\partial J_{i,1}^f(\hat{W}_i(t))/\partial \hat{W}_i(t)$ can be rewritten as follows:

$$\frac{\partial J_{i,1}^f\left(\hat{W}_i(t)\right)}{\partial \hat{W}_i(t)} = F_{i,f}(t) + E_{i,f}(t)\cdot \hat{W}_i(t). \tag{19}$$

Therefore, one has

$$\begin{aligned}\frac{\partial J_{i,f}}{\partial \hat{W}_i} &= \frac{\partial J_{i,1}^f\left(\hat{W}_i(t)\right)}{\partial \hat{W}_i(t)} - \sum_{j=1}^N a_{ij}\varpi_{i,j}(t) \\ &= F_{i,f}(t) + E_{i,f}(t)\cdot \hat{W}_i(t) - \sum_{j=1}^N a_{ij}\varpi_{i,j}(t).\end{aligned} \tag{20}$$

To this end, the distributed critic learning can be designed as follows:

$$\dot{\hat{W}}_i(t) = -\gamma\left[F_{i,f}(t) + E_{i,f}(t)\cdot \hat{W}_i(t) - \sum_{j=1}^N a_{ij}\varpi_{i,j}(t)\right]. \tag{21}$$

Theorem 1: Suppose that $\bar{\psi}_{i,f}(\tau) = \psi_{i,f}(\tau)/(1 + [\psi_{i,f}(\tau)]^T\psi_{i,f}(\tau))$ satisfies the type-I cFE condition in the sense that there exists $\alpha > 0$ such that

$$E_{i,f}(t) = \sum_{i=1}^N \int_0^t \bar{\psi}_{i,f}(\tau)\left[\bar{\psi}_{i,f}(\tau)\right]^T\mathrm{d}\tau \geq \alpha I_p \ \forall t \geq t_s. \tag{22}$$

Given the critic design (21), one has the following.

1) There exists a positive constant α_c such that $E_f(t) + (L \otimes I_p) \geq \alpha_c I_{Np}$ $\forall t \geq t_s$, where $E_f(t) = \mathrm{diag}(E_{1,f}(t), \ldots, E_{N,f}(t))$.
2) All the closed-loop signals for each agent are L_∞-stable and the state of each agent asymptotically converges to a small neighborhood of the origin on $t \in [0, +\infty)$.
3) The critic weight error $\tilde{W}_i$ of each agent exponentially converge to small neighborhoods of the origin on $t \in [t_s, +\infty)$.

Proof: See Appendix B. ■

Remark 1: For the existing adaptive optimal learning control with relaxed excitation [34], [46], [47], the critic learning involves a singular-value-based condition on the collected data matrix, where the time complexity for computing the data matrix singular value is $\mathcal{O}(p^3)$ with p being the dimension of the basis function for value function approximation. In contrast, for the presented cooperative critic learning, the data matrix singular value calculation can be avoided. As shown in (21), the critic learning depends on the filters design (16) with time complexity being $\mathcal{O}(p^2)$, which is similar to least-squares temporal difference (LSTD) algorithm [48] and more efficient than the existing relaxed excited critic learning design [34], [46], [47].

V. Cooperative Actor–Critic Learning for Policy Optimization

In the following, based on the value function representation using function approximators in Section II-C, we will employ the Type-II cFE condition in Definition 5 to construct the distributed actor–critic learning algorithm to learn the solution to the HJI equation (9). As investigated in Theorem 2, the distributed actor–critic learning algorithm would converge toward the Nash equilibrium without the requirement of the PE/FE/cPE condition.

A. Composite Actor–Critic Design

In this section, we design two types of filters, as shown in (23), to collect the online data. On this basis, a novel cooperative critic learning algorithm is constructed to learn the Nash feedback equilibrium for each agent.

1) Online Filters Design: Define the filtered signals

$$\begin{aligned}\dot{Q}_{i,f}(t) &= -\kappa\cdot Q_{i,f}(t) + Q(x_i(t))\\ \dot{\omega}_{i,f}(t) &= -\kappa\cdot \omega_{i,f}(t) + \omega(x_i(t))\\ \dot{\Gamma}_{i,f}(t) &= -\kappa\cdot \Gamma_{i,f}(t) + \Gamma(x_i(t))\\ \dot{\Xi}_{i,f}(t) &= -\kappa\cdot \Xi_{i,f}(t) + \Xi(x_i(t))\\ \dot{\sigma}^{\star}_{i,f}(t) &= -\kappa\cdot \sigma^{\star}_{i,f}(t) + \sigma^{\star}(x_i(t))\end{aligned} \tag{23}$$

with zero initial conditions. The signals $Q_{i,f}(t)$, $\omega_{i,f}(t)$, $\Gamma_{i,f}(t)$, $\Xi_{i,f}(t)$, and $\sigma_{i,f}(t)$ can be interpreted as frequency domain representation of the signals $Q(x_i(t))$, $\omega(x_i(t))$, $\Gamma(x_i(t))$, $\Xi(x_i(t))$, and $\sigma(x_i(t))$ passing through the filter $1/s+\kappa$.

For the subsequent designs, we denote $a_{i,\omega\omega}(t) = \int_0^t \omega_{i,f}(\tau)\omega_{i,f}^{\mathrm{T}}(\tau)\mathrm{d}\tau$, $a_{i,\omega\Gamma}(t) = \int_0^t \Gamma_{i,f}(\tau)\otimes\omega_{i,f}(\tau)\mathrm{d}\tau$, $a_{i,\Gamma\omega}(t) = \int_0^t \omega_{i,f}(\tau)\otimes\Gamma_{i,f}(\tau)\mathrm{d}\tau$, $a_{i,\Gamma\Gamma}(t) = \int_0^t \Gamma_{i,f}(\tau)\otimes\Gamma_{i,f}(\tau)\mathrm{d}\tau$, $a_{i,\omega\Xi}(t) = \int_0^t \Xi_{i,f}(\tau)\otimes\omega_{i,f}(\tau)\mathrm{d}\tau$, $a_{i,\Xi\omega}(t) = \int_0^t \omega_{i,f}(\tau)\otimes\Xi_{i,f}(\tau)\mathrm{d}\tau$, $a_{i,\Gamma\Xi}(t) = \int_0^t \Xi_{i,f}(\tau)\otimes\Gamma_{i,f}(\tau)\mathrm{d}\tau$, $a_{i,\Xi\Gamma}(t) = \int_0^t \Gamma_{i,f}(\tau)\otimes\Xi_{i,f}(\tau)\mathrm{d}\tau$, $a_{i,\Xi\Xi}(t) = \int_0^t \Xi_{i,f}(\tau)\otimes\Xi_{i,f}(\tau)\mathrm{d}\tau$, $d_{i,Q\omega}(t) = \int_0^t Q_{i,f}(\tau)\omega_{i,f}(\tau)\mathrm{d}\tau$, $d_{i,Q\Gamma}(t) = \int_0^t Q_{i,f}(\tau)\Gamma_{i,f}(\tau)\mathrm{d}\tau$, $d_{i,Q\Xi}(t) = \int_0^t Q_{i,f}(\tau)\Xi_{i,f}(\tau)\mathrm{d}\tau$, $b_{i,\omega\omega}(t) = \int_0^t \omega_{i,f}^{\mathrm{T}}(\tau)\omega_{i,f}(\tau)\mathrm{d}\tau$, $b_{i,\Gamma\Gamma}(t) = \int_0^t \Gamma_{i,f}(\tau)\Gamma_{i,f}(\tau)\mathrm{d}\tau$, $b_{i,\Gamma\omega}(t) = \int_0^t \Gamma_{i,f}(\tau)\omega_{i,f}(\tau)\mathrm{d}\tau$, $b_{i,\omega\Gamma}(t) = \int_0^t \omega_{i,f}^{T}(\tau)\Gamma_{i,f}(\tau)\mathrm{d}\tau$, $b_{i,\omega\Xi}(t) = \int_0^t \omega_{i,f}^{T}(\tau)\Xi_{i,f}(\tau)\mathrm{d}\tau$, $b_{i,\Gamma\Xi}(t) = \int_0^t \Gamma_{i,f}(\tau)\Xi_{i,f}(\tau)\mathrm{d}\tau$, $b_{i,\Xi\omega}(t) = \int_0^t \Xi_{i,f}(\tau)\omega_{i,f}(\tau)\mathrm{d}\tau$, $b_{i,\Xi\Gamma}(t) = \int_0^t \Xi_{i,f}(\tau)\Gamma_{i,f}(\tau)\mathrm{d}\tau$, and $b_{i,\Xi\Xi}(t) = \int_0^t \Xi_{i,f}(\tau)\Xi_{i,f}(\tau)\mathrm{d}\tau$.

Given the fact in (14), one has

$$\begin{aligned}\sigma^{\star}_{i,f}(t) = Q_{i,f}(t) + W_{\star}^{\mathrm{T}}\omega_{i,f}(t) - \frac{1}{4}W_{\star}^{\mathrm{T}}\Gamma_{i,f}(t)W_{\star}\\ + \frac{1}{4\gamma^2}W_{\star}^{\mathrm{T}}\Xi_{i,f}(t)W_{\star}.\end{aligned} \tag{24}$$

Denote $\psi^{\star}_{i,f}(t) = \omega_{i,f}(t) - (1/2)\Gamma_{i,f}(t)W_{\star} + 1/2\gamma^2\Xi_{i,f}(t)W_{\star}$ and $r^{\star}_{i,f}(t) = Q_{i,f}(t) + 1/4W_{\star}^{\mathrm{T}}\Gamma_{i,f}(t)W_{\star}$. Then, (24) can be rewritten in a compact form as follows:

$$\sigma^{\star}_{i,f}(t) = W_{\star}^{\mathrm{T}}\psi^{\star}_{i,f}(t) + r^{\star}_{i,f}(t). \tag{25}$$

2) Distributed Critic Learning: First, with the notations in (23), define

$$\begin{aligned}\psi^{a}_{i,f}(t,\tau) &= \omega_{i,f}(\tau) - \frac{1}{2}\Gamma_{i,f}(\tau)\hat{W}_{i,u}(t)\\ &\quad + \frac{1}{2\gamma^2}\Xi_{i,f}(\tau)\hat{W}_{i,v}(t)\\ r^{a}_{i,f}(t,\tau) &= Q_{i,f}(\tau) + \frac{1}{4}\hat{W}_{i,u}^{T}(t)\Gamma_{i,f}(\tau)\hat{W}_{i,u}(t)\\ &\quad - \frac{1}{4\gamma^2}\hat{W}_{i,v}^{T}(t)\Xi_{i,f}(\tau)\hat{W}_{i,v}(t).\end{aligned} \tag{26}$$

Then, the composite critic learning error with online actors $u_i(t)$ and $v_i(t)$ is considered as follows:

$$e_{i,c}(t,\tau) = \hat{W}_{i,c}^{\mathrm{T}}(t)\psi^{a}_{i,f}(t,\tau) + r^{a}_{i,f}(t,\tau). \tag{27}$$

The critic learning objective for each agent is considered as follows:

$$J_{i,c}\Big(\hat{W}_{i,c}(t)\Big) = \bar{J}^{c}_{i,1}(t) + J^{c}_{i,2}(t)$$

with $\bar{J}^{c}_{i,1}(t) = J^{c}_{i,1}(\hat{W}_{i,c}(t))/[B_{i,c}(t)]^2$, $J^{c}_{i,1}(\hat{W}_{i,c}(t)) = (1/2)\int_0^t \|e_{i,c}(t,\tau)\|^2\mathrm{d}\tau$, $J^{c}_{i,2}(t) = (1/2)\sum_{j\in N_i} a_{ij}\|\varpi^{c}_{i,j}(t)\|^2$, the normalization term $B_{i,c}(t) = 1 + \int_0^t \|\psi^{a}_{i,f}(t,\tau)\|^2\mathrm{d}\tau$ and $\varpi^{c}_{i,j} = \hat{W}_{j,c}(t) - \hat{W}_{i,c}(t)$.

