Electronic Research Archive

https://www.aimspress.com/journal/era

ERA, 34(8): 5496–5524.
DOI: 10.3934/era.2026246
Received: 28 April 2026
Revised: 16 June 2026
Accepted: 18 June 2026
Published: 30 June 2026

Research article

Actor–critic–identifier reinforcement learning with gradient-based integral concurrent learning for continuous-time nonlinear nonzero-sum differential graphical games

Lei Guo*, Fei Li and Yuan Song

School of Intelligent Engineering and Automation, Beijing University of Posts and Telecommunications, Beijing 100876, China

* **Correspondence:** Email: guolei@bupt.edu.cn.

Abstract: In this paper, we investigate multiplayer nonzero-sum differential graphical games for continuous-time nonlinear systems with unknown dynamics. Conventional reinforcement-learning-based optimal control methods for such problems usually suffer from the curse of dimensionality in parameter estimation and high online computational cost. To address these issues, an indirect adaptive control algorithm with distributed sparse identification is developed within an actor–critic–identifier architecture. Specifically, graphical game structural priors are incorporated to reduce the dimension of the unknown parameter space, and a gradient-based integral concurrent learning identifier using a historical data stack is designed for online system identification. In contrast to conventional least-squares-based concurrent learning methods, the proposed identifier avoids covariance matrix inversion and reduces the per-step computational complexity from $O(q^2)$ to $O(q)$. By combining the identifier with integral reinforcement learning, the proposed method achieves online approximation of the Nash equilibrium solution to the coupled Hamilton–Jacobi–Bellman equations without requiring an exact system model. A Lyapunov-based analysis is further conducted to establish exponential convergence of the identifier parameters under finite excitation and uniformly ultimately bounded stability of the closed-loop system under a persistence-of-excitation condition on the critic regressor. Comparative simulation results for the considered benchmark example show that, while maintaining a weight approximation error on the order of 10^{-3}, the proposed algorithm reduces the convergence time by 75%, thereby improving learning efficiency and alleviating the online computational burden.

Keywords: nonzero-sum differential games; differential graphical games; actor–critic–identifier; integral concurrent learning; integral reinforcement learning

1. Introduction

Game theory has attracted extensive attention in recent years [1, 2]. In the real world, interactive decision-making problems in which cooperation and competition coexist among multiple agents can be modeled as multiplayer nonzero-sum (NZS) differential games [3]. For continuous-time (CT) NZS games, optimal control methods are often used to analyze the strategy of each player. By solving the coupled Hamilton–Jacobi–Bellman (HJB) equations, each player can obtain a Nash equilibrium (NE) strategy with respect to its predefined performance index. If the system dynamics are nonlinear, the coupled HJB equations become a class of nonlinear partial differential equations, which are notoriously difficult, if not impossible, to solve analytically [4]. Therefore, neural network (NN) approximation techniques have been widely used to obtain approximate optimal solutions, including constrained-input non-zero-sum games [1], event-triggered robust nonzero-sum games with unknown dynamics [2], and unknown nonlinear multiplayer nonzero-sum games [3].

Reinforcement learning (RL) is currently a popular approach for handling complex decision-making and control problems. The concept of RL is mainly inspired by animal behavior and is an effective machine learning method [5]. In recent years, learning-based optimization and soft-computing methods have also been widely investigated in complex decision-making problems, such as distributed energy-efficient parallel-machine scheduling [6], assembly job-shop scheduling with uncertain assembly times [7], and dynamic flexible job-shop scheduling based on heterogeneous graph reinforcement learning [8]. These studies show the potential of RL-related and learning-based optimization methods in complex systems. For NZS problems, many RL-based methods have also been developed after the early online learning study on unknown-dynamics NZS games [9] and the neural-network feedback-control foundation summarized in [10], such as industrial process optimal control [11], event-triggered control [12], output regulation [13], and optimal regulation of immune systems [14]. In the field of control, RL is also referred to as adaptive dynamic programming (ADP), which mainly studies the optimal control of deterministic CT systems [1, 2] and discrete-time (DT) systems [3]. Different from RL methods for discrete decision-making and scheduling problems, ADP for continuous-time control systems usually emphasizes admissible policy design, Hamilton–Jacobi–Bellman equation approximation, closed-loop stability, and convergence analysis based on Lyapunov theory. In this paper, we focus on the optimal control of deterministic CT systems for NZS problems.

In the control field, RL/ADP methods can be divided into two categories: model-free and model-based, with the most significant difference being whether system dynamics are required [15]. Model-free algorithms do not require knowledge of system dynamics, and representative studies include event-triggered ADP for unknown nonlinear systems [12], approximate N-player nonzero-sum game solutions for uncertain nonlinear systems [16], and online adaptive learning solutions of coupled Hamilton–Jacobi equations [17]. Therefore, these methods usually exhibit better adaptability to different environments. By contrast, model-based algorithms require complete system dynamics to solve the coupled HJB equations. Although their computational burden is lower [18], they have the advantages of reducing training time and improving learning efficiency [19]. However, due to the mismatch that often exists between prior knowledge and the actual environment, their adaptability to different environments may be limited.

To eliminate the dependence on system models, two main technical routes have been developed:

integral reinforcement learning (IRL) [20] and the actor–critic–identifier (ACI) architecture. The former was proposed by Vrabie and Lewis and has been further developed in recent years [20, 21]. IRL uses the integral form of the Bellman equation to remove the dependence on the drift dynamics. However, standard IRL methods usually still assume that the input dynamics are known. Although some recent model-free IRL studies [22] attempt to circumvent this limitation by introducing exploration noise, this usually leads to a significant increase in computational burden. By contrast, the latter was proposed by Bhasin et al. [23]. It is an indirect adaptive control (IAC) method that introduces an independent identifier neural network to estimate the dynamics online, thereby forming the ACI closed-loop structure. The key advantage of the ACI architecture lies in its ability to simultaneously handle the uncertainties in both the drift and input dynamics, and to use the identifier output to assist the updates of the critic and actor, without requiring precise knowledge of the system dynamics.

Because the ACI architecture has shown excellent performance in single-system settings, its extension to multiplayer differential games has attracted increasing attention. In this paper, we investigate an RL algorithm based on the ACI architecture, in which gradient-type integral concurrent learning is employed as the identifier to solve nonzero-sum graphical games. Concurrent learning (CL) was proposed by Chowdhary and Johnson [24], and was applied to NZS problems by Kamalapurkar et al. [25]. For nonlinear differential graphical games with unknown dynamics and external disturbances, Tatari et al. [26] developed an online distributed adaptive learning method in which actor–critic structures, identifiers, and recorded past observations were used to approximate the solution of coupled Hamilton–Jacobi–Isaacs equations. To relax the excitation requirement and avoid direct use of state derivatives, Parikh et al. [27] developed integral concurrent learning (ICL) and established parameter convergence under finite excitation. However, similar to Kamalapurkar's work, most existing CL algorithms still adopt least-squares-type update laws, whose core mathematical form involves covariance-matrix inversion. For a local identifier with parameter dimension q, the computational time complexity can be as high as $O(q^2)$. Building on this line of research, Le et al. [28] began to explore accelerated-gradient-based ICL methods in an attempt to avoid matrix inversion, but their application to continuous-time nonlinear nonzero-sum differential graphical games remains underexplored. In recent years, research on the ACI architecture for optimal control of uncertain nonlinear systems has continued to advance [29].

Another major challenge in multiplayer nonzero-sum games is the curse of dimensionality. In a general N-player game, the performance index V_i of each player depends not only on its own state x_i and control input u_i, but also on the states and inputs of all the other $N-1$ players. This means that the input dimension of the HJB equations grows linearly with the number of players, whereas the solution complexity increases exponentially. The theory of differential graphical games proposed by Vamvoudakis et al. [30] provides a possible way to address this issue. Graphical games use a communication topology graph to restrict the interaction range among players, so that the performance index of each player depends only on its neighbor set. This local interaction property allows the high-dimensional coupled HJB equations to be decomposed into N lower-dimensional locally coupled HJB equations. Since then, studies on the solution structure and distributed implementation of differential graphical games have been further extended in terms of the stability and robustness analysis of minmax solutions [31], the construction of distributed global Nash solutions [32], and alternative solution concepts for graphical games when Nash equilibria are

unattainable [33]. In the past two years, graphical-game research has been further extended to more complicated and more engineering-oriented scenarios, such as robust differential graphical games with external disturbances and nonzero leader inputs [34], Nash–minmax graphical-game solutions combined with reinforcement learning [35], and distributed global Nash solutions and online reinforcement learning implementations in the presence of adversarial inputs [36].

In summary, although ACI-based learning, concurrent-learning-based identification, and differential graphical games have been separately or partially investigated in the literature, these results do not directly provide a scalable online solution for continuous-time nonlinear nonzero-sum differential graphical games with unknown dynamics. In particular, ACI-based learning with recorded data has been studied for nonlinear differential graphical games under unknown dynamics [26]; however, in that work the graph topology is mainly used to formulate local tracking-error coupling and distributed zero-sum disturbance-rejection games. For the nonzero-sum Nash learning problem considered in this paper, the coupled Hamilton–Jacobi–Bellman equations, unknown nonlinear dynamics, and graph-induced local interactions must be handled simultaneously.

Motivated by these observations, this paper develops an indirect adaptive reinforcement learning method that integrates graph-structure-driven sparse identification, gradient-type integral concurrent learning, and the ACI architecture into a unified framework. The rationale behind this design is threefold. First, since the drift and input dynamics are unknown, an indirect adaptive learning structure is adopted so that the identifier can provide online model estimates for the subsequent critic and actor updates. Second, different from using the graph topology only to describe local information exchange or tracking-error coupling, the known interaction topology is further exploited as structural prior information to construct sparse local identifiers and reduce redundant parameter estimation. This sparsity is particularly beneficial for concurrent-learning-based identification, since the finite-excitation condition and history-stack construction are imposed on lower-dimensional local regressors rather than on a globally parameterized model. Third, a gradient-type ICL update law is used instead of a conventional least-squares-type concurrent-learning update, so that finite-excitation-based parameter convergence can be retained while avoiding covariance-matrix inversion. Therefore, the main methodological contribution of this work is to develop a computationally efficient ACI-based online Nash learning scheme, in which the known graphical structure is embedded into the gradient-type ICL identifier to reduce redundant parameter estimation for unknown continuous-time nonlinear nonzero-sum differential graphical games.

The main contributions of this paper are summarized as follows.

- For the high-dimensional globally coupled estimation problem in multiplayer nonzero-sum differential graphical games, a graph-structure-driven distributed sparse identification mechanism is developed. By exploiting the sparsity and topological prior information of the graph, the original unified parameter-estimation problem for globally unknown coupled dynamics is transformed into local distributed estimation problems involving only nonzero adjacent coupling weights. From the identifier-design viewpoint, this provides a topology-consistent structural mechanism for reducing the parameter-estimation dimension: the local regressor of each player retains only the basis functions associated with its neighbor set and the input channels over $\mathcal{N}_i \cup \{i\}$, while redundant non-neighbor coupling terms are excluded from the parameterization. The corresponding reduction in parameter dimension is theoretically clarified through the local parameterization and the dimension analysis. This design also

provides a structural basis for applying concurrent-learning-based identification to lower-dimensional local regressors.

- A gradient-type integral concurrent learning identifier based on a historical data stack is designed. Different from conventional least-squares-based concurrent learning methods [24, 25], the proposed identifier avoids covariance-matrix inversion and reduces the per-step computational complexity of parameter updating from $O(q^2)$ to $O(q)$, thereby retaining the advantages of low computational cost and easy online implementation of gradient-based updates. It also preserves the finite-excitation-based learning feature of ICL while avoiding the main matrix-inversion operation in least-squares-type CL schemes.
- Based on the above distributed sparse parameter-estimation mechanism and gradient-type ICL, an online indirect adaptive reinforcement learning algorithm under the ACI architecture is proposed. Without requiring an exact system dynamics model, it realizes online approximation of the Nash equilibrium solution to the coupled HJB equations and closed-loop control. Theoretically, exponential convergence of the identifier parameters under finite excitation conditions is established, and the uniform ultimate boundedness of the closed-loop system is proved under a persistence-of-excitation condition on the critic regressor. Simulation results for the considered benchmark verify the closed-loop Nash learning performance and illustrate consistency with the UUB stability analysis. Compared with existing model-free methods [15, 22], the proposed method can shorten the convergence time and reduce the online computational burden while maintaining the same order of weight-approximation error.

The remainder of this paper is organized as follows. Section 2 discusses the continuous-time nonlinear NZS problem and its Nash equilibrium strategy, introduces the general form of graphical games, and derives the coupled HJB equations. Section 3 proposes an indirect adaptive control algorithm based on ACI-architecture reinforcement learning and applies it to nonzero-sum differential graphical games. The ICL parameter update law and the adaptive update laws of the actor and critic are designed to estimate the value function and the Nash equilibrium strategy simultaneously. Based on Lyapunov analysis, the closed-loop stability of the proposed method is guaranteed. Section 4 verifies the closed-loop learning performance and computational efficiency of the proposed algorithm through simulation and comparative analysis of a two-player nonzero-sum game. Detailed Lyapunov analysis is provided in the Appendix.

Throughout this paper, $\|\cdot\|$ denotes the Euclidean norm of a vector or the induced norm of a matrix, and $(\cdot)^\top$ denotes the transpose of a vector or matrix. For a symmetric matrix X, $\lambda_{\max}(X)$ and $\lambda_{\min}(X)$ denote its maximum and minimum eigenvalues, respectively. The symbols I and 0 denote an identity matrix and a zero matrix with compatible dimensions, respectively. The operator $\mathrm{col}\{\cdot\}$ denotes column-wise stacking. For a differentiable scalar function $f(x)$, $\nabla f(x) = \partial f/\partial x$ denotes its gradient with respect to x. For ease of notation, the time argument is omitted when no confusion arises; for example, x, x_i, and u denote $x(t)$, $x_i(t)$, and $u(t)$, respectively.

2. Preliminaries and problem formulation

2.1. Graph

Consider a multiplayer system composed of N players. The communication topology among the players can be described by a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, A)$, where $\mathcal{V} = \{1, \ldots, N\}$ denotes the set

of players and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ denotes the edge set. The adjacency matrix $A = [a_{ij}] \in \mathbb{R}^{N\times N}$ is used to characterize the generalized coupling weights among the players. If the state or control input of player j affects the dynamical information of player i, then $(j, i) \in \mathcal{E}$ and $a_{ij} \neq 0$; otherwise, $a_{ij} = 0$. In this study, the diagonal entry a_{ii} may represent the self-feedback gain of player i. The edge set $\mathcal{E}$ is assumed to be known. In this paper, the adjacency pattern is used as structural prior information to determine which coupling terms are retained in the local parameterization; the numerical coupling strengths are absorbed into the unknown parameter vectors θ_i introduced later. The direction of each edge indicates the propagation direction of information or dynamical influence.

2.2. Problem formulation

The multiplayer system is first abstracted as the following continuous-time nonlinear system

$$\dot{x} = F(x) + G(x)u. \tag{2.1}$$

Here, $x = [x_1^\top, \ldots, x_i^\top, \ldots, x_N^\top]^\top \in \mathbb{R}^n$ denotes the global joint state vector, where $x_i \in \mathbb{R}^{n_i}$ denotes the local state of player i, and $n = \sum_{i=1}^N n_i$. Here, x_i also denotes the i-th state block of the global state vector x, and this notation will be used throughout the subsequent subsystem dynamics and identifier design. Similarly, $u = [u_1^\top, \ldots, u_i^\top, \ldots, u_N^\top]^\top \in \mathbb{R}^m$ denotes the joint control input, where $u_i \in \mathbb{R}^{m_i}$ denotes the control input of player i, and $m = \sum_{i=1}^N m_i$. The global drift vector field $F(x) \in \mathbb{R}^n$ is composed of the local drift terms and the state-coupling terms of all players stacked according to the state dimensions. Accordingly, $F(x)$ can be partitioned according to the players' state blocks as $F(x) = [F_1^\top(x), \ldots, F_i^\top(x), \ldots, F_N^\top(x)]^\top$, where $F_i(x)$ corresponds to the drift component of the i-th player's subsystem. The input matrix $G(x) = [g_1(x), \ldots, g_i(x), \ldots, g_N(x)] \in \mathbb{R}^{n\times m}$ is partitioned according to the players' control inputs, where $g_i(x) \in \mathbb{R}^{n\times m_i}$ denotes the input block associated with player i. Owing to neighbor-input coupling, $G(x)$ is not necessarily block diagonal. It is assumed that, on the compact operating set Ω, both $F(x)$ and $G(x)$ satisfy the local Lipschitz condition, and $F(0) = 0$.

On the basis of system (2.1), globally parameterizing and estimating the unknown dynamics online generally leads to an excessively high parameter dimension and a substantial increase in computational complexity, thereby impairing real-time performance and scalability. In practical multiplayer systems, each player typically interacts with only a limited number of neighbors, and such interaction relations can often be described by a known topology. If this structural prior is ignored and the unknown dynamics are estimated at the global level, a large number of redundant parameters will be introduced, which significantly reduces online learning efficiency and weakens the feasibility of distributed implementation. Motivated by this observation, the graphical-game modeling idea is further introduced so that the system dynamics can be structurally characterized under a known interaction topology, while preserving sufficient generality for online learning and game-theoretic control design.

For any player i, define the stacked neighbor-state vector as $x_{\mathcal{N}_i} = \mathrm{col}\{x_j : j \in \mathcal{N}_i\}$ and the stacked local-neighborhood input vector as $u_{\mathcal{N}_i\cup\{i\}} = \mathrm{col}\{u_j : j \in \mathcal{N}_i \cup \{i\}\}$, where $\mathrm{col}\{\cdot\}$ denotes the column-wise stacking operator. Then, the dynamics of the i-th subsystem depend only on its own state, the states of its neighbors, and the control inputs over its local interaction set.

According to the above graphical-game description and structural prior knowledge, the original system can be rewritten in the following graphical-game form. Inspired by the decomposition idea

for unknown nonlinear systems in Ming et al. [37], and combined with the known communication topology of multiplayer systems, the dynamics of player i can be expressed as

$$\dot{x}_i = f_i(x_i) + \sum_{j\in\mathcal{N}_i} h_{ij}(x_i, x_j) + \sum_{j\in\mathcal{N}_i\cup\{i\}} g_{ij}(x_i, x_j)u_j. \tag{2.2}$$

In (2.2), $f_i : \mathbb{R}^{n_i} \to \mathbb{R}^{n_i}$ denotes the local drift dynamics of player i, $h_{ij} : \mathbb{R}^{n_i} \times \mathbb{R}^{n_j} \to \mathbb{R}^{n_i}$ denotes the state-coupling effect of neighboring player j on subsystem i, and $g_{ij} : \mathbb{R}^{n_i} \times \mathbb{R}^{n_j} \to \mathbb{R}^{n_i\times m_j}$ characterizes the influence of control input u_j on subsystem i. When $j = i$, $g_{ii}(\cdot)$ corresponds to the local input channel of player i; when $j \in \mathcal{N}_i$, $g_{ij}(\cdot)$ describes the sparse neighbor-input coupling induced by the interaction topology. It follows from (2.2) that the overall multiplayer system remains input-affine with respect to the joint control input u, except that the input gain matrix now exhibits a topology-consistent sparse structure rather than a block-diagonal one. This representation explicitly encodes the sparsity of the graph topology and converts the original globally coupled estimation problem into a local estimation problem involving only nonzero neighbor interaction terms, thereby reducing the parameter dimension and facilitating distributed implementation of the identifier.

To ensure controllability of the practical physical system and establish the subsequent convergence analysis, the following standard assumptions are imposed on the system dynamics and approximation properties.