The distributed critic learning for each agent is designed as follows:

$$\begin{aligned}\dot{\hat{W}}_{i,c}(t) = \gamma_c\frac{1}{\big[B_{i,c}(t)\big]^2}\frac{\partial J^{c}_{i,1}\Big(\hat{W}_{i,c}(t)\Big)}{\partial\hat{W}_{i,c}(t)}\\ + \gamma_c\sum_{j=1}^{N}a_{ij}\Big[\hat{W}_{j,c}(t) - \hat{W}_{i,c}(t)\Big].\end{aligned} \tag{28}$$

In the following, the filters design are employed to online implement the above adaptive critic learning algorithm as:

$$\begin{aligned}B_{i,c}(t) &= 1 + \frac{1}{4}\hat{W}_{i,u}^{\mathrm{T}}(t)b_{i,\Gamma\Gamma}(t)\hat{W}_{i,u}(t)\\ &\quad + b_{i,\omega\omega}(t) - \frac{1}{2}\hat{W}_{i,u}^{\mathrm{T}}(t)b_{i,\Gamma\omega}(t) - \frac{1}{2}b_{i,\omega\Gamma}(t)\hat{W}_{i,u}(t)\\ &\quad + \frac{1}{2\gamma^2}b_{i,\omega\Xi}(t)\hat{W}_{i,v}(t) - \frac{1}{4\gamma^2}\hat{W}_{i,u}^{T}(t)b_{i,\Gamma\Xi}(t)\hat{W}_{i,v}(t)\\ &\quad + \frac{1}{2\gamma^2}\hat{W}_{i,v}^{T}(t)b_{i,\Xi\omega}(t) - \frac{1}{4\gamma^2}\hat{W}_{i,v}^{T}(t)b_{i,\Xi\Gamma}(t)\hat{W}_{i,u}(t)\\ &\quad + \frac{1}{4\gamma^4}\hat{W}_{i,v}^{T}(t)b_{i,\Xi\Xi}(t)\hat{W}_{i,v}(t).\end{aligned}$$

Next, the objective gradient can be further expressed as $\partial J^{c}_{i,1}(\hat{W}_{i,c}(t))/\partial\hat{W}_{i,c}(t) = A_{i,c}(t)\cdot\hat{W}_{i,c}(t) + D_{i,c}(t)$ with $A_{i,c}(t) = \int_0^t \psi^{a}_{i,f}(t,\tau)[\psi^{a}_{i,f}(t,\tau)]^{\mathrm{T}}\mathrm{d}\tau$ and $D_{i,c}(t) = \int_0^t \psi^{a}_{i,f}(t,\tau)r^{a}_{i,f}(t,\tau)\mathrm{d}\tau$.

With the Kronecker product property, one has $\int_0^t \omega_{i,f}(\tau)\hat{W}_{i,u}^{T}(t)\Gamma_{i,f}(\tau)\mathrm{d}\tau = \mathrm{vec}^{-1}_{M,M}(a_{i,\omega\Gamma}(t)\cdot\mathrm{vec}(\hat{W}_{i,u}^{T}(t)))$, $\int_0^t \Gamma_{i,f}\hat{W}_{i,u}(t)\omega_{i,f}^{T}(\tau)\mathrm{d}\tau = \mathrm{vec}^{-1}_{M,M}[a_{i,\Gamma\omega}(t)\cdot\mathrm{vec}(\hat{W}_{i,u}(t))]$, $\int_0^t \Gamma_{i,f}(\tau)\hat{W}_{i,u}(t)\hat{W}_{i,u}^{T}(t)\Gamma_{i,f}(\tau)\mathrm{d}\tau = \mathrm{vec}^{-1}_{M,M}(a_{i,\Gamma\Gamma}(t)\mathrm{vec}(\hat{W}_{i,u}(t)\hat{W}_{i,u}^{T}(t)))$, $\int_0^t \omega_{i,f}(\tau)\hat{W}_{i,v}^{T}(t)\cdot\Xi_{i,f}(\tau)\mathrm{d}\tau = \mathrm{vec}^{-1}_{M,M}(a_{i,\omega\Xi}(t)\hat{W}_{i,v}(t))$, $\int_0^t \Xi_{i,f}(\tau)\hat{W}_{i,v}(t)\ \omega_{i,f}(\tau)\mathrm{d}\tau = \mathrm{vec}^{-1}_{M,M}(a_{i,\Xi\omega}(t)\hat{W}_{i,v}(t))$, $\int_0^t \Gamma_{i,f}(\tau)\hat{W}_{i,u}(t)\ \hat{W}_{i,v}^{T}(t)\ \Xi_{i,f}(\tau)\mathrm{d}\tau = \mathrm{vec}^{-1}_{M,M}(a_{i,\Gamma\Xi}(t)\mathrm{vec}[\hat{W}_{i,u}(t)\cdot\ \hat{W}_{i,v}^{T}(t)])$, $\int_0^t \Xi_{i,f}(\tau)\ \hat{W}_{i,v}(t)\ \hat{W}_{i,u}^{T}(t)\Gamma_{i,f}(\tau)\mathrm{d}\tau = \mathrm{vec}^{-1}_{M,M}(a_{i,\Xi\Gamma}(t)\mathrm{vec}\ [\hat{W}_{i,v}(t)\ \hat{W}_{i,u}^{T}(t)])$, $\int_0^t \Xi_{i,f}(\tau)\hat{W}_{i,v}(t)\cdot\ \hat{W}_{i,v}^{T}(t)\Xi_{i,f}(\tau)\mathrm{d}\tau = \mathrm{vec}^{-1}_{M,M}\ (a_{i,\Xi\Xi}(t)\ \mathrm{vec}[\hat{W}_{i,v}(t)\hat{W}_{i,v}^{T}(t)])$. Therefore, $A_{i,c}(t)$ can be further rewritten as follows:

$$\begin{aligned}A_{i,c}(t) &= a_{i,\omega\omega}(t) - \frac{1}{2}\mathrm{vec}^{-1}_{M,M}\Big(a_{i,\omega\Gamma}(t)\cdot\hat{W}_{i,u}(t)\Big)\\ &\quad - \frac{1}{2}\mathrm{vec}^{-1}_{M,M}\Big(a_{i,\Gamma\omega}(t)\cdot\hat{W}_{i,u}(t)\Big)\\ &\quad + \frac{1}{4}\mathrm{vec}^{-1}_{M,M}\Big(a_{i,\Gamma\Gamma}(t)\cdot\mathrm{vec}\Big(\hat{W}_{i,u}(t)\hat{W}_{i,u}^{\mathrm{T}}(t)\Big)\Big)\\ &\quad + \frac{1}{2\gamma^2}\mathrm{vec}^{-1}_{M,M}\Big(a_{i,\omega\Xi}(t)\cdot\hat{W}_{i,v}(t)\Big)\end{aligned}$$

$$+\frac{1}{2\gamma^2}\mathrm{vec}^{-1}_{M,M}\Big(a_{i,\Xi\omega}(t)\cdot\hat{W}_{i,v}(t)\Big)$$
$$-\frac{1}{4\gamma^2}\mathrm{vec}^{-1}_{M,M}\Big(a_{i,\Gamma\Xi}(t)\cdot\mathrm{vec}\Big[\hat{W}_{i,u}(t)\hat{W}^T_{i,v}(t)\Big]\Big)$$
$$-\frac{1}{4\gamma^2}\mathrm{vec}^{-1}_{M,M}\Big(a_{i,\Xi\Gamma}(t)\cdot\mathrm{vec}\Big[\hat{W}_{i,v}(t)\hat{W}^T_{i,u}(t)\Big]\Big)$$
$$+\frac{1}{4\gamma^4}\mathrm{vec}^{-1}_{M,M}\Big(a_{i,\Xi\Xi}(t)\cdot\mathrm{vec}\Big[\hat{W}_{i,v}(t)\hat{W}^T_{i,v}(t)\Big]\Big).$$

Similarly, using the Kronecker product, one has

$$D_{i,c}(t)=d_{i,Q\omega}(t)-\frac{1}{2}d_{i,Q\Gamma}(t)\hat{W}_{i,u}(t)+\frac{1}{2\gamma^2}d_{i,Q\Xi}(t)\hat{W}_{i,v}(t)$$
$$+\frac{1}{4}\mathrm{vec}^{-1}_{M,M}\Big(a_{i,\omega\Gamma}(t)\hat{W}_{i,u}(t)\Big)\hat{W}_{i,u}(t)$$
$$-\frac{1}{4\gamma^2}\mathrm{vec}^{-1}_{M,M}\Big(a_{i,\omega\Xi}(t)\hat{W}_{i,v}(t)\Big)\hat{W}_{i,v}(t)$$
$$-\frac{1}{8}\mathrm{vec}^{-1}_{M,M}\Big(a_{i,\Gamma\Gamma}(t)\mathrm{vec}\Big(\hat{W}_{i,u}(t)\hat{W}^T_{i,u}(t)\Big)\Big)\hat{W}_{i,u}(t)$$
$$+\frac{1}{8\gamma^2}\mathrm{vec}^{-1}_{M,M}\Big(a_{i,\Xi\Gamma}(t)\mathrm{vec}\Big(\hat{W}_{i,v}(t)\hat{W}^T_{i,u}(t)\Big)\Big)\hat{W}_{i,u}(t)$$
$$+\frac{1}{8\gamma^2}\mathrm{vec}^{-1}_{M,M}\Big(a_{i,\Gamma\Xi}(t)\mathrm{vec}\Big(\hat{W}_{i,u}(t)\hat{W}^T_{i,v}(t)\Big)\Big)\hat{W}_{i,v}(t)$$
$$-\frac{1}{8\gamma^4}\mathrm{vec}^{-1}_{M,M}\Big(a_{i,\Xi\Xi}(t)\mathrm{vec}\Big(\hat{W}_{i,v}(t)\hat{W}^T_{i,v}(t)\Big)\Big)\hat{W}_{i,v}(t).$$