Assumption 1.

(a) *Continuity and boundedness of the system dynamics:* On the compact set Ω, the drift term $F(x)$ is locally Lipschitz and linearly bounded, that is, there exists a positive constant b_f such that $\|F(x)\| \le b_f\|x\|$ for all $x \in \Omega$. Moreover, for each player i, the corresponding input block $g_i(x)$ is bounded on Ω, namely, there exists a positive constant b_{g_i} such that $\|g_i(x)\| \le b_{g_i}$ for all $x \in \Omega$.

(b) *Neural-network approximation properties:* The neural-network basis function $\phi_i(x)$ used to approximate the value function and its gradient $\nabla\phi_i(x)$ are bounded on Ω. In addition, the ideal critic weight W_{1i}, the value-function approximation error $\varepsilon_i(x)$, and its gradient $\nabla\varepsilon_i(x)$ are all bounded.

(c) *Boundedness of the exploration signal:* To guarantee the persistence of excitation required for online learning, an artificially designed exploration signal $e_i(t)$ is introduced into the control input. It is assumed to be a bounded continuous signal, i.e., there exists a positive constant $\bar{e}_i$ such that $\|e_i(t)\| \le \bar{e}_i$ for all $t \ge 0$.

These are the standard assumptions in nonlinear approximation and learning control. Based on them, the infinite-horizon performance index of player i is defined as

$$V_i(x(t)) = \int_t^\infty r_i(x(\tau), u(\tau))\, d\tau = \int_t^\infty \left(S_i(x) + \sum_{j=1}^{N} u_j^\top R_{ij} u_j \right) d\tau. \tag{2.3}$$

Here, $Q_i \in \mathbb{R}^{n_i\times n_i}$ is symmetric positive definite and $S_i(x) = x_i^\top Q_i x_i$. For $j \in \mathcal{N}_i \cup \{i\}$, $R_{ij} \in \mathbb{R}^{m_j\times m_j}$ is symmetric positive definite and characterizes the influence of player j's control effort on player i's cost function; if $j \notin \mathcal{N}_i \cup \{i\}$, then $R_{ij} = 0_{m_j\times m_j}$. Hence, only the control inputs of player i and its neighbors appear in the local cost functional.

Assuming that $V_i(x)$ is continuously differentiable, differentiation of (2.3) along the system trajectory yields

$$S_i(x) + \sum_{j=1}^{N} u_j^\top R_{ij} u_j + \nabla V_i^\top(x)(F(x) + G(x)u) = 0, \tag{2.4}$$

where $V_i(0) = 0$, and $\nabla V_i(x)$ denotes the gradient of the value function with respect to the state.

According to the certainty-equivalence principle [38], the online estimated parameters or model are substituted into the control law derived under the assumption that the system model is known. However, certainty equivalence alone usually guarantees only boundedness of the closed-loop signals or convergence of the tracking error, rather than convergence of the estimated dynamical parameters to the true physical parameters. Classical adaptive control theory shows that asymptotic convergence of the parameter estimation error requires the regressor signal to satisfy the stringent persistence of excitation (PE) condition [38], namely, there exist positive constants T_0 and α_0 such that

$$\lambda_{\min}\left[\int_t^{t+T_0} \omega(\tau)\omega^\top(\tau)\, d\tau\right] \geq \alpha_0, \qquad \forall t \geq 0, \tag{2.5}$$

where $\omega(\tau) \in \mathbb{R}^\ell$ denotes a generic regressor vector. This condition requires the signal to remain sufficiently rich in all directions of the parameter space over every sliding time window of length T_0.

To relax this restrictive requirement at the identifier layer, the ICL identifier developed in this paper only requires the mild finite excitation (FE) condition, namely, that the recorded history stack be full rank over a finite time interval. In particular, if

$$\lambda_{\min}\left(\sum_{k=1}^{M} \mathcal{Y}_{ik}^\top \mathcal{Y}_{ik}\right) \geq \lambda > 0,$$

then exact parameter convergence can be established theoretically. By contrast, the closed-loop uniformly ultimately bounded (UUB) result in Theorem 2 additionally relies on PE of the critic regressor.

Specifically, the previously estimated input dynamics $\hat{G}(x)$ are used to replace the actual input dynamics $G(x)$, where $\hat{G}(x) = [\hat{g}_1(x), \ldots, \hat{g}_N(x)]$. Then, the Hamiltonian function of player i is defined as

$$H_i(x, \nabla V_i, u) = S_i(x) + \sum_{j=1}^{N} u_j^\top R_{ij} u_j + \nabla V_i^\top(x)(F(x) + \hat{G}(x)u). \tag{2.6}$$

Accordingly, the coupled HJB equations can be written as

$$H_i(x, \nabla V_i, u) = 0. \tag{2.7}$$

At the Nash equilibrium, given the strategies of the other players, player i obtains its optimal feedback control by minimizing its own Hamiltonian, i.e.,

$$\frac{\partial H_i}{\partial u_i} = 2R_{ii}u_i + \hat{g}_i^\top(x)\nabla V_i(x) = 0. \tag{2.8}$$

Thus, the optimal strategy of player i is given by

$$u_i^* = -\frac{1}{2} R_{ii}^{-1} \hat{g}_i^\top(x)\nabla V_i(x). \tag{2.9}$$

This shows that the optimal control of each player is a linear combination of the estimated input vector field $\hat{g}_i(x)$ and the corresponding value-function gradient $\nabla V_i(x)$. Substituting the optimal control strategy u^* into (2.7) yields the expanded form of the coupled HJB equations

$$\begin{aligned} 0 = S_i(x) + \nabla V_i^\top(x) F(x) - \frac{1}{2}\nabla V_i^\top(x) \sum_{j=1}^{N} \hat{g}_j(x) R_{jj}^{-1} \hat{g}_j^\top(x) \nabla V_j(x) \\ + \frac{1}{4} \sum_{j=1}^{N} \nabla V_j^\top(x) \hat{g}_j(x) R_{jj}^{-1} R_{ij} R_{jj}^{-1} \hat{g}_j^\top(x) \nabla V_j(x). \end{aligned} \tag{2.10}$$

3. Reinforcement learning algorithm based on the ACI architecture

3.1. Identifier design

According to the linear parameterization assumption in adaptive control theory [39], the unknown dynamics of a nonlinear system can, under suitable conditions, be represented as the product of a known regressor matrix and an unknown constant parameter vector. Similar modeling assumptions have been widely adopted in adaptive control and concurrent learning frameworks [24]. For the topological weights and input-dynamics uncertainty appearing in (2.2), an online-identifiable mathematical model is required. In this paper, it is assumed that the nonlinear system satisfies the linear parameterization condition.

Assume that the dynamics of player i in (2.2) can be rewritten as

$$\dot{x}_i = Y_i(x_i, x_{\mathcal{N}_i}, u_{\mathcal{N}_i \cup \{i\}})\, \theta_i, \tag{3.1}$$

where $\theta_i \in \mathbb{R}^{p_i}$ is an unknown constant parameter vector to be estimated, containing the unknown coefficients in the local drift term f_i, the local input gain, and the sparse neighbor-coupling terms associated with $\mathcal{N}_i$. Correspondingly, the regressor matrix $Y_i(\cdot) \in \mathbb{R}^{n_i \times p_i}$ is constructed from known basis functions associated with these unknown parameters.

Since $Y_i(\cdot)$ depends only on the local state x_i, the neighboring states $x_{\mathcal{N}_i}$, and the control inputs over the local interaction set $u_{\mathcal{N}_i \cup \{i\}}$, the resulting identifier naturally inherits the sparsity of the graph topology and therefore avoids redundant estimation of globally coupled terms.

To make the parameter-reduction mechanism explicit, we further compare the full parameterization with the topology-consistent local parameterization. Suppose that the local drift basis dimension is p_f, each state-coupling basis dimension is p_h, and each input-channel basis dimension is p_g. If the graph structure is not exploited, player i has to include all possible state-coupling terms from the other $N-1$ players and all possible input channels, and the corresponding local parameter dimension is

$$p_i^{\text{full}} = p_f + (N-1)p_h + Np_g.$$

By contrast, under the proposed topology-consistent sparse identifier, only the terms associated with the neighbor set $\mathcal{N}_i$ and the input channels over $\mathcal{N}_i \cup \{i\}$ are retained. Therefore, the local parameter dimension becomes

$$p_i^{\text{sp}} = p_f + |\mathcal{N}_i| p_h + (|\mathcal{N}_i| + 1) p_g.$$

When the graph is sparse, i.e., $|\mathcal{N}_i| \ll N-1$, it follows that $p_i^{\mathrm{sp}} \ll p_i^{\mathrm{full}}$. This analysis shows that the known adjacency pattern is explicitly used to construct lower-dimensional local regressors, rather than only to describe information exchange among players.

To handle the uncertainty in (3.1), an ICL parameter estimation scheme is developed. Since direct measurement of the state derivative $\dot{x}_i$ may introduce high-frequency noise in practice, the differential-form dynamics are converted into an algebraic form by using integral transformation. Specifically, integrating both sides of (3.1) over the sliding time window $[t-\Delta T, t]$ yields

$$\int_{t-\Delta T}^{t} \dot{x}_i(\tau)\, d\tau = \int_{t-\Delta T}^{t} Y_i\big(x_i(\tau), x_{\mathcal{N}_i}(\tau), u_{\mathcal{N}_i\cup\{i\}}(\tau)\big)\, d\tau\, \theta_i, \tag{3.2}$$

where $\Delta T > 0$ denotes the integration-window length.

By the fundamental theorem of calculus, the left-hand side of (3.2) can be converted into the state difference between the current instant and a historical instant. Since θ_i is constant, it can be extracted from the integral. Define the state-difference vector over the sliding window and the integral regressor matrix as

$$\Delta x_i(t) = x_i(t) - x_i(t-\Delta T) \in \mathbb{R}^{n_i}, \tag{3.3}$$

and

$$\mathcal{Y}_i(t) = \int_{t-\Delta T}^{t} Y_i\big(x_i(\tau), x_{\mathcal{N}_i}(\tau), u_{\mathcal{N}_i\cup\{i\}}(\tau)\big)\, d\tau \in \mathbb{R}^{n_i\times p_i}. \tag{3.4}$$

Then, the original differential dynamics can be rewritten in the algebraic form

$$\Delta x_i(t) = \mathcal{Y}_i(t)\theta_i, \tag{3.5}$$

which does not explicitly involve the state derivative.

The key idea of concurrent learning is to exploit historical data so as to relax the excitation requirement on the instantaneous signal. To this end, a cyclic history stack containing M data samples is constructed as $\{\Delta x_{ik}, \mathcal{Y}_{ik}\}_{k=1}^{M}$, where the subscript k denotes the k-th sampled historical data point.

The history stack is updated according to the following rule: whenever the minimum eigenvalue of the information matrix formed by the newly collected historical regressor exceeds a prescribed positive threshold λ, the update law is activated so as to guarantee the FE condition. This mechanism ensures that the stored data contain effective dynamical information.

To analyze parameter convergence, define the parameter estimation error as

$$\tilde{\theta}_i(t) = \theta_i - \hat{\theta}_i(t). \tag{3.6}$$

Following Parikh et al. [27], construct the performance index based on the prediction error of historical data as

$$J_i(\hat{\theta}_i) = \frac{1}{2}\sum_{k=1}^{M} \left\|\Delta x_{ik} - \mathcal{Y}_{ik}\hat{\theta}_i\right\|^2, \tag{3.7}$$

whose gradient with respect to $\hat{\theta}_i$ is

$$\nabla_{\hat{\theta}_i} J_i = -\sum_{k=1}^{M} \mathcal{Y}_{ik}^{\top}\left(\Delta x_{ik} - \mathcal{Y}_{ik}\hat{\theta}_i\right). \tag{3.8}$$

Since both the critic and actor update laws in the overall ACI framework are gradient-based, the identifier is also designed in a gradient form to maintain structural consistency. Therefore, the pure ICL-based adaptive parameter update law using the history stack is given by

$$\dot{\hat{\theta}}_i(t) = \Gamma_i \sum_{k=1}^{M} \mathcal{Y}_{ik}^{\top} \left(\Delta x_{ik} - \mathcal{Y}_{ik} \hat{\theta}_i \right), \tag{3.9}$$

where $\Gamma_i \in \mathbb{R}^{p_i \times p_i}$ is a constant symmetric positive-definite learning-rate matrix, and $k_{\mathrm{ICL}} > 0$ is the concurrent-learning gain absorbed into Γ_i for notational simplicity. The history stack is updated whenever the replacement of an old sample by a new one satisfies

$$\lambda_{\min}\left(\sum_{k=1}^{M} \mathcal{Y}_{ik}^{\top} \mathcal{Y}_{ik} \right) \geq \lambda > 0.$$

Theorem 1 (ICL parameter convergence): Consider the linearly parameterized system (3.1) with the identifier update law (3.9). If the history stack satisfies the FE condition, namely, there exists a constant $\lambda > 0$ such that

$$\lambda_{\min}\left(\sum_{k=1}^{M} \mathcal{Y}_{ik}^{\top} \mathcal{Y}_{ik} \right) \geq \lambda,$$

then the parameter estimation error $\tilde{\theta}_i(t)$ converges to zero exponentially.

Proof: See Appendix A.

The above convergence result provides the foundation for the subsequent actor–critic learning process. As the identification error decreases, the estimated input-gain matrix $\hat{G}(x)$ approaches the true input dynamics $G(x)$ more closely, thereby improving the accuracy of the Bellman residual constructed from the identified model and facilitating the online approximation of the Nash equilibrium strategy.

3.2. Actor–critic update law design

Although the coupled HJB equations provide the necessary conditions for the Nash equilibrium solution, these nonlinear partial differential equations are generally difficult, if not impossible, to solve analytically. Moreover, to ensure parameter convergence during online learning, exploration noise must be introduced into the control input so that the persistence of excitation condition can be satisfied. Therefore, based on the system input dynamics estimated by the ICL identifier, an integral reinforcement-learning algorithm with exploration is developed under the actor–critic–identifier architecture to approximately solve the coupled HJB equations online.

For the affine dynamics in (2.2), after introducing a bounded exploration signal $e \in \mathbb{R}^m$, the system dynamics can be written as

$$\dot{x} = F(x) + G(x)(u + e). \tag{3.10}$$

To realize online learning relying only on measured data and the identified model, denote the input-matrix estimate generated by the ICL identifier as $\hat{G}(x) \in \mathbb{R}^{n \times m}$, and define the input-matrix estimation error as

$$\tilde{G}(x) = G(x) - \hat{G}(x).$$

Then (3.10) can be rewritten as

$$\dot{x} = F(x) + \hat{G}(x)(u + e) + \tilde{G}(x)(u + e). \tag{3.11}$$

Substituting the above decomposition into the Bellman equation leads to

$$\dot{V}_i = -\Big[S_i + \sum_{j=1}^{N} u_j^\top R_{ij} u_j\Big] + \nabla V_i^\top \hat{G}(x)\, e + \varepsilon_{\mathrm{id},i}, \tag{3.12}$$

where $\varepsilon_{id,i} = \nabla V_i^\top(x)\tilde{G}(x)e$ is the residual term induced by the model estimation error. Since $\tilde{G}(x)$ converges during learning, this residual remains bounded and asymptotically vanishes.

To avoid direct use of state derivatives, integrating (3.12) over the interval $[t - T, t]$ gives

$$V_i(x(t)) - V_i(x(t-T)) - \int_{t-T}^{t} \nabla V_i^\top \hat{G}(x)\, e\, d\tau = -\int_{t-T}^{t} \Big[S_i(x) + \tfrac{1}{4}\sum_{j=1}^{N} \nabla V_j^\top \hat{g}_j R_{jj}^{-1} R_{ij} R_{jj}^{-1} \hat{g}_j^\top \nabla V_j\Big] d\tau + \Delta\varepsilon_{\mathrm{id},i}, \tag{3.13}$$

where $\Delta\varepsilon_{id,i} = \int_{t-T}^{t} \varepsilon_{id,i}(\tau)\, d\tau$ denotes the integrated model-error term. Since $\hat{G}(x)$ is available and $\Delta\varepsilon_{id,i}$ decreases as the identification process converges, the term $\varepsilon_{id,i}$ is treated as a bounded perturbation and is absorbed into the residual terms in the subsequent critic and Lyapunov analyses.

Benefiting from the previously designed ICL identifier, the estimate $\hat{G}(x)$ is already available. Therefore, it can be directly incorporated into the integral Bellman equation, thereby avoiding estimation of redundant coupling terms. Under this framework, only the value functions of the players need to be approximated, which effectively reduces the computational burden and improves learning efficiency.

To realize online solution, a neural network is employed to parameterize each value function. By the Weierstrass approximation theorem, for any smooth function $V_i(x)$ over the compact set Ω, there exists a sufficiently large number of neurons N_i and an activation-function vector $\phi_i(x) \in \mathbb{R}^{N_i}$ such that the value function and its gradient can be uniformly approximated. Hence, the value function of player i and its gradient are represented as

$$V_i(x) = W_{1i}^\top \phi_i(x) + \varepsilon_i(x), \qquad \nabla V_i(x) = \nabla \phi_i^\top(x) W_{1i} + \nabla \varepsilon_i(x), \tag{3.14}$$

where $W_{1i} \in \mathbb{R}^{N_i}$ is the ideal critic weight, $\phi_i(x) \in \mathbb{R}^{N_i}$ is the basis-function vector, and $\nabla\phi_i(x) \in \mathbb{R}^{N_i \times n}$ so that $\nabla\phi_i^\top(x)W_{1i} \in \mathbb{R}^n$. The terms $\varepsilon_i(x)$ and $\nabla\varepsilon_i(x)$ denote the reconstruction errors of the value function and its gradient, respectively, both of which are uniformly bounded on Ω.

Since the ideal weight W_{1i} is unknown, a critic network is constructed for online estimation. Define the estimated value function of player i as

$$\hat{V}_i(x) = \hat{W}_{1i}^\top \phi_i(x),$$

where $\hat{W}_{1i}$ is the current estimate of the critic weight. Then define the measurable critic residual as

$$E_{1i} = \Delta\phi_i^\top \hat{W}_{1i} + \rho_i(t), \tag{3.15}$$

where

$$\Delta\phi_i = \phi_i(x(t)) - \phi_i(x(t-T)) - \int_{t-T}^{t} \nabla\phi_i^\top(x(\tau))\hat{G}(x(\tau))e(\tau)\, d\tau,$$

and

$$\rho_i(t) = \int_{t-T}^{t} \left(S_i(x) + \sum_{j=1}^{N} u_j^\top R_{ij} u_j \right) d\tau.$$

Substituting (3.14) into the integral Bellman equation shows that the ideal critic weight satisfies

$$\Delta\phi_i^\top W_{1i} + \rho_i(t) + \varepsilon_{B_i}(t) = 0,$$

where

$$\varepsilon_{B_i}(t) = \varepsilon_i(x(t)) - \varepsilon_i(x(t-T)) - \int_{t-T}^{t} \nabla\varepsilon_i^\top(x(\tau))\hat{G}(x(\tau))e(\tau)\,d\tau - \Delta\varepsilon_{id,i}(t).$$

Hence, the measurable residual (3.15) can be rewritten as

$$E_{1i} = -\Delta\phi_i^\top \tilde{W}_{1i} - \varepsilon_{B_i}(t), \qquad \tilde{W}_{1i} = W_{1i} - \hat{W}_{1i}.$$

To minimize the estimation error in (3.15), define the objective function of the i-th critic network as

$$l_i = \frac{1}{2} E_{1i}^2. \tag{3.16}$$

Based on (3.16) and the chain rule, the critic weight update law is designed via normalized gradient descent as

$$\dot{\hat{W}}_{1i} = -\alpha_{1i} \frac{\Delta\phi_i}{m_{si}} E_{1i}, \qquad m_{si} = 1 + \Delta\phi_i^\top \Delta\phi_i, \tag{3.17}$$

where $\alpha_{1i} > 0$ is the learning rate of the critic network. The normalization factor m_{si} is introduced to suppress excessively large update gains when the regressor norm becomes large, thereby improving boundedness and numerical stability of the learning process.