3) Distributed Actor Learning: The online actors for the min–max players are designed as follows:

$$u_i(t)=-\frac{1}{2}R^{-1}g^{\mathrm{T}}(x_i)[\nabla\phi(x_i)]^{\mathrm{T}}\hat{W}_{i,u}(t) \tag{29}$$
$$v_i(t)=\frac{1}{2\gamma^2}h^T(x_i)[\nabla\phi(x_i)]^T\hat{W}_{i,v}(t) \tag{30}$$

with the adaptive learning for the minimizing player u_i and the maximizing player v_i as follows:

$$\dot{\hat{W}}_{i,u}(t)=\gamma_u\Bigg\{-K_{i,\mathrm{uu}}\hat{W}_{i,u}(t)+\frac{1}{B_{i,c}(t)}K_{i,\mathrm{cu}}F_{i,c}(t)$$
$$+\frac{1}{4}\frac{1}{\big[B_{i,c}(t)\big]^2}F_{i,u}(t)+\sum_{j=1}^{N}a_{ij}\Big[\hat{W}_{j,u}(t)-\hat{W}_{i,u}(t)\Big]\Bigg\} \tag{31}$$

and

$$\dot{\hat{W}}_{i,v}(t)=\gamma_v\Bigg\{-K_{i,\mathrm{vv}}\hat{W}_{i,v}(t)+\frac{1}{B_{i,c}(t)}K_{i,\mathrm{cv}}F_{i,c}(t)$$
$$-\frac{1}{4\gamma^2}\frac{1}{\big[B_{i,c}(t)\big]^2}F_{i,v}(t)$$
$$+\gamma_v\sum_{j=1}^{N}a_{ij}\Big[\hat{W}_{j,v}(t)-\hat{W}_{i,v}(t)\Big]\Bigg\} \tag{32}$$

where $F_{i,u}(t)=\int_0^t\Gamma_{i,f}(\tau)\hat{W}_{i,u}(t)\ [\psi^a_{i,f}(t,\tau)]^{\mathrm{T}}\hat{W}_{i,c}(t)\mathrm{d}\tau$, $F_{i,v}(t)=\int_0^t\Xi_{i,f}(\tau)\hat{W}_{i,v}(t)\ [\psi^a_{i,f}(t,\tau)]^T\hat{W}_{i,c}(t)\mathrm{d}\tau$, and $F_{i,c}(t)=\int_0^t[\psi^a_{i,f}(t,\tau)]^{\mathrm{T}}\hat{W}_{i,c}(t)\mathrm{d}\tau$. From the definition of $\psi^a_{i,f}(t,\tau)$ in (26), the terms $F_{i,c}(t)$, $F_{i,u}(t)$, and $F_{i,v}(t)$ can be equivalently obtained as $F_{i,c}(t)=\int_0^t\omega^T_{i,f}(\tau)\mathrm{d}\tau\hat{W}_{i,c}(t)-(1/2)\hat{W}^T_{i,u}(t)\int_0^t\Gamma_{i,f}(\tau)\mathrm{d}\tau\hat{W}_{i,c}(t)+(1/2)\gamma^2\cdot\hat{W}^T_{i,v}(t)\int_0^t\Xi_{i,f}(\tau)\mathrm{d}\tau\hat{W}_{i,c}(t)$, and $F_{i,u}(t)=\mathrm{vec}^{-1}_{M,M}(a_{i,\Gamma\omega}(t)\cdot\hat{W}_{i,u}(t))\ \hat{W}_{i,c}(t)-(1/2)\mathrm{vec}^{-1}_{M,M}(a_{i,\Gamma\Gamma}(t)\mathrm{vec}(\hat{W}_{i,u}(t)\hat{W}^{\mathrm{T}}_{i,u}(t)))\hat{W}_{i,c}(t)+(1/2)\gamma^2\mathrm{vec}^{-1}_{M,M}(a_{i,\Gamma\Xi}(t)\mathrm{vec}(\hat{W}_{i,u}(t)\hat{W}^{\mathrm{T}}_{i,v}(t)))\hat{W}_{i,c}(t)$, and $F_{i,v}(t)=\mathrm{vec}^{-1}_{M,M}(a_{i,\Xi\omega}(t)\hat{W}_{i,v}(t))\hat{W}_{i,c}(t)-(1/2)\mathrm{vec}^{-1}_{M,M}(a_{i,\Xi\Gamma}(t)\mathrm{vec}(\hat{W}_{i,v}(t)\hat{W}^{\mathrm{T}}_{i,u}(t)))\hat{W}_{i,c}(t)+(1/2)\gamma^2\mathrm{vec}^{-1}_{M,M}(a_{i,\Xi\Xi}(t)\mathrm{vec}(\hat{W}_{i,v}(t)\hat{W}^{\mathrm{T}}_{i,v}(t)))\hat{W}_{i,c}(t)$.

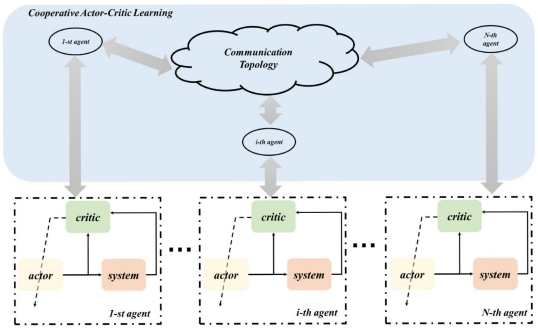


Fig. 1. Cooperative actor–critic learning for H_∞ control problem.

Theorem 2: Suppose that the signal $\psi^a_{i,f}(t,\tau)$ defined in (26) satisfies the type-II cFE condition in the sense that there exists $\alpha>0$ such that

$$\sum_{i=1}^{N}\int_0^t\psi^a_{i,f}(t,\tau)\Big[\psi^a_{i,f}(t,\tau)\Big]^T\mathrm{d}\tau\geq\alpha\cdot I_p\quad\forall t\geq t_s. \tag{33}$$

Consider the zero-sum game with performance (2) and dynamics (1). Design the critic learning (28), and actor learning (31) and (32) with $K_{i,cu}$, $K_{i,uu}$, $K_{i,cv}$, $K_{i,vv}$ satisfying $K_{i,cu}\ll K_{i,uu}$ and $K_{i,cv}\ll K_{i,vv}$. Then, one has the following conclusions.

1) Denote $A_c(t)=\mathrm{diag}[A_{1,c}(t)\cdots A_{N,c}(t)]$. Then, there exists a positive constant α such that $A_c(t)\geq\alpha I_{Np}$ for $t\geq t_s$.
2) All the closed-loop signals for each agent are L_∞-stable and the state of each agent asymptotically converges to a small neighborhood of the origin on $t\in[0,+\infty)$.
3) For each agent, the state x_i, the critic weight error $\tilde{W}_{i,c}$, the actor weight errors $\tilde{W}_{i,u}$ and $\tilde{W}_{i,v}$ are uniformly ultimately bounded (UUB).[1]

Proof: See Appendix C. ■

The overall design framework for the cooperative actor–critic algorithm can be summarized as in Fig. 1. Each agent implements its own composite actor–critic learning, where the local interaction of the actor–critic information among agents is considered. This corresponds to the terms $\sum_{j=1}^N a_{ij}[\hat{W}_{j,c}(t)-\hat{W}_{i,c}(t)]$, $\sum_{j=1}^N a_{ij}[\hat{W}_{j,u}(t)-\hat{W}_{i,u}(t)]$, and $\sum_{j=1}^N a_{ij}[\hat{W}_{j,v}(t)-\hat{W}_{i,v}(t)]$ in the cooperative actor–critic learning design (28), (31), and (32), which captures the interaction between agent i and agent j through the communication topology.

Remark 2: As shown in the proof of Theorems 1 and 2, the convergence rate of the critic learning depends on the excitation degree of $\sum_{i=1}^N\int_0^t\bar{\psi}_{i,f}(\tau)[\bar{\psi}_{i,f}(\tau)]^T\mathrm{d}\tau$ and

[1]The concept of UUB is defined in [49].

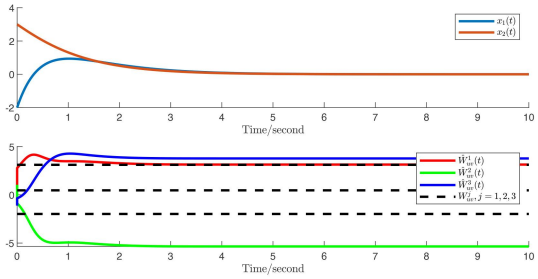


Fig. 2. Policy evaluation with gradient descent learning [26].

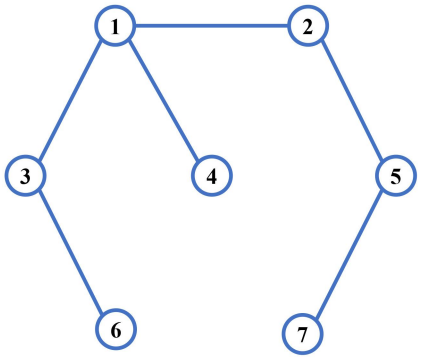


Fig. 3. Communication topology that dictates the information exchange between the agents.

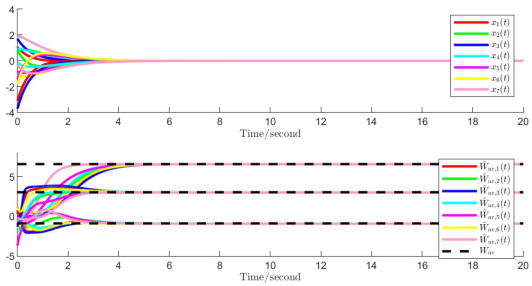


Fig. 4. Policy evaluation with the proposed cooperative critic learning.

$\sum_{i=1}^{N}\int_0^t \psi_{i,f}^a(t,\tau)[\psi_{i,f}^a(t,\tau)]^T \mathrm{d}\tau$. Therefore, with sufficient excited historical data, the critic learning could converge with satisfactory speed. In addition, as the function approximator dimension increases, the value function approximation residual would decrease accordingly, which could further improve the performance of critic learning.

Remark 3: As shown in the filters design (16) and (23), the distributed critic learning requires the knowledge of system dynamics because $f(\cdot)$, $g(\cdot)$, and $h(\cdot)$ are involved. Therefore, the presented distributed critic learning relies on the system model. Model-free extension of the distributed critic learning can be established by utilizing additional learning-based identifier as investigated in [36] and [50].

VI. Simulations

In this section, we apply the cooperative composite actor–critic design to the zero-sum game with two cases.