Next, an actor network is introduced to generate the real-time control input for each player. Using the estimated input block $\hat{g}_i(x)$ and the ideal critic weight W_{1i}, the optimal strategy can be expressed as

$$u_i^* = -\frac{1}{2} R_{ii}^{-1} \hat{g}_i^\top(x) \nabla\phi_i^\top(x) W_{1i}. \tag{3.18}$$

Since the ideal weight W_{1i} is unknown, it is replaced by the current critic estimate $\hat{W}_{1i}$. Accordingly, define the reference policy as

$$u_i = -\frac{1}{2} R_{ii}^{-1} \hat{g}_i^\top(x) \nabla\phi_i^\top(x) \hat{W}_{1i}. \tag{3.19}$$

The signal u_i in (3.19) is a virtual target generated from the critic and is used only for actor learning. The actual control applied to the plant is $\hat{u}_i + e_i$, where $\hat{u}_i$ is the nominal actor output and e_i is the bounded exploration signal introduced in (3.10). Accordingly, in the Bellman-residual derivation, the symbol u denotes the nominal policy component, whereas the exploration enters additively through e.

The actor network then generates the real-time control input according to

$$\hat{u}_i = -\frac{1}{2} R_{ii}^{-1} \hat{g}_i^\top(x) \nabla\phi_i^\top(x) \hat{W}_{2i}, \tag{3.20}$$

where $\hat{W}_{2i} \in \mathbb{R}^{N_i}$ is the actor weight. By driving $\hat{u}_i$ to track the reference policy u_i, the actor network indirectly approximates the optimal strategy u_i^*.

In the proposed algorithm, the actor and critic networks are updated simultaneously and continuously. The goal of the actor is to drive $\hat{u}_i$ toward the optimal policy while guaranteeing closed-loop stability and convergence of the estimation error. Define the loss function of the i-th actor network as

$$E_{2i} = 2(\hat{u}_i - u_i)^\top R_{ii}^2(\hat{u}_i - u_i) = \frac{1}{2}(\hat{W}_{1i} - \hat{W}_{2i})^\top B_i(\hat{W}_{1i} - \hat{W}_{2i}), \tag{3.21}$$

where

$$B_i = \nabla\phi_i(x)\hat{g}_i(x)\hat{g}_i^\top(x)\nabla\phi_i^\top(x).$$

The actor weight update law is designed as

$$\dot{\hat{W}}_{2i} = -\alpha_{2i}\left[\tanh\Big(B_i(\hat{W}_{1i} - \hat{W}_{2i})\Big) + \left(c_i - \frac{T}{4}\sum_{j=1}^{N}\frac{\Delta\phi_j}{m_{sj}^2}\hat{W}_{1j}^\top D_{3ij}\right)\hat{W}_{2i}\right], \tag{3.22}$$

where $\alpha_{2i} > 0$ is the actor learning rate, and $c_i > 0$ is the designed parameter. The first term in (3.22) employs the hyperbolic tangent function to guarantee boundedness of the weight update while descending along the negative gradient direction to minimize the loss function E_{2i}. The second term, together with the designed parameter c_i, is introduced to ensure the stability of the real-time learning process. More specifically, c_i acts as a damping-like design gain in the actor update law and is selected sufficiently large to dominate the bounded coupling term induced by the critic update and the cross terms appearing in the Lyapunov analysis. Here,

$$D_{3ij} = \nabla\phi_j(x)\hat{g}_j(x)R_{jj}^{-1}R_{ij}R_{jj}^{-1}\hat{g}_j^\top(x)\nabla\phi_j^\top(x).$$

Theorem 2 (UUB): Consider the multiplayer nonzero-sum differential graphical game described by (2.1) and (2.2). Suppose that Assumption 1 holds, that the history stack used in the identifier update law (3.9) satisfies the FE condition in Theorem 1, and that the normalized critic regressor satisfies a PE condition for all $t \geq t_{\mathrm{FE}}$. If the critic and actor update laws are given by (3.17) and (3.22), respectively, then there exist properly chosen design parameters $T > 0$ and $c_i > 0$ such that the closed-loop system state x and the neural-network weight estimation errors $\tilde{W}_{1i}$ and $\tilde{W}_{2i}$ are uniformly ultimately bounded.

Proof: See Appendix B.

Theorem 2 establishes the closed-loop stability of the proposed ACI learning architecture. Specifically, under appropriate excitation conditions, the system state and the neural-network weight errors remain bounded and enter a bounded neighborhood of the origin. This result shows that the coupled update mechanism of the identifier, critic, and actor networks is theoretically stable and also provides the analytical basis for the numerical simulations presented in the next section.

The pseudocode for the proposed algorithm is presented in Algorithm 1. The block diagram of the proposed algorithm is illustrated in Figure 1.

Algorithm 1 Proposed ACI-based learning algorithm with gradient-type ICL

1: **Initialization:** Set the initial state $x(0)$, the initial identifier parameters $\hat{\theta}_i(0)$, the critic weights $\hat{W}_{1i}(0)$, and the actor weights $\hat{W}_{2i}(0)$. Choose the reinforcement interval T, the ICL integration window ΔT, the history-stack size M, the learning gains, and the exploration signals $e_i(t)$.

2: **Online data collection:** Apply the control input $\hat{u}_i(t) + e_i(t)$ to the system. Collect the local state and input data of each player and its neighbors according to the known graph topology.

3: **Identifier update:** Construct the integral data pair $\{\Delta x_i, \mathcal{Y}_i\}$ and update the history stack. If the FE condition is satisfied, update $\hat{\theta}_i$ using the gradient-type ICL law and obtain the estimated dynamics $\hat{G}(x)$.

4: **Critic and actor update:** For each player, calculate the critic residual E_{1i} and the actor error E_{2i} using the online data and $\hat{G}(x)$. Update $\hat{W}_{1i}$ and $\hat{W}_{2i}$ using the critic and actor tuning laws, respectively.

5: Repeat Steps 2–4 until, for each player $i = 1, \ldots, N$, $\|\hat{\theta}_i(t) - \hat{\theta}_i(t - \Delta t_c)\| < \varepsilon_\theta$, $\|\hat{W}_{1i}(t) - \hat{W}_{1i}(t - \Delta t_c)\| < \varepsilon_1$, and $\|\hat{W}_{2i}(t) - \hat{W}_{2i}(t - \Delta t_c)\| < \varepsilon_2$, where Δt_c is the checking interval, and ε_θ, ε_1, and ε_2 are small positive constants. Then the learning process is regarded as convergent.

It should be noted that the additional online computation introduced by the proposed learning procedure mainly comes from the history-stack update and the gradient-type ICL identifier update. Compared with least-squares-type CL methods, the proposed update law does not require covariance-matrix inversion. Therefore, when the history-stack size is fixed, the dominant per-step parameter-update complexity is reduced from $O(q^2)$ to $O(q)$, where q denotes the local parameter dimension.

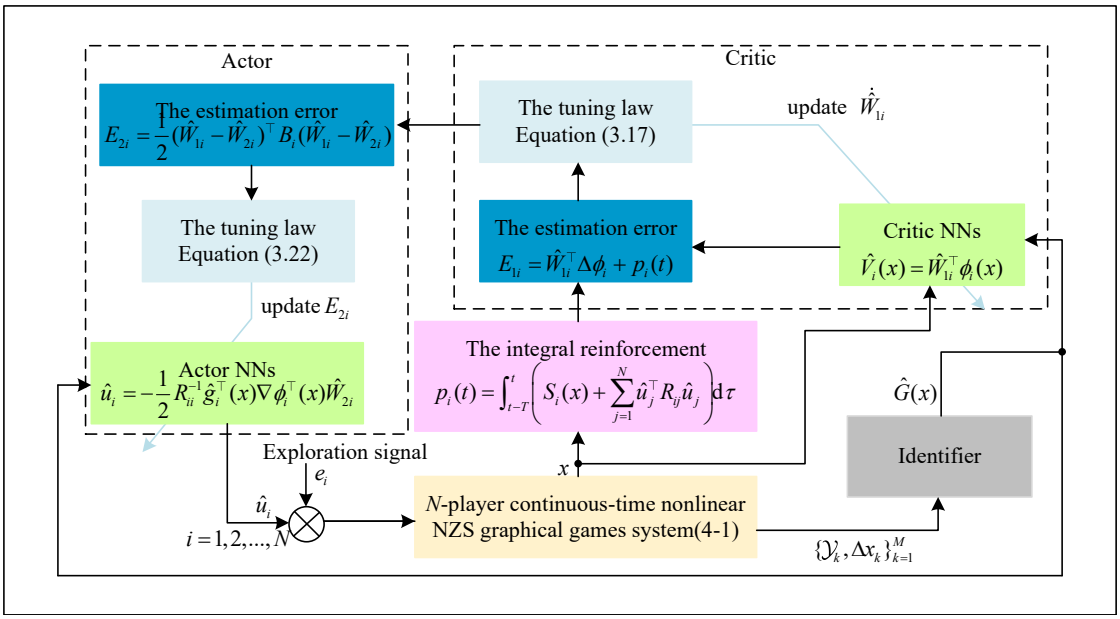


Figure 1. Architecture of the proposed indirect adaptive reinforcement learning algorithm for nonzero-sum differential graphical games based on the ACI architecture.

Figure 1 illustrates the implementation procedure of the proposed method. The closed-loop system consists of three main modules: the identifier, the critic network, and the actor network. First, to deal with the uncertainty of the system dynamics, the identifier module introduces the ICL mechanism and utilizes the state-input data stored in the history stack $\{\Delta x_k, \mathcal{Y}_k\}_{k=1}^{M}$ to estimate the unknown system dynamics online. The resulting estimate $\hat{G}(x)$ is then transmitted to the critic and actor networks, thereby removing the requirement for an exact model in optimal-control design. On this basis, the critic network constructs the Bellman residual E_{1i} from the integral reinforcement signal and updates the weight $\hat{W}_{1i}$ through gradient descent to approximate the value function of each player. The actor network computes the control law using the identified dynamics $\hat{G}(x)$ and the gradient information provided by the critic, and then updates the strategy by minimizing the deviation from the critic weight. The entire process operates in closed loop under the injected exploration signal e_i, achieving online approximation of the Nash equilibrium solution to the coupled HJB equations in a model-unknown environment.

4. Simulation studies

4.1. Simulation example

The proposed algorithm is validated numerically using a two-player shared-state differential graphical game. In this example, the two players correspond to the two control channels u_1 and u_2, respectively, while $x = [x_1, x_2]^\top$ denotes the common plant state. The interaction graph is chosen as a two-node complete directed graph with $\mathcal{V} = \{1, 2\}$, $\mathcal{N}_1 = \{2\}$, and $\mathcal{N}_2 = \{1\}$. Therefore, this benchmark represents the smallest nontrivial instance of the graphical-game setting, where the graphical interaction reduces to the full two-player case.

The considered two-player benchmark has been commonly used in related nonzero-sum ADP studies [14, 16, 21] and enables a fair comparison with representative existing methods. Since all possible edges are present in the two-node complete graph, this example mainly verifies the closed-loop Nash learning performance, convergence behavior, and computational efficiency of the proposed ACI-based learning scheme. The sparsity-induced parameter-reduction mechanism is supported by the topology-consistent parameterization and dimension analysis in Section 3.

The plant dynamics are given as

$$\begin{aligned} \dot{x}_1 &= -2x_1 + x_2, \\ \dot{x}_2 &= -\frac{1}{2}x_1 - x_2 + \frac{1}{4}x_2\Big((\cos(2x_1) + 2)^2 + (\sin(4x_1^2) + 2)^2\Big) \\ &\quad + (\cos(2x_1) + 2)u_1 + (\sin(4x_1^2) + 2)u_2. \end{aligned} \tag{4.1}$$

Rewriting (4.1) in the form of (2.1) gives

$$F(x) = \begin{bmatrix} -2x_1 + x_2 \\ -\frac{1}{2}x_1 - x_2 \\ \quad + \frac{1}{4}x_2\Big((\cos(2x_1) + 2)^2 + (\sin(4x_1^2) + 2)^2\Big) \end{bmatrix}, \tag{4.2}$$

and

$$G(x) = \begin{bmatrix} 0 & 0 \\ \cos(2x_1) + 2 & \sin(4x_1^2) + 2 \end{bmatrix}, \qquad u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}. \tag{4.3}$$

The system can be further rewritten as

$$\dot{x} = Y(x, u)\theta, \tag{4.4}$$

where the regressor matrix $Y(x, u)$ is constructed from the known basis functions of the system and is given by

$$Y(x, u) = \begin{bmatrix} Y_1 & 0_{1\times 6} \\ 0_{1\times 2} & Y_2 \end{bmatrix},$$

which contains all known state and input terms, and $\theta = [\theta_1^\top, \theta_2^\top]^\top$ is the parameter vector to be estimated. For this example,

$$Y_1 = [\, x_1 \quad x_2 \,],$$

and

$$Y_2 = [x_1,\ x_2,\ x_2(\cos(2x_1) + 2)^2,\ x_2(\sin(4x_1^2) + 2)^2,$$
$$(\cos(2x_1) + 2)u_1,\ (\sin(4x_1^2) + 2)u_2].$$

The corresponding true parameters are set to

$$\theta_1^* = [\, \theta_{f1}^* \,] = [-2 \quad 1]^\top,$$

and

$$\theta_2^* = [\theta_{f2}^* \quad \theta_g^*]^\top = [-0.5 \quad -1 \quad 0.25 \quad 0.25 \quad 1 \quad 1]^\top.$$

The above reformulation shows that the considered benchmark system satisfies the linear-in-parameters identifier structure adopted in the proposed method. Specifically, the known graph topology specifies the interaction relation between the two players, while the regressor matrix $Y(x, u)$ contains the corresponding known state and input basis functions. By substituting the true parameter vector θ^* into $Y(x, u)\theta^*$, the original nonlinear dynamics in (4.1) can be exactly recovered. It should be emphasized that θ^* is used only to evaluate the identification accuracy in the simulation and is not used in the online control design.

The initial condition is $x(0) = [1, -1]^\top$, the simulation horizon is 20 s, and the sampling period is $T_s = 0.001$ s. For the ICL identifier, the sliding-window width, history-stack size, and history-stack sampling period are set to $\Delta T = 0.3$ s, $M = 20$, and 0.05 s, respectively. The gains are set to $k_{\mathrm{ICL}} = 1$ and $\Gamma = 15$, and parameter updating is enabled only when $\lambda_{\min} \geq 10^{-4}$. Although the concurrent-learning gain k_{ICL} is absorbed into Γ_i in the theoretical derivation for compact notation, it is listed separately from Γ in the simulation to explicitly show the numerical gain selection.

The simulation results are illustrated in Figures 2–5. The state trajectories, parameter-estimation results, and critic and actor weight evolutions are illustrated in Figures 2–5, respectively. During the simulation, the exploration noise is removed at $t = 15$ s. Under the learned optimal control policy, the system states rapidly converge to a neighborhood of the origin, which verifies the closed-loop stability.

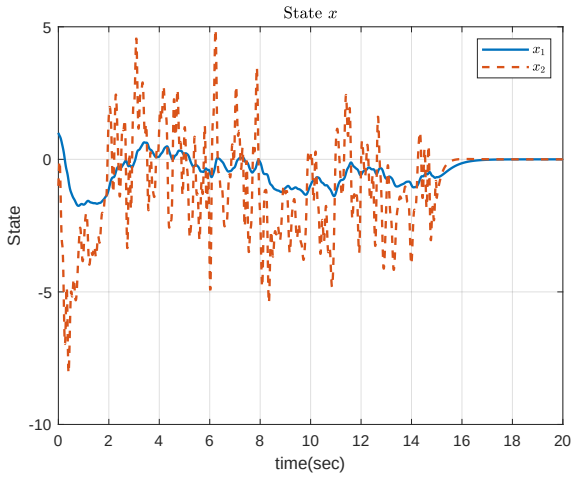


Figure 2. State trajectories of the underactuated nonlinear system.

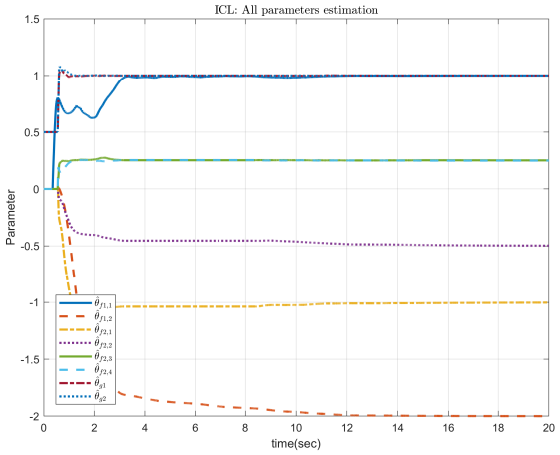
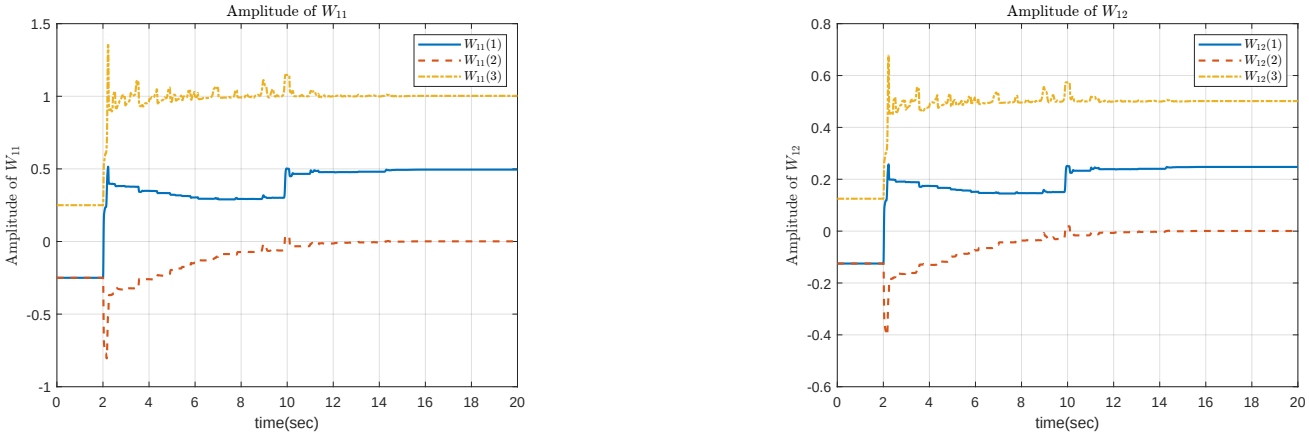


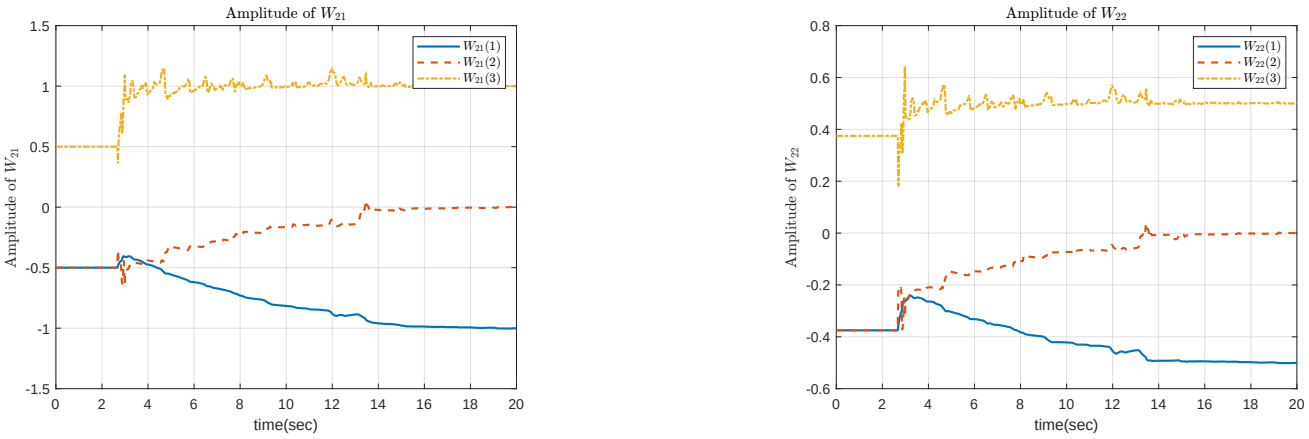
Figure 3. Evolution of the parameter estimates.