Case 1: In this example, we consider the dynamical systems in the form of (1) as follows:

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\end{bmatrix}=\begin{bmatrix}0 & 1\\ -2 & 3\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix}+\begin{bmatrix}0\\ 1\end{bmatrix}u+\begin{bmatrix}1\\ 0\end{bmatrix}v\ \forall t\geq 0$$

where $x=[x_1\ x_2]^{\mathrm{T}}$ is the system state, u is the maximizing player, and v is the minimizing player. For the min–max zero-sum differential game, the reward function of interest is considered as $r(x,u,v)=x^{\mathrm{T}}Qx+u^{\mathrm{T}}Ru-\gamma^2\|v\|^2$ with $Q=I_2$, $R=0.5$, and $\gamma=0.2$.

In this example, we investigate the policy evaluation problem to solve the Bellman equation (5) with a given pair of policies $\{u,v\}$. The policy pair under consideration is $u=[1.85\quad -3.8]\begin{bmatrix}x_1\\ x_2\end{bmatrix}$ and $v=[-2.2\quad 0.6]\begin{bmatrix}x_1\\ x_2\end{bmatrix}$. Based on the linear optimal control theory [16], the corresponding value function can be explicitly expressed as $V(x)=x^{\mathrm{T}}Px$ with $P=\begin{bmatrix}0.4638 & -0.9918\\ -0.9918 & 3.1090\end{bmatrix}$. In addition, with the polynomial basis function $\phi(x)=[x_1^2\quad x_1x_2\quad x_2^2]^{\mathrm{T}}$, the value function $V(x)=x^{\mathrm{T}}Px$ can be equivalently expressed as $V(x)=W^T\phi(x)$ with $W=[0.4638\quad -1.9836\quad 3.1090]^T$.

As shown in Fig. 2, one can observe that, with the conventional critic learning [26], the critic weight vector could not converge to the desired values because the PE condition is not satisfied. In contrast, to implement the presented cooperative critic learning, the interaction among the multiple agents are shown in Fig. 3, from which one can observe that the multiple agents learn to solve the Bellman equation through local interaction, which is different from the centralized learning mechanism. The evolution of multiple agents states with the cooperative critic learning start from distinct initial conditions, as shown in the top of Fig. 4. With the value function parameterization using polynomial basis function, the group of multiple agents learn the desired critic weight vector $W=[0.4638\quad -1.9836\quad 3.1090]^T$ in a distributed fashion through local interaction. The cooperative critic learning process is shown in the bottom of Fig. 4, where the cooperative critic learning achieve consensus on the desired weight vector $W=[0.4638\quad -1.9836\quad 3.1090]^T$ within 5 s.

Collecting the results in Figs. 2 and 4, the strict PE condition for convergence of conventional critic learning with single agent [26] can be relaxed by the cooperative critic learning with multiple agents. In addition, with the presented cooperative critic learning design, the probing noise to guarantee the PE condition is also avoided to maintain smooth system operation.

Case 2: In this case, we consider an affine nonlinear dynamical system in the form (1) with

$$f(x)=\begin{bmatrix}-x_1+x_1x_2^2\\ -x_2\end{bmatrix},\ g(x)=\begin{bmatrix}0\\ x_1\end{bmatrix},\ h(x)=\begin{bmatrix}0\\ x_1\end{bmatrix}.$$

For the zero-sum differential game, the reward function of interest is considered as $r(x,u,v)=x^{\mathrm{T}}Qx+u^{\mathrm{T}}Ru-\gamma^2\|v\|^2$ with minimizing player u, maximizing player v, $Q=2I_2$, $R=0.4902$, and $\gamma=5$. With the inverse optimal design procedure investigated in [51], the optimal value function is $V^\star(x)=x_1^2+x_2^2$, and the optimal control policy and worst-case disturbance policy are $u^\star=-2.04x_1x_2$ and $v^\star=-0.04x_1x_2$,

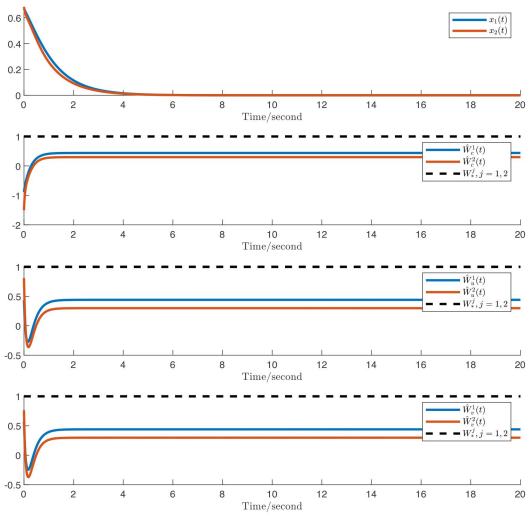


Fig. 5. Evolution of system state and actor–critic learning with conventional design [26] for nonlinear H_∞ control.

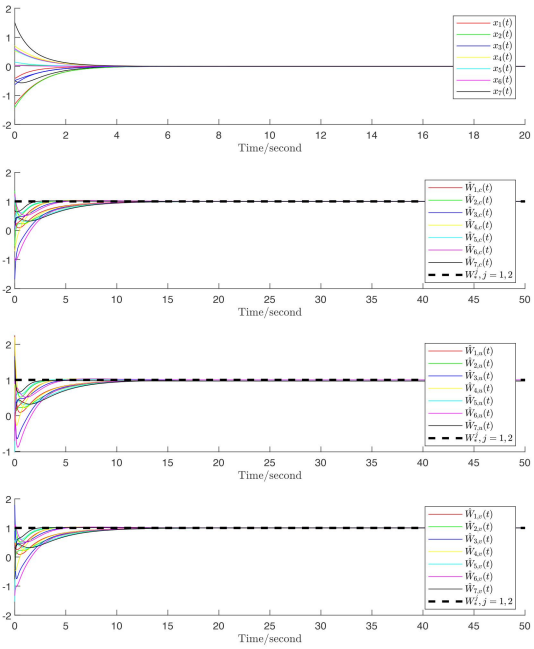


Fig. 6. Evolution of system state and actor–critic learning with presented design for nonlinear H_∞ control.

respectively. In addition, the optimal value function can be parameterized using the polynomial basis function $\phi(x) = [x_1^2 \quad x_2^2]^{\mathrm{T}}$ as $V^\star(x) = W_\star^{\mathrm{T}}\phi(x)$ with the ideal weight vector $W_\star = [1 \quad 1]^{\mathrm{T}}$.

We first apply the conventional actor–critic learning [26], where the results are shown in Fig. 5, from which one can observe that the conventional actor–critic learning does not converge to the desired optimal weight vectors because the PE condition is not satisfied during the online learning process.

As a comparison, we then apply the presented cooperative actor–critic learning with multiple learning agents discussed in Section V, where the communication topology among the agents is the same as in Fig. 3. As shown in the top of Fig. 6, the multiple states evolution starts from distinct initial conditions. The critic learning process of the cooperative critic learning is illustrated in the second row of Fig. 6, where the adaptive critic weights of each agent reach consensus on the desired critic weight vector $W_\star = [1 \quad 1]^{\mathrm{T}}$. Under the cFE condition, the online learning of the maximizing player u_i and the minimizing player v_i for each agent is depicted in the third and fourth row in Fig. 6, respectively. From Fig. 6, the actor weights of all the agents could be synchronized to the Nash equilibrium policies $\{u^\star, v^\star\}$.

VII. Conclusion

A novel actor–critic learning algorithm is developed to find the Nash equilibrium of the H_∞ zero-sum game of systems evolving in continuous time. The online critic learning is designed based on the experience replay method by the combination of historical and current data. In the cooperative actor–critic leaning, the local interaction among agents contributes to relax the stringent PE condition, which is difficult to be satisfied and verified. In contrast, in the distributed actor–critic learning, the relaxed cooperative excitation condition can be verified during online learning and ensure the exponent convergence to the Nash equilibrium of zero-sum games. Theoretical analysis is investigated about the stability for the adaptive learning system and the convergence of the actor–critic networks to the solution of the HJI equation for the H_∞ control problem. Application of the presented cooperative actor–critic learning to the policy evaluation problem and policy optimization are conducted in the simulation study.

Future work will extend the presented design to non-zero-sum games.

Appendix A
Proof of Lemma 1

1) The first proposition can be referred to [43].

2) First, denote λ_i with $\lambda_i \leq \lambda_{i+1}$ as the eigenvalues of $\mathcal{L}\otimes I_p$, and w_i as the unit eigenvectors of $\mathcal{L}\otimes I_p$ corresponding to eigenvalue λ_i, for $i = 1, \ldots, Np$. In addition, ζ_i represents the ith column of I_p for $i = 1, \ldots, p$. From Assumption 3 and the property of the Laplacian matrix $\mathcal{L}$, one has $\lambda_i = 0$ and w_i satisfies $w_i = 1/\sqrt{N}1_N \otimes \zeta_i$ for $i = 1, \ldots, p$. For $i \geq p+1$, one has $\lambda_i > 0$. In addition, any nonzero vector $\xi \in \mathbb{R}^{Np}$ can be expressed as $\xi = \sum_{i=1}^{Np} c_i w_i$, with c_i being constants satisfying $\sum_{i=1}^{Np} c_i^2 > 0$. In the following, we investigate two different cases.

Case 1 ($\sum_{i=p+1}^{Np} c_i^2 > 0$): For the matrix $G(t)+(\mathcal{L}\otimes I_p)$, one has $\xi^T[G(t) + (L \otimes I_p)]\xi = \xi^T G(t)\xi + \sum_{i=p+1}^{Np} c_i^2\lambda_i w_i^T w_i \geq \sum_{i=p+1}^{Np} c_i^2\lambda_i w_i^T w_i > 0 \quad \forall t \geq t_s$, which implies that the matrix $G(t) + (\mathcal{L} \otimes I_p)$ is positive definite $\forall t \geq t_s$.

Case 2 ($\sum_{i=p+1}^{Np} c_i^2 = 0$): For $\xi \neq 0$, the fact $\sum_{i=1}^{Np} c_i^2 > 0$ implies that $\sum_{i=1}^{p} c_i^2 > 0$. Denote $c = [c_1^{\mathrm{T}}, \ \ldots, \ c_p^{\mathrm{T}}]^T$ and $W = [w_1, \ \ldots, \ w_p] = 1_N \otimes I_p$. Then, $\xi^T[G(t) + (L \otimes I_p)]\xi = c^T W^{\mathrm{T}} G(t) W c = 1/N c^T \sum_{i=1}^{N} G_i(t) c > 0 \quad \forall t \geq t_s$, where

the inequality is guaranteed by condition (11). Therefore, the matrix $G(t)+(\mathcal{L}\otimes I_p)\succ 0$ for all $t\geq t_s$ in the second case.

To further show that there exists a positive constant $\bar{\beta}$ such that $G(t)+(\mathcal{L}\otimes I_p)\geq \bar{\beta} I_{Np}\quad \forall t\geq t_s$, it is referred to [43]. This completes the proof.

Appendix B
Proof of Theorem 1

1) The condition (22) is the Type-I cFE condition and the first conclusion can be referred to the first proposition in Lemma 1.

2) This conclusion is a standard result in adaptive control and can be referred to [52].

3) From the facts in (18) and (19), one has $\dot{\tilde{W}}_i(t)=-\gamma\{[\delta_{i,f}(t)-E_{i,f}(t)\cdot\tilde{W}]-\sum_{j=1}^{N}a_{ij}[\tilde{W}_i(t)-\tilde{W}_j(t)]\}$. Consider the Lyapunov candidate as $V(\tilde{W}(t))=(1/2)\sum_{i=1}^{N}\tilde{W}_i^{\mathrm{T}}(t)\gamma^{-1}\tilde{W}_i(t)$. Differentiating $V(\tilde{W}(t))$ yields

$$\dot{V}\big(\tilde{W}(t)\big)=-\tilde{W}^T(t)\Big[E^f(t)+(L\otimes I_n)\Big]\tilde{W}(t)+\tilde{W}^T(t)\delta_f(t) \tag{34}$$

where $\delta_f(t)=[\,\delta_{1,f}^T(t)\cdots\delta_{N,f}^T(t)\,]^T$ and $\tilde{W}(t)=[\tilde{W}_1^T(t).\cdots.\tilde{W}_N^T(t)]^T$. From Assumption 3, the filtered signal $\delta_f(t)$ is bounded in the sense that $\|\delta_f(t)\|\leq\bar{\delta}_f$ with constant $\bar{\delta}_f>0$. From the first proposition, $\dot{V}(\tilde{W}(t))$ further satisfies

$$\begin{aligned}\dot{V}\big(\tilde{W}(t)\big)&\leq\bar{\delta}_f\big\|\tilde{W}(t)\big\|-\alpha_c\big\|\tilde{W}(t)\big\|^2\\&=-(1-\beta)\alpha_c\big\|\tilde{W}(t)\big\|^2-\beta\alpha_c\big\|\tilde{W}(t)\big\|^2+\bar{\delta}_f\big\|\tilde{W}(t)\big\|\end{aligned} \tag{35}$$

with $\beta\in(0,1)$. With Young's inequality $2ab-b^2\leq a^2$, one has $-\beta\alpha_c\|\tilde{W}(t)\|^2+\bar{\delta}_f\|\tilde{W}(t)\|\leq(\bar{\delta}_f)^2/4\beta\alpha_c$. Then, $\dot{V}(\tilde{W}(t))\leq -a_cV(\tilde{W}(t))+b_c$ with $a_c=(1-\beta)\alpha_c\gamma$ and $b_c=(\bar{\delta}_f)^2/4\beta\alpha_c$. Based on the standard Lyapunov extension theorem [49], the conclusion holds. This completes the proof.

Appendix C
Proof of Theorem 2

1) This can be referred to the second proposition in Lemma 1.

2) This conclusion is a standard result in adaptive control and can be referred to [52].

3) Consider the Lyapunov candidate for all the agents as $\mathcal{V}(\chi)=\sum_{i=1}^{N}L(\chi_i)$ with $L(\chi_i)=L_c(\tilde{W}_{i,c})+L_u(\tilde{W}_{i,u})+L_v(\tilde{W}_{i,v})+V^\star(x_i)$, where $L_c(\tilde{W}_{i,c})=(1/2)\tilde{W}_{i,c}^T\gamma_c^{-1}\tilde{W}_{i,c}$, $L_u(\tilde{W}_{i,u})=(1/2)\tilde{W}_{i,u}^T\gamma_u^{-1}\tilde{W}_{i,u}$, $L_v(\tilde{W}_{i,v})=(1/2)\tilde{W}_{i,v}^T\gamma_v^{-1}\tilde{W}_{i,v}$, $\chi_i=[\,x_i^{\mathrm{T}}\ \tilde{W}_{i,c}^{\mathrm{T}}\ \tilde{W}_{i,u}^{\mathrm{T}}\ \tilde{W}_{i,v}^{\mathrm{T}}\,]^{\mathrm{T}}$, and $\chi=[\,\chi_1^{\mathrm{T}}\cdots\chi_N^{\mathrm{T}}\,]^T$. First, the agent dynamics with distributed online actors can be denoted as follows:

$$\dot{x}_i=f(x_i)-\frac{1}{2}R^{-1}g^T(x_i)[\nabla\phi(x_i)]^T\hat{W}_{i,u}(t)+\frac{1}{2\gamma^2}h^T(x_i)[\nabla\phi(x_i)]^T\hat{W}_{i,v}(t). \tag{36}$$

With the fact in (14), differentiating $V^\star(x_i)$ along above dynamics yields

$$\begin{aligned}\dot{V}^\star(x_i)=&-Q(x_i)-\frac{1}{4}W_\star^T\Gamma(x_i)W_\star+\frac{1}{4\gamma^2}W_\star^T\Xi(x_i)W_\star\\&+\frac{1}{2}W_\star^T\Gamma_i(t)\tilde{W}_{i,u}(t)-\frac{1}{2\gamma^2}W_\star^T\Xi_i(t)\tilde{W}_{i,v}(t)\\&+\sigma^\star(x_i)+\zeta_i\end{aligned} \tag{37}$$

with

$$\zeta_i=\big[\nabla\epsilon^\star(x_i)\big]^{\mathrm{T}}\Big\{f(x_i)-\frac{1}{2}R^{-1}g^T(x_i)[\nabla\phi(x_i)]^T\hat{W}_{i,u}(t)+\frac{1}{2\gamma^2}h^T(x_i)[\nabla\phi(x_i)]^T\hat{W}_{i,v}(t)\Big\}. \tag{38}$$

According to Assumption 4, the term ζ_i satisfies

$$\begin{aligned}\|\zeta_i\|\leq&\,\eta_{d\epsilon}\eta_f\cdot\|x_i\|+\frac{1}{2}\eta_{d\epsilon}\eta_g^2\eta_{d\phi}\lambda_{\min}(R)\big\|\tilde{W}_{i,u}\big\|\\&+\frac{1}{2\gamma^2}\eta_{d\epsilon}\eta_h^2\eta_{d\phi}\big\|\tilde{W}_{i,v}\big\|+\frac{1}{2}\eta_{d\epsilon}\eta_g^2\eta_{d\phi}\eta_W\lambda_{\min}(R)\\&+\frac{1}{2\gamma^2}\eta_{d\epsilon}\eta_h^2\eta_{d\phi}\eta_W.\end{aligned} \tag{39}$$

Recall the critic learning (28) and the fact that $\partial J_{i,1}^c(\hat{W}_{i,c}(t))/\partial\hat{W}_{i,c}(t)=A_{i,c}(t)\cdot\hat{W}_{i,c}(t)+D_{i,c}(t)$, one has

$$\dot{\hat{W}}_{i,c}(t)=-\gamma_c\frac{1}{\big[B_{i,c}(t)\big]^2}\int_0^t\psi_{i,f}^a(t,\tau)e_{i,c}(t,\tau)\mathrm{d}\tau+\gamma_c\sum_{j=1}^{N}a_{ij}\Big[\hat{W}_{j,c}(t)-\hat{W}_{i,c}(t)\Big]. \tag{40}$$

From the definitions $\psi_{i,f}^a(t,\tau)$ and $r_{i,f}^a(t,\tau)$ and the fact in (25), one has

$$\begin{aligned}e_{i,c}(t,\tau)=&-\tilde{W}_{i,c}^T(t)\psi_{i,f}^a(t,\tau)+\sigma_{i,f}^\star(\tau)\\&+\frac{1}{4}\tilde{W}_{i,u}^T(t)\Gamma_{i,f}(\tau)\tilde{W}_{i,u}(t)\\&-\frac{1}{4\gamma^2}\tilde{W}_{i,v}^T(t)\Xi_{i,f}(\tau)\tilde{W}_{i,v}(t).\end{aligned} \tag{41}$$

Then, taking $e_{i,c}(t,\tau)$ into the critic weight update (40) yields

$$\begin{aligned}\dot{\hat{W}}_{i,c}(t)=&\,\gamma_c\sum_{j=1}^{N}a_{ij}\Big[\hat{W}_{j,c}(t)-\hat{W}_{i,c}(t)\Big]\\&+\gamma_c\frac{1}{\big[B_{i,c}(t)\big]^2}\int_0^t\Big\{\psi_{i,f}^a(t,\tau)\Big[\psi_{i,f}^a(t,\tau)\Big]^{\mathrm{T}}\Big\}\mathrm{d}\tau\tilde{W}_{i,c}(t)\\&-\gamma_c\frac{1}{\big[B_{i,c}(t)\big]^2}\int_0^t\Big[\psi_{i,f}^a(t,\tau)\Big]\mathrm{d}\tau\sigma_{i,f}^\star(t)\\&-\frac{\gamma_c}{4}\frac{1}{\big[B_{i,c}(t)\big]^2}\int_0^t\Big[\psi_{i,f}^a(t,\tau)\tilde{W}_{i,u}^{\mathrm{T}}(t)\Gamma_{i,f}(\tau)\tilde{W}_{i,u}(t)\Big]\mathrm{d}\tau\\&+\frac{\gamma_c}{4\gamma^2\big[B_{i,c}(t)\big]^2}\int_0^t\Big[\psi_{i,f}^a(t,\tau)\tilde{W}_{i,v}^T(t)\Xi_{i,f}(\tau)\tilde{W}_{i,v}(t)\Big]\mathrm{d}\tau.\end{aligned}$$