(a) Player 1

(b) Player 2

Figure 4. Evolution of the critic weights.



(a) Player 1 (b) Player 2

Figure 5. Evolution of the actor weights.

From the viewpoint of parameter estimation, the minimum eigenvalue of the history stack remains positive throughout the simulation, which guarantees convergence of the parameter estimates. The estimation error of the input-dynamics parameter θ_g rapidly converges to the order of 1×10^{-4} and remains at that level. The drift-dynamics parameters θ_{f1} and θ_{f2} also exhibit gradual convergence, and the final estimation errors remain at the order of 1×10^{-3}. For the critic network, the final estimated weights are

$$\hat{W}_{11}(t_f) = [0.4993,\ 0.0004,\ 1.0007]^\top,$$
$$\hat{W}_{12}(t_f) = [0.2497,\ 0.0002,\ 0.5003]^\top,$$

which agree well with the theoretical optimal values

$$W_{11}^* = [0.5,\ 0,\ 1]^\top, \qquad W_{12}^* = [0.25,\ 0,\ 0.5]^\top.$$

The simulation results can be interpreted together with the Lyapunov-based theoretical analysis as follows. The convergence of the identifier parameters indicates that the gradient-type ICL mechanism and the history stack provide effective information for estimating the unknown dynamics under the FE condition. The convergence of the critic weights shows that the integral Bellman residual can be reduced with the aid of the estimated dynamics. The convergence of the actor weights further indicates that the learned policy can track the critic-induced reference policy. Moreover, the state trajectories converge to a neighborhood of the origin, which is consistent with the uniformly ultimately bounded stability result established in the theoretical analysis. These observations illustrate the closed-loop cooperation among the graph-structure-driven identifier, the critic network, and the actor network, and explain why the proposed ACI-based learning algorithm is effective in the considered simulation.

4.2. Comparison with other algorithms

A comparison with several representative algorithms is summarized in Table 1.

For the proposed method, the learning convergence time in Table 1 is determined by Algorithm 1, where $\Delta t_c = 0.1$ s and $\varepsilon_\theta = \varepsilon_1 = \varepsilon_2 = 10^{-3}$. The weight approximation error is defined as $e_W = \max_{i=1,\dots,N} \|\hat{W}_{1i}(t_f) - W_{1i}^*\| / \|W_{1i}^*\|$, where $\hat{W}_{1i}(t_f)$ is the final learned critic weight and W_{1i}^* is the theoretical optimal critic weight.

Table 1. Comparison with other algorithms.

Metric	Ref. [16]	Ref. [14] Algorithm 1	Ref. [14] Algorithm 2	Ref. [21]	Ours
Paradigm	○	●	●	●	◑
Learning convergence time	300 s	100 s	200 s	80 s	20 s
No. of critic NNs	N	$N^2 + N$	$2N$	$3N$	N
Weight approx. error	10^{-3}	10^{-2}	10^{-3}	10^{-3}	10^{-3}

Note: ○ model-based, ● strictly model-free, ◑ indirect adaptive.

To ensure a fair and rigorous comparison, the compared algorithms were re-evaluated using the codes provided by the original authors under the same software and hardware environment, benchmark system, and initial conditions. The reproduced results were consistent with the reported results. The convergence time in Table 1 denotes the online learning time required for the neural-network weights to converge, rather than the wall-clock computational time.

In terms of learning convergence time, the proposed method significantly improves upon references [14, 16, 21]. Meanwhile, while maintaining a weight approximation error at the order of 10^{-3}, the proposed method requires only N critic networks, which reduces the online computational burden and realizes indirect adaptive control without requiring any exact system model.

Compared with the method in reference [25], once the FE condition is satisfied, the unknown parameters and neural-network weights in that method are updated using least squares. Its advantage is fast convergence, but the price is a relatively high per-step computational complexity of $O(q^2)$, where q is the parameter dimension, together with the need for matrix inversion. As a result, the online computation and storage cost increase rapidly with the number of parameters.

By contrast, the proposed method also first collects history data and then uses the concurrent-learning idea to extract informative samples. The difference is that, once the FE condition is satisfied, the actor network, critic network, and identifier all adopt gradient-type update laws. Consequently, the per-step computational complexity is typically only $O(q)$. Therefore, from a mechanism viewpoint, the proposed algorithm preserves the relaxed-excitation advantage of concurrent learning while retaining the low computational cost and ease of online implementation associated with gradient descent.

5. Conclusions

In this paper, we investigated the multiplayer nonzero-sum differential game problem for continuous-time input-affine nonlinear systems and designed an indirect adaptive optimal control algorithm based on the ACI architecture. The main contribution of this work lies in developing a graph-structure-driven sparse ACI learning framework with a gradient-type ICL identifier, which enables online Nash learning for continuous-time nonlinear nonzero-sum differential graphical games with unknown dynamics. In this framework, the graph topology is embedded into the identifier parameterization as a structural prior, and the resulting parameter-reduction mechanism is clarified through the topology-consistent local model and dimension analysis.

Compared with conventional concurrent learning methods, the proposed gradient-based ICL identifier does not require least-squares updating or covariance-matrix inversion, thereby reducing the

dominant per-step computational cost from $O(q^2)$ to $O(q)$ when the history-stack size and the local basis dimension are fixed. By exploiting the sparsity and structural prior knowledge of the graph topology, the proposed method performs distributed estimation at the identifier layer, while the value-function approximation and policy learning are still written in terms of the global state in the current formulation. The identifier requires only FE for parameter convergence, whereas the closed-loop UUB result additionally relies on PE of the critic regressor.

On this basis, a composite Lyapunov analysis was used to study the boundedness of the closed-loop signals. A representative two-player nonzero-sum game example based on an underactuated nonlinear system was provided, and comparative simulations with existing algorithms were carried out. The simulation results verify the closed-loop Nash learning performance and are consistent with the UUB stability analysis. The reported results indicate that the proposed method can achieve satisfactory control performance without requiring exact system dynamics, while reducing the online computational burden.

Future work will consider larger-scale non-complete sparse graphical games to more comprehensively evaluate the scalability of the proposed ACI-based Nash learning framework and to further demonstrate the sparsity-induced parameter-reduction benefit in closed-loop simulations. In addition, for continuous-time nonlinear systems with input constraints, developing synchronous integral reinforcement-learning algorithms for nonzero-sum games will also be an important direction for further study.

Use of AI tools declaration

The authors declare that artificial intelligence (AI) tools were used for language polishing and limited checking of the presentation and consistency of some theoretical derivations in this manuscript. No AI tools were used for research design, simulation data generation, or drawing the final conclusions.

Conflict of interest

The authors declare there are no conflicts of interest.

References

1. Y. Huo, D. Wang, J. Qiao, M. Li, Off-policy model-free learning for multi-player non-zero-sum games with constrained inputs, *IEEE Trans. Circuits Syst. I Reg. Pap.*, **70** (2023), 910–920. https://doi.org/10.1109/TCSI.2022.3221274
2. Y. Zhang, B. Zhao, D. Liu, S. Zhang, Adaptive dynamic programming-based event-triggered robust control for multiplayer nonzero-sum games with unknown dynamics, *IEEE Trans. Cybern.*, **53** (2023), 5151–5164. https://doi.org/10.1109/TCYB.2022.3175650
3. Q. Wei, L. Zhu, R. Song, P. Zhang, D. Liu, J. Xiao, Model-free adaptive optimal control for unknown nonlinear multiplayer nonzero-sum game, *IEEE Trans. Neural Netw. Learn. Syst.*, **33** (2022), 879–892. https://doi.org/10.1109/TNNLS.2020.3030127

4. D. Liu, S. Xue, B. Zhao, B. Luo, Q. Wei, Adaptive dynamic programming for control: A survey and recent advances, *IEEE Trans. Syst. Man Cybern. Syst.*, **51** (2021), 142–160. https://doi.org/10.1109/TSMC.2020.3042876

5. R. S. Sutton, A. G. Barto, *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, MA, USA, 2018.

6. Z. Pan, D. Lei, L. Wang, A knowledge-based two-population optimization algorithm for distributed energy-efficient parallel machines scheduling, *IEEE Trans. Cybern.*, **52** (2022), 5051–5063. https://doi.org/10.1109/TCYB.2020.3026571

7. H. Wang, B. R. Sarker, J. Li, J. Li, Adaptive scheduling for assembly job shop with uncertain assembly times based on dual Q-learning, *Int. J. Prod. Res.*, **59** (2021), 5867–5883. https://doi.org/10.1080/00207543.2020.1794075

8. F. Zhao, Z. Fu, L. Wang, H. Sang, A heterogeneous graph reinforcement learning framework with question-aware neighborhood aggregation and interoption prompt attention for dynamic flexible job shop scheduling problem, *IEEE Trans. Ind. Inform.*, **22** (2026), 2863–2874. https://doi.org/10.1109/TII.2025.3646962