Therefore, differentiating $L_c(\tilde{W}_{i,c}) = (1/2)\tilde{W}_{i,c}^T\gamma_c^{-1}\tilde{W}_{i,c}$ yields

$$\begin{aligned}\dot{L}_c(t) = &-\tilde{W}_{i,c}^T(t)\sum_{j=1}^{N} a_{ij}\Big[\hat{W}_{j,c}(t) - \hat{W}_{i,c}(t)\Big] \\ &- \tilde{W}_{i,c}^T(t)\frac{\int_0^t \Big\{\psi_{i,f}^a(t,\tau)\Big[\psi_{i,f}^a(t,\tau)\Big]^T\Big\}\mathrm{d}\tau}{\big[B_{i,c}(t)\big]^2}\tilde{W}_{i,c}(t) \\ &+ \tilde{W}_{i,c}^T(t)\frac{\int_0^t \Big[\psi_{i,f}^a(t,\tau)\Big]\mathrm{d}\tau}{\big[B_{i,c}(t)\big]^2}\sigma_{i,f}^\star(t) \\ &+ \frac{\Lambda_{i,u}(t)}{4\big[B_{i,c}(t)\big]^2} - \frac{\Lambda_{i,v}(t)}{4\gamma^2\big[B_{i,c}(t)\big]^2}\end{aligned}$$

where $\Lambda_{i,u}$ and $\Lambda_{i,v}(t)$ are defined in (42), at the bottom of the page. Considering the actor learning (31) for the minimizing player u_i, differentiating $L_u(\tilde{W}_{i,u}) = (1/2)\tilde{W}_{i,u}^T\gamma_u^{-1}\tilde{W}_{i,u}$ yields

$$\begin{aligned}\dot{L}_u(t) = &\tilde{W}_{i,u}^T(t)\sum_{j=1}^{N} a_{ij}\big[\tilde{W}_{j,u}(t) - \tilde{W}_{i,u}(t)\big] \\ &+ \tilde{W}_{i,u}^T(t)K_{i,\mathrm{uu}}W_\star - \tilde{W}_{i,u}^T(t)K_{i,\mathrm{uu}}\tilde{W}_{i,u}(t) \\ &+ \tilde{W}_{i,u}^T(t)K_{i,\mathrm{cu}}\frac{\Big[\int_0^t \psi_{i,f}^a(t,\tau)\mathrm{d}\tau\Big]^T}{\big[B_{i,c}(t)\big]^2}\tilde{W}_{i,c}(t) \\ &- \tilde{W}_{i,u}^T(t)K_{i,\mathrm{cu}}\frac{\Big[\int_0^t \psi_{i,f}^a(t,\tau)\mathrm{d}\tau\Big]^T}{\big[B_{i,c}(t)\big]^2}W_\star - \frac{\tilde{W}_{i,u}^T(t)F_{i,u}(t)}{4\big[B_{i,c}(t)\big]^2}.\end{aligned} \tag{43}$$

Similarly, for the maximizing player v_i, one has

$$\begin{aligned}\dot{L}_v(\tilde{W}_{i,v}) = &\tilde{W}_{i,v}^T(t)\sum_{j=1}^{N} a_{ij}\big[\tilde{W}_{j,v}(t) - \tilde{W}_{i,v}(t)\big] \\ &+ \tilde{W}_{i,v}^T(t)K_{i,\mathrm{vv}}W_\star - \tilde{W}_{i,v}^T(t)K_{i,\mathrm{vv}}\tilde{W}_{i,v}(t) \\ &+ \tilde{W}_{i,v}^T(t)K_{i,\mathrm{cv}}\frac{\Big[\int_0^t \psi_{i,f}^a(t,\tau)\mathrm{d}\tau\Big]^T}{\big[B_{i,c}(t)\big]^2}\tilde{W}_{i,c}(t) \\ &- \tilde{W}_{i,v}^T(t)K_{i,\mathrm{cv}}\frac{\Big[\int_0^t \psi_{i,f}^a(t,\tau)\mathrm{d}\tau\Big]^T}{\big[B_{i,c}(t)\big]^2}W_\star \\ &+ \tilde{W}_{i,v}^T(t)\frac{F_{i,v}(t)}{4\gamma^2\big[B_{i,c}(t)\big]^2}.\end{aligned} \tag{44}$$

Collecting (37), (47), (43), and (44), one has $\dot{L}(\chi_i) = -\chi_i^{\mathrm{T}}P_i\chi_i + \chi_i^{\mathrm{T}}p_i + c_i + d_i$, with $d_i = \tilde{W}_{i,c}^T(t)\sum_{j=1}^N a_{ij}[\tilde{W}_{j,c}(t) - \tilde{W}_{i,c}(t)] + \tilde{W}_{i,u}^T(t)\sum_{j=1}^N$

$$\begin{aligned}\Lambda_{i,u}(t) = &-\int_0^t \tilde{W}_{i,u}^T(t)\Gamma_{i,f}(\tau)W_\star\Big[\psi_{i,f}^a(t,\tau)\Big]^T W_\star\mathrm{d}\tau + \int_0^t \tilde{W}_{i,u}^T(t)\Gamma_{i,f}(\tau)W_\star\Big[\psi_{i,f}^a(t,\tau)\Big]^T\tilde{W}_{i,c}(t)\mathrm{d}\tau \\ &+ \int_0^t \tilde{W}_{i,u}^T(t)\Gamma_{i,f}(\tau)\tilde{W}_{i,u}(t)\Big[\psi_{i,f}^a(t,\tau)\Big]^T W_\star\mathrm{d}\tau + \tilde{W}_{i,u}^T(t)F_{i,u}(t) \\ \Lambda_{i,v}(t) = &-\int_0^t \tilde{W}_{i,v}^T(t)\Xi_{i,f}(\tau)W_\star\Big[\psi_{i,f}^a(t,\tau)\Big]^T W_\star\mathrm{d}\tau + \int_0^t \tilde{W}_{i,v}^T(t)\Xi_{i,f}(\tau)W_\star\Big[\psi_{i,f}^a(t,\tau)\Big]^T\tilde{W}_{i,c}(t)\mathrm{d}\tau \\ &+ \int_0^t \tilde{W}_{i,v}^T(t)\Xi_{i,f}(\tau)\tilde{W}_{i,v}(t)\Big[\psi_{i,f}^a(t,\tau)\Big]^T W_\star\mathrm{d}\tau + \tilde{W}_{i,v}^T(t)F_{i,v}(t)\end{aligned} \tag{42}$$

$$\begin{aligned}&P_{i,\mathrm{uu}} = K_{i,\mathrm{uu}} - \frac{\int_0^t \Gamma_{i,f}(\tau)\Big[\psi_{i,f}^a(t,\tau)\Big]^T W_\star\mathrm{d}\tau}{4\big[B_{i,c}(t)\big]^2}, P_{i,\mathrm{uc}} = P_{i,\mathrm{cu}}^{\mathrm{T}} = -\frac{\int_0^t \Gamma_{i,f}(\tau)W_\star\Big[\psi_{i,f}^a(t,\tau)\Big]^T\mathrm{d}\tau}{8\big[B_{i,c}(t)\big]^2} - K_{i,\mathrm{cu}}\frac{\Big[\int_0^t \psi_{i,f}^a(t,\tau)\mathrm{d}\tau\Big]^T}{2\big[B_{i,c}(t)\big]^2} \\ &P_{i,\mathrm{vv}} = K_{i,\mathrm{vv}} + \frac{\int_0^t \Xi_{i,f}(\tau)\Big[\psi_{i,f}^a(t,\tau)\Big]^T W_\star\mathrm{d}\tau}{4\gamma^2\big[B_{i,c}(t)\big]^2}, P_{i,\mathrm{vc}} = P_{i,\mathrm{cv}}^{\mathrm{T}} = \frac{\int_0^t \Xi_{i,f}(\tau)W_\star\Big[\psi_{i,f}^a(t,\tau)\Big]^T\mathrm{d}\tau}{8\gamma^2\big[B_{i,c}(t)\big]^2} - K_{i,\mathrm{cv}}\frac{\Big[\int_0^t \psi_{i,f}^a(t,\tau)\mathrm{d}\tau\Big]^T}{2\big[B_{i,c}(t)\big]^2} \\ &p_{i,x} = \eta_{d\epsilon}\eta_f, p_{i,c} = \frac{\int_0^t \Big[\psi_{i,f}^a(t,\tau)\Big]\mathrm{d}\tau}{4\big[B_{i,c}(t)\big]^2}\sigma_{i,f}^\star(t), P_{i,\mathrm{xx}} = qI_{n\times n}, P_{i,\mathrm{cc}} = \frac{\int_0^t \Big\{\psi_{i,f}^a(t,\tau)\Big[\psi_{i,f}^a(t,\tau)\Big]^T\Big\}\mathrm{d}\tau}{\big[B_{i,c}(t)\big]^2} \\ &p_{i,u} = K_{i,\mathrm{uu}}W_\star + \frac{1}{2}\Gamma_i(t)W_\star + \frac{1}{2}\eta_{d\epsilon}\eta_g^2\eta_{d\phi}\lambda_{\min}(R) - K_{i,\mathrm{cu}}\frac{\Big[\int_0^t \psi_{i,f}^a(t,\tau)\mathrm{d}\tau\Big]^T}{\big[B_{i,c}(t)\big]^2}W_\star - \frac{\int_0^t \Gamma_{i,f}(\tau)W_\star\Big[\psi_{i,f}^a(t,\tau)\Big]^T W_\star\mathrm{d}\tau}{4\big[B_{i,c}(t)\big]^2} \\ &p_{i,v} = K_{i,\mathrm{vv}}W_\star - \frac{1}{2\gamma^2}\Xi_i(t)W_\star + \frac{1}{2\gamma^2}\eta_{d\epsilon}\eta_h^2\eta_{d\phi} - K_{i,\mathrm{cv}}\frac{\Big[\int_0^t \psi_{i,f}^a(t,\tau)\mathrm{d}\tau\Big]^T}{\big[B_{i,c}(t)\big]^2}W_\star + \frac{\int_0^t \Xi_{i,f}(\tau)W_\star\Big[\psi_{i,f}^a(t,\tau)\Big]^T W_\star\mathrm{d}\tau}{4\gamma^2\big[B_{i,c}(t)\big]^2} \\ &c_i = -\frac{1}{4}W_\star^T\Gamma(x_i)W_\star + \frac{1}{4\gamma^2}W_\star^T\Xi(x_i)W_\star + \sigma^\star(x_i) + \frac{1}{2}\eta_{d\epsilon}\eta_g^2\eta_{d\phi}\eta_W\lambda_{\min}(R) + \frac{1}{2\gamma^2}\eta_{d\epsilon}\eta_h^2\eta_{d\phi}\eta_W\end{aligned} \tag{46}$$

$a_{ij}[\tilde{W}_{j,u}(t)-\tilde{W}_{i,u}(t)]+\tilde{W}_{i,v}^{T}(t)\sum_{j=1}^{N}a_{ij}[\tilde{W}_{j,v}(t)-\tilde{W}_{i,v}(t)]$ and

$$P_i=\begin{bmatrix} qI_{n\times n} & 0 & 0 & 0\\ 0 & P_{i,\mathrm{cc}} & P_{i,\mathrm{cu}} & P_{i,\mathrm{cv}}\\ 0 & P_{i,\mathrm{uc}} & P_{i,\mathrm{uu}} & 0\\ 0 & P_{i,\mathrm{vc}} & 0 & P_{i,\mathrm{vv}}\end{bmatrix},\ p_i=\begin{bmatrix} p_{i,x}\\ p_{i,c}\\ p_{i,u}\\ p_{i,v}\end{bmatrix} \tag{45}$$

with notations in (46), at the bottom of the previous page. Denote

$$\begin{aligned}\tilde{W}_c(t)&\coloneqq\begin{bmatrix}\tilde{W}_{1,c}^{\mathrm{T}}(t)\cdots\tilde{W}_{N,c}^{\mathrm{T}}(t)\end{bmatrix}^{\mathrm{T}}\\ \tilde{W}_u(t)&\coloneqq\begin{bmatrix}\tilde{W}_{1,u}^{\mathrm{T}}(t)\cdots\tilde{W}_{N,u}^{\mathrm{T}}(t)\end{bmatrix}^{\mathrm{T}}\\ \tilde{W}_v(t)&\coloneqq\begin{bmatrix}\tilde{W}_{1,v}^{\mathrm{T}}(t)\cdots\tilde{W}_{N,v}^{\mathrm{T}}(t)\end{bmatrix}^{\mathrm{T}}\\ \bar{x}&\coloneqq\begin{bmatrix}x_1^{\mathrm{T}}(t)\cdots x_N^{\mathrm{T}}(t)\end{bmatrix}^{\mathrm{T}}.\end{aligned} \tag{47}$$

Note that $\sum_{i=1}^{N}d_i=-\tilde{W}_c^{\mathrm{T}}(L\otimes I_p)\tilde{W}_c-\tilde{W}_u^{\mathrm{T}}(L\otimes I_p)\tilde{W}_u-\tilde{W}_v^{\mathrm{T}}(L\otimes I_p)\tilde{W}_v\le 0$. Therefore

$$\sum_{i=1}^{N}\dot{L}(\chi_i)\le\sum_{i=1}^{N}\left(-\chi_i^{T}P_i\chi_i+\chi_i^{T}p_i+c_i\right). \tag{48}$$

Finally, one has $\dot{V}(\chi)\le-\chi^{\mathrm{T}}P\chi+\chi^{\mathrm{T}}p+c$, where

$$\begin{aligned}P_{\mathrm{cc}}&=\mathrm{diag}\left(P_{i,cc},\ldots,P_{N,\mathrm{cc}}\right)\\ P_{\mathrm{uu}}&=\mathrm{diag}\left(P_{i,uu},\ldots,P_{N,\mathrm{uu}}\right)\\ P_{\mathrm{vv}}&=\mathrm{diag}\left(P_{i,vv},\ldots,P_{N,\mathrm{vv}}\right)\\ P_{\mathrm{cu}}&=\mathrm{diag}\left(P_{i,cu},\ldots,P_{N,\mathrm{cu}}\right)\\ P_{\mathrm{uc}}&=\mathrm{diag}\left(P_{i,uc},\ldots,P_{N,\mathrm{uc}}\right)\\ P_{\mathrm{cv}}&=\mathrm{diag}\left(P_{i,cv},\ldots,P_{N,\mathrm{cv}}\right)\\ P_{\mathrm{vc}}&=\mathrm{diag}\left(P_{i,\mathrm{vc}},\ldots,P_{N,\mathrm{vc}}\right),\ c=\sum_{i=1}^{N}c_i\\ p_x&=\begin{bmatrix}p_{1,x}^{T}\cdots p_{N,x}^{T}\end{bmatrix}^{T},\ p_c=\begin{bmatrix}p_{1,c}^{T}\cdots p_{N,c}^{T}\end{bmatrix}^{T}\\ p_u&=\begin{bmatrix}p_{1,u}^{T}\cdots p_{N,u}^{T}\end{bmatrix}^{T},\ p_v=\begin{bmatrix}p_{1,v}^{T}\cdots p_{N,v}^{T}\end{bmatrix}^{T}\\ P&=\begin{bmatrix}qI_{nN\times nN} & 0 & 0 & 0\\ 0 & P_{\mathrm{cc}} & P_{\mathrm{cu}} & P_{\mathrm{cv}}\\ 0 & P_{\mathrm{uc}} & P_{\mathrm{uu}} & 0\\ 0 & P_{\mathrm{vc}} & 0 & P_{\mathrm{vv}}\end{bmatrix},\ p=\begin{bmatrix}p_x\\ p_c\\ p_u\\ p_v\end{bmatrix}.\end{aligned}$$

Note that the matrix P can be rewritten in a block-wise form as follows:

$$P=\begin{bmatrix}qI_{nN\times nN} & 0 & 0\\ 0 & P_{\mathrm{cc}} & P_{\mathrm{ca}}\\ 0 & P_{\mathrm{ac}} & P_{\mathrm{aa}}\end{bmatrix} \tag{49}$$

with $P_{\mathrm{aa}}=\mathrm{diag}[P_{\mathrm{uu}},P_{\mathrm{vv}}]$, $P_{\mathrm{ca}}=P_{\mathrm{ac}}^{\mathrm{T}}=[P_{\mathrm{cu}}\ P_{\mathrm{cv}}]$. Since $\sum_{i=1}^{N}\int_0^t\psi_{i,f}^{a}(t,\tau)[\psi_{i,f}^{a}(t,\tau)]^{T}\mathrm{d}\tau\ge\alpha\cdot I_{M\times M}$, the matrix P_{cc} is positive definite. In addition, the Shur complement of P_{cc} is positive definite, that is, $P_{\mathrm{cc}}-P_{\mathrm{ca}}P_{\mathrm{aa}}^{-1}P_{\mathrm{ac}}\succ 0$, provided that $K_{i,\mathrm{cu}}\ll K_{i,\mathrm{uu}}$ and $K_{i,\mathrm{cv}}\ll K_{i,\mathrm{vv}}$. Combined with the facts that $q>0$, then, the matrix $P\succ 0$, that is, $\lambda_{\min}(P)>0$. From Assumption 4, one has $\|p\|\le p_{\max}$ and $\|c\|\le c_{\max}$. Therefore, one has $\dot{\mathcal{V}}(\chi)\le-\lambda_{\min}(P)\cdot\|\chi\|^2+p_{\max}\cdot\|\chi\|+c_{\max}$ and $\dot{\mathcal{V}}(\chi)$ is negative if the augment variable χ exceeds the set $\{\chi|\|\chi\|>b_\chi.\}$ with $b_\chi\triangleq p_{\max}/2\lambda_{\min}(P)+\sqrt{p_{\max^2}/4\lambda_{\max^2}(P)+c_{\max}/\lambda_{\min}(P)}$. According to the standard Lyapunov extension theorem [49], the agents' states $x_i(t)$ and the actor–critic weights learning error $\tilde{W}_{i,c}(t),\tilde{W}_{i,u}(t)$, and $\tilde{W}_{i,v}(t)$ are UUB. This completes the proof.

REFERENCES

[1] V. G. Lopez, F. L. Lewis, Y. Wan, E. N. Sanchez, and L. Fan, "Solutions for multiagent pursuit-evasion games on communication graphs: Finite-time capture and asymptotic behaviors," *IEEE Trans. Autom. Control*, vol. 65, no. 5, pp. 1911–1923, May 2020.

[2] K. G. Vamvoudakis, F. L. Lewis, and G. R. Hudas, "Multi-agent differential graphical games: Online adaptive learning solution for synchronization with optimality," *Automatica*, vol. 48, no. 8, pp. 1598–1611, 2012.

[3] S. Manfredi, "Design of a multi-hop dynamic consensus algorithm over wireless sensor networks," *Control Eng. Pract.*, vol. 21, no. 4, pp. 381–394, 2013.

[4] R. M. Murray, "Recent research in cooperative control of multivehicle systems," *J. Dyn. Syst. Meas. Control*, vol. 129, no. 5, pp. 571–583, May 2007.

[5] W. Chen, C. Wen, S. Hua, and C. Sun, "Distributed cooperative adaptive identification and control for a group of continuous-time systems with a cooperative PE condition via consensus," *IEEE Trans. Autom. Control*, vol. 59, no. 1, pp. 91–106, Jan. 2014.

[6] H. Zhang, H. Su, K. Zhang, and Y. Luo, "Event-triggered adaptive dynamic programming for non-zero-sum games of unknown nonlinear systems via generalized fuzzy hyperbolic models," *IEEE Trans. Fuzzy Syst.*, vol. 27, no. 11, pp. 2202–2214, Nov. 2019.

[7] G. Zames, "Feedback and optimal sensitivity: Model reference transformations, multiplicative seminorms, and approximate inverses," *IEEE Trans. Autom. Control*, vol. 26, no. 2, pp. 301–320, Apr. 1981.

[8] A. Isidori and A. Astolfi, "Disturbance attenuation and H_∞-control via measurement feedback in nonlinear systems," *IEEE Trans. Autom. Control*, vol. 37, no. 9, pp. 1283–1293, Sep. 1992.

[9] T. Başar and P. Bernhard, *H-Infinity Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach*. Boston, MA, USA: Springer, 2008.

[10] A. J. Van Der Schaft, "L_2-gain analysis of nonlinear systems and nonlinear state feedback H_∞ control," *IEEE Trans. Autom. Control*, vol. 37, no. 6, pp. 770–784, Jun. 1992.

[11] X. Zhong, H. He, D. Wang, and Z. Ni, "Model-free adaptive control for unknown nonlinear zero-sum differential game," *IEEE Trans. Cybern.*, vol. 48, no. 5, pp. 1633–1646, May 2018.

[12] Y. Yang, K. G. Vamvoudakis, H. Ferraz, and H. Modares, "Dynamic intermittent Q-learning-based model-free suboptimal co-design of L_2-stabilization," *Int. J. Robust Nonlinear Control*, vol. 29, no. 9, pp. 2673–2694, 2019.

[13] M. Abu-Khalaf, J. Huang, and F. L. Lewis, *Nonlinear H_2/H_∞ Constrained Feedback Control: A Practical Design Approach Using Neural Networks*. London, U.K.: Springer, 2006.

[14] B. Luo, Y. Yang, and D. Liu, "Policy iteration Q-learning for data-based two-player zero-sum game of linear discrete-time systems," *IEEE Trans. Cybern.*, vol. 51, no. 7, pp. 3630–3640, Jul. 2021.

[15] H. Zhang, C. Qin, B. Jiang, and Y. Luo, "Online adaptive policy learning algorithm for H_∞ state feedback control of unknown affine nonlinear discrete-time systems," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2706–2718, Dec. 2014.

[16] M. Abu-Khalaf, F. L. Lewis, and J. Huang, "Policy iterations on the Hamilton-Jacobi-Isaacs equation for H_∞ state feedback control with input saturation," *IEEE Trans. Autom. Control*, vol. 51, no. 12, pp. 1989–1995, Dec. 2006.

[17] R. W. Bea, "Successive Galerkin approximation algorithms for nonlinear optimal and robust control," *Int. J. Control*, vol. 71, no. 5, pp. 717–743, 1998.

[18] D. Prokhorov and D. Wunsch, "Adaptive critic designs," *IEEE Trans. Neural Netw.*, vol. 8, no. 5, pp. 997–1007, Sep. 1997.

[19] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, 4th ed. Belmont, MA, USA: Athena Sci., 2011.

[20] F. L. Lewis and D. Liu, *Reinforcement Learning and Approximate Dynamic Programming for Feedback Control*. Hoboken, NJ, USA: Wiley, 2012.

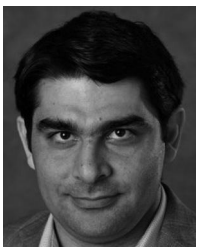
[21] X. Yang and H. He, "Event-driven H_∞-constrained control using adaptive critic learning," *IEEE Trans. Cybern.*, vol. 51, no. 10, pp. 4860–4872, Oct. 2021.

[22] Y. Lv and X. Ren, "Approximate Nash solutions for multiplayer mixed-zero-sum game with reinforcement learning," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 49, no. 12, pp. 2739–2750, Dec. 2019.
[23] J. Zhao, Y. Lv, and Z. Zhao, "Adaptive learning based output-feedback optimal control of CT two-player zero-sum games," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 69, no. 3, pp. 1437–1441, Mar. 2022.
[24] 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.
[25] Y. Yang, W. Gao, H. Modares, and C.-Z. Xu, "Robust actor-critic learning for continuous-time nonlinear systems with unmodeled dynamics," *IEEE Trans. Fuzzy Syst.*, vol. 30, no. 6, pp. 2101–2112, Jun. 2022.
[26] K. G. Vamvoudakis and F. Lewis, "Online solution of nonlinear two-player zero-sum games using synchronous policy iteration," *Int. J. Robust Nonlinear Control*, vol. 22, no. 13, pp. 1460–1483, 2012.
[27] Y. Li, T. Yang, and S. Tong, "Adaptive neural networks finite-time optimal control for a class of nonlinear systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 11, pp. 4451–4460, Nov. 2020.
[28] Y. Li, X. Min, and S. Tong, "Observer-based fuzzy adaptive inverse optimal output feedback control for uncertain nonlinear systems," *IEEE Trans. Fuzzy Syst.*, vol. 29, no. 6, pp. 1484–1495, Jun. 2021.
[29] H. Su, H. Zhang, H. Jiang, and Y. Wen, "Decentralized event-triggered adaptive control of discrete-time nonzero-sum games over wireless sensor-actuator networks with input constraints," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 10, pp. 4254–4266, Oct. 2020.
[30] Y. Yang, H. Modares, K. G. Vamvoudakis, W. He, C.-Z. Xu, and D. C. Wunsch, "Hamiltonian-driven adaptive dynamic programming with approximation errors," *IEEE Trans. Cybern.*, vol. 52, no. 12, pp. 13762–13773, Dec. 2022.
[31] G. Chowdhary, T. Yucelen, M. Mühlegg, and E. N. Johnson, "Concurrent learning adaptive control of linear systems with exponentially convergent bounds," *Int. J. Adapt. Control Signal Process.*, vol. 27, no. 4, pp. 280–301, 2013.
[32] Y. Yang, D.-W. Ding, H. Xiong, Y. Yin, and D. C. Wunsch, "Online barrier-actor-critic learning for H_∞ control with full-state constraints and input saturation," *J. Franklin Inst.*, vol. 357, no. 6, pp. 3316–3344, 2020.
[33] Y. Yang, K. G. Vamvoudakis, and H. Modares, "Safe reinforcement learning for dynamical games," *Int. J. Robust Nonlinear Control*, vol. 30, no. 9, pp. 3706–3726, 2020.
[34] R. Kamalapurkar, P. Walters, and W. E. Dixon, "Model-based reinforcement learning for approximate optimal regulation," *Automatica*, vol. 64, pp. 94–104, Feb. 2016.
[35] G. Chowdhary and E. Johnson, "A singular value maximizing data recording algorithm for concurrent learning," in *Proc. Amer. Control Conf.*, 2011, pp. 3547–3552.
[36] R. Kamalapurkar, B. Reish, G. Chowdhary, and W. E. Dixon, "Concurrent learning for parameter estimation using dynamic state-derivative estimators," *IEEE Trans. Autom. Control*, vol. 62, no. 7, pp. 3594–3601, Jul. 2017.
[37] Y. Pan and H. Yu, "Composite learning from adaptive dynamic surface control," *IEEE Trans. Autom. Control*, vol. 61, no. 9, pp. 2603–2609, Sep. 2016.
[38] N. Cho, H. Shin, Y. Kim, and A. Tsourdos, "Composite model reference adaptive control with parameter convergence under finite excitation," *IEEE Trans. Autom. Control*, vol. 63, no. 3, pp. 811–818, Mar. 2018.
[39] Y. Yang, Y. Pan, C.-Z. Xu, and D. C. Wunsch, "Hamiltonian-driven adaptive dynamic programming with efficient experience replay," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Oct. 25, 2022, doi: 10.1109/TNNLS.2022.3213566
[40] A. Van der Schaft, *L_2-Gain and Passivity Techniques in Nonlinear Control*. Cham, Switzerland: Springer, 2000.
[41] D. Vrabie, K. G. Vamvoudakis, and F. L. Lewis, *Optimal Adaptive Control and Differential Games by Reinforcement Learning Principles*. Stevenage, U.K.: Inst. Eng. Technol., 2012.
[42] P. Ioannou and J. Sun, *Robust Adaptive Control*. Hoboken, NJ, USA: Prentice-Hall, 1995.
[43] C. Yuan, P. Stegagno, H. He, and W. Ren, "Cooperative adaptive containment control with parameter convergence via cooperative finite-time excitation," *IEEE Trans. Autom. Control*, vol. 66, no. 11, pp. 5612–5618, Nov. 2021.
[44] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Netw.*, vol. 2, no. 5, pp. 359–366, 1989.
[45] F. L. Lewis, H. Zhang, K. Hengster-Movric, and A. Das, *Cooperative Control of Multi-Agent Systems: Optimal and Adaptive Design Approaches*. London, U.K.: Springer, 2013.
[46] R. Kamalapurkar, J. A. Rosenfeld, and W. E. Dixon, "Efficient model-based reinforcement learning for approximate online optimal control," *Automatica*, vol. 74, pp. 247–258, Dec. 2016.
[47] K. G. Vamvoudakis, M. F. Miranda, and J. P. Hespanha, "Asymptotically stable adaptive–optimal control algorithm with saturating actuators and relaxed persistence of excitation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 11, pp. 2386–2398, Nov. 2016.
[48] S. J. Bradtke and A. G. Barto, "Linear least-squares algorithms for temporal difference learning," *Mach. Learn.*, vol. 22, no. 1, pp. 33–57, 1996.
[49] F. L. Lewis, S. Jagannathan, and A. Yesildirak, *Neural Network Control of Robot Manipulators and Non-Linear Systems*. Boca Raton, FL, USA: CRC Press, 1999.
[50] 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.
[51] W. Haddad and V. Chellaboina, *Nonlinear Dynamical Systems and Control: A Lyapunov-Based Approach*. Princeton, NJ, USA: Princeton Univ. Press, 2008.
[52] Y. Pan and H. Yu, "Composite learning robot control with guaranteed parameter convergence," *Automatica*, vol. 89, pp. 398–406, Mar. 2018.

Yongliang Yang (Member, IEEE) received the B.S. degree in electrical engineering from Hebei University, Baoding, China, in 2011, and the Ph.D. degree in control theory and control engineering from the University of Science and Technology Beijing (USTB), Beijing, China, in 2018.

From 2015 to 2017, he was a Visiting Scholar with the Missouri University of Science and Technology, Rolla, MO, USA, sponsored by China Scholarship Council. He was an Assistant Professor with USTB from 2018 to 2020. From 2020 to 2021, he was an independent Postdoctoral Research Fellow with the State Key Laboratory of Internet of Things for Smart City, Faculty of Science and Technology, University of Macau, Macau, China. He is currently an Associate Professor with USTB. His research interests include reinforcement learning theory, robotics, distributed optimization and control for cyber–physical systems.

Dr. Yang was a recipient of the Best Ph.D. Dissertation of the China Association of Artificial Intelligence, the Best Ph.D. Dissertation of USTB, the Chancellor's Scholarship in USTB, the Excellent Graduates Awards in Beijing, and the UM Macao Talent Program in Macau. He is an Associate Editor for IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS.



Hamidreza Modares (Senior Member, IEEE) received the B.S. degree in electrical engineering from the University of Tehran, Tehran, Iran, in 2004, the M.S. degree in electrical engineering from the Shahrood University of Technology, Shahrood, Iran, in 2006, and the Ph.D. degree in electrical engineering from The University of Texas at Arlington, Arlington, TX, USA, in 2015.

He was a Senior Lecturer with the Shahrood University of Technology from 2006 to 2009, and a Faculty Research Associate with The University of Texas at Arlington from 2015 to 2016. He is currently an Assistant Professor with the Mechanical Engineering Department, Michigan State University, East Lansing, MI, USA. His current research interests include cyber–physical systems, reinforcement learning, distributed control, robotics, and machine learning.

Dr. Modares was the recipient of the Best Paper Award from the 2015 IEEE International Symposium on Resilient Control Systems. He is an Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS.

Kyriakos G. Vamvoudakis (Senior Member, IEEE) received the Diploma degree in electronic and computer engineering from the Technical University of Crete, Chania, Greece, in 2006, and the M.S. and Ph.D. degrees in electrical engineering from The University of Texas at Arlington, Arlington, TX, USA, in 2008 and 2011, respectively.

From 2012 to 2016, he was a Project Research Scientist with the Center for Control, Dynamical Systems and Computation, University of California at Santa Barbara, Santa Barbara, CA, USA. He was an Assistant Professor with the Kevin T. Crofton Department of Aerospace and Ocean Engineering, Virginia Tech, Blacksburg, VA, USA, until 2018. He currently serves as an Assistant Professor with The Daniel Guggenheim School of Aerospace Engineering, Georgia Institute of Technology (Georgia Tech), Atlanta, GA, USA. His research interests include approximate dynamic programming, game theory, cyber–physical security, networked control, smart grid, and safe autonomy.

Dr. Vamvoudakis was a recipient of the 2019 ARO YIP Award and the 2018 NSF CAREER Award. He is currently an Associate Editor of *Automatica* and IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS.

Frank L. Lewis (Life Fellow, IEEE) received the bachelor's degree in physics/electrical engineering and the M.E.E. degree from Rice University, Houston, TX, USA, in 1971 and 1971, respectively, the M.S. degree in aeronautical engineering from the University of West Florida, Pensacola, FL, USA, in 1977, and the Ph.D. degree from the Georgia Institute of Technology, Atlanta, GA, USA, in 1988.

He is the UTA Distinguished Scholar Professor, the UTA Distinguished Teaching Professor, and the Moncrief-O'Donnell Chair with The University of Texas at Arlington Research Institute, Arlington, TX, USA. He has seven U.S. patents, numerous journal special issues, numerous journal articles, and 20 books. He is involved in feedback control, reinforcement learning, intelligent systems, and distributed control systems.

Dr. Lewis is the PE Texas, the U.K. Chartered Engineer, a member of the National Academy of Inventors and Fellow Institute Measurement Control, and an IFAC Fellow.