9. D. Liu, H. Li, D. Wang, Online synchronous approximate optimal learning algorithm for multi-player non-zero-sum games with unknown dynamics, *IEEE Trans. Syst. Man Cybern. Syst.*, **44** (2014), 1015–1027. https://doi.org/10.1109/TSMC.2013.2295351

10. K. G. Vamvoudakis, F. L. Lewis, S. S. Ge, Neural networks in feedback control systems, in *Mechanical Engineers' Handbook, Volume 2: Design, Instrumentation, and Controls*, Wiley, (2014), 843–894. https://doi.org/10.1002/9781118985960.meh223

11. J. Li, J. Ding, T. Chai, F. L. Lewis, Nonzero-sum game reinforcement learning for performance optimization in large-scale industrial processes, *IEEE Trans. Cybern.*, **50** (2020), 4132–4145. https://doi.org/10.1109/TCYB.2019.2950262

12. Q. Zhao, J. Sun, G. Wang, J. Chen, Event-triggered ADP for nonzero-sum games of unknown nonlinear systems, *IEEE Trans. Neural Netw. Learn. Syst.*, **33** (2022), 1905–1913. https://doi.org/10.1109/TNNLS.2021.3071545

13. A. Odekunle, W. Gao, M. Davari, Z. P. Jiang, Reinforcement learning and non-zero-sum game output regulation for multi-player linear uncertain systems, *Automatica*, **112** (2020), 108672. https://doi.org/10.1016/j.automatica.2019.108672

14. J. Sun, H. Zhang, Y. Yan, S. Xu, X. Fan, Optimal regulation strategy for nonzero-sum games of the immune system using adaptive dynamic programming, *IEEE Trans. Cybern.*, **53** (2023), 1475–1484. https://doi.org/10.1109/TCYB.2021.3103820

15. L. Guo, H. Zhao, Model-free adaptive optimal control of continuous-time nonlinear non-zero-sum games based on reinforcement learning, *IET Control Theory Appl.*, **17** (2023), 223–239. https://doi.org/10.1049/cth2.12376

16. M. Johnson, R. Kamalapurkar, S. Bhasin, W. E. Dixon, Approximate *N*-player nonzero-sum game solution for an uncertain continuous nonlinear system, *IEEE Trans. Neural Netw. Learn. Syst.*, **26** (2015), 1645–1658. https://doi.org/10.1109/TNNLS.2014.2350835

17. K. G. Vamvoudakis, F. L. Lewis, Multi-player non-zero-sum games: Online adaptive learning solution of coupled Hamilton–Jacobi equations, *Automatica*, **47** (2011), 1556–1569. https://doi.org/10.1016/j.automatica.2011.03.005

18. T. M. Moerland, J. Broekens, A. Plaat, C. M. Jonker, Model-based reinforcement learning: A survey, *Found. Trends Mach. Learn.*, **16** (2023), 1–118. https://doi.org/10.1561/2200000086

19. D. Vrabie, F. L. Lewis, Neural network approach to continuous-time direct adaptive optimal control for partially unknown nonlinear systems, *Neural Netw.*, **22** (2009), 237–246. https://doi.org/10.1016/j.neunet.2009.03.008

20. J. Y. Lee, J. B. Park, Y. H. Choi, Integral reinforcement learning for continuous-time input-affine nonlinear systems with simultaneous invariant explorations, *IEEE Trans. Neural Netw. Learn. Syst.*, **26** (2015), 916–932. https://doi.org/10.1109/TNNLS.2014.2328590

21. L. Guo, H. Zhao, Online adaptive optimal control algorithm based on synchronous integral reinforcement learning with explorations, *Neurocomputing*, **520** (2023), 250–261. https://doi.org/10.1016/j.neucom.2022.11.055

22. L. Guo, W. Xiong, Y. Song, D. Gan, An efficient model-free adaptive optimal control of continuous-time nonlinear non-zero-sum games based on integral reinforcement learning with exploration, *IET Control Theory Appl.*, **18** (2024), 748–763. https://doi.org/10.1049/cth2.12610

23. S. Bhasin, R. Kamalapurkar, M. Johnson, K. G. Vamvoudakis, F. L. Lewis, W. E. Dixon, A novel actor-critic-identifier architecture for approximate optimal control of uncertain nonlinear systems, *Automatica*, **49** (2013), 82–92. https://doi.org/10.1016/j.automatica.2012.09.019

24. G. Chowdhary, E. Johnson, Concurrent learning for convergence in adaptive control without persistency of excitation, in *49th IEEE Conference on Decision and Control*, IEEE, (2010), 3674–3679. https://doi.org/10.1109/CDC.2010.5717148

25. R. Kamalapurkar, J. R. Klotz, W. E. Dixon, Concurrent learning-based approximate feedback-Nash equilibrium solution of *N*-player nonzero-sum differential games, *IEEE/CAA J. Autom. Sin.*, **1** (2014), 239–247. https://doi.org/10.1109/JAS.2014.7004681

26. F. Tatari, K. G. Vamvoudakis, M. Mazouchi, Optimal distributed learning for disturbance rejection in networked non-linear games under unknown dynamics, *IET Control Theory Appl.*, **13** (2019), 2838–2848. https://doi.org/10.1049/iet-cta.2018.5832

27. A. Parikh, R. Kamalapurkar, W. E. Dixon, Integral concurrent learning: Adaptive control with parameter convergence using finite excitation, *Int. J. Adapt. Control Signal Process.*, **33** (2019), 1775–1787. https://doi.org/10.1002/acs.2945

28. D. M. Le, O. S. Patil, P. M. Amy, W. E. Dixon, Integral concurrent learning-based accelerated gradient adaptive control of uncertain Euler–Lagrange systems, in *2022 American Control Conference*, IEEE, (2022), 806–811. https://doi.org/10.23919/ACC53348.2022.9867710

29. W. Si, S. Gao, M. Zhang, T. Wen, Y. Bai, H. Wang, Approximate optimal control for uncertain nonlinear systems: A reinforcement relearning framework, *Nonlinear Dyn.*, **113** (2025), 24821–24848. https://doi.org/10.1007/s11071-025-11393-9

30. K. G. Vamvoudakis, F. L. Lewis, G. R. Hudas, Multi-agent differential graphical games: Online adaptive learning solution for synchronization with optimality, *Automatica*, **48** (2012), 1598–1611. https://doi.org/10.1016/j.automatica.2012.05.074

31. V. G. Lopez, F. L. Lewis, Y. Wan, M. Liu, G. A. Hewer, K. Estabridis, Stability and robustness analysis of minmax solutions for differential graphical games, *Automatica*, **121** (2020), 109177. https://doi.org/10.1016/j.automatica.2020.109177

32. M. Liu, Y. Wan, V. G. Lopez, F. L. Lewis, G. A. Hewer, K. Estabridis, Differential graphical game with distributed global Nash solution, *IEEE Trans. Control Netw. Syst.*, **8** (2021), 1371–1382. https://doi.org/10.1109/TCNS.2021.3065654

33. V. G. Lopez, F. L. Lewis, M. Liu, Y. Wan, Beyond Nash solutions for differential graphical games, *IEEE Trans. Autom. Control*, **68** (2023), 5791–5797. https://doi.org/10.1109/TAC.2022.3230734

34. G. Huang, Z. Zhang, W. Yan, X. Guo, Differential graphical games of multiagent systems with nonzero leader's control input and external disturbances, *Int. J. Robust Nonlinear Control*, **34** (2024), 8144–8162. https://doi.org/10.1002/rnc.7378

35. B. Lian, W. Xue, F. L. Lewis, A. Davoudi, Nash-minmax strategy for multiplayer multiagent graphical games with reinforcement learning, *IEEE Trans. Control Netw. Syst.*, **12** (2025), 763–775. https://doi.org/10.1109/TCNS.2024.3419823

36. Q. Liu, H. Yan, K. Chen, M. Wang, Z. Li, Distributed Nash equilibrium solution for multi-agent game in adversarial environment: A reinforcement learning method, *Automatica*, **178** (2025), 112342. https://doi.org/10.1016/j.automatica.2025.112342

37. Z. Ming, H. Zhang, J. Zhang, X. Xie, A novel actor-critic-identifier architecture for nonlinear multiagent systems with gradient descent method, *Automatica*, **155** (2023), 111128. https://doi.org/10.1016/j.automatica.2023.111128

38. K. S. Narendra, A. M. Annaswamy, *Stable Adaptive Systems*, Prentice Hall, Englewood Cliffs, NJ, USA, 1989.

39. J. J. E. Slotine, W. Li, *Applied Nonlinear Control*, Prentice Hall, Englewood Cliffs, NJ, USA, 1991.

Appendix

A. Proof of Theorem 1

Since the proof is identical for each player, the subscript i is omitted in this appendix. By (3.5), each recorded data pair satisfies

$$\Delta x_k = \mathcal{Y}_k \theta, \qquad k = 1, \ldots, M.$$

Let the parameter estimation error be

$$\tilde{\theta}(t) = \theta - \hat{\theta}(t).$$

Choose the Lyapunov candidate

$$V_\theta(\tilde{\theta}) = \frac{1}{2}\tilde{\theta}^\top \Gamma^{-1} \tilde{\theta}, \tag{A.1}$$

where Γ is symmetric positive definite. From (3.9), one has

$$\dot{\hat{\theta}} = \Gamma \sum_{k=1}^{M} \mathcal{Y}_k^\top \left(\Delta x_k - \mathcal{Y}_k \hat{\theta}\right) = \Gamma \sum_{k=1}^{M} \mathcal{Y}_k^\top \mathcal{Y}_k \tilde{\theta},$$

and hence

$$\dot{\tilde{\theta}} = -\dot{\hat{\theta}} = -\Gamma \sum_{k=1}^{M} \mathcal{Y}_k^\top \mathcal{Y}_k \tilde{\theta}.$$

Taking the time derivative of (A.1) gives

$$\dot{V}_\theta = \tilde{\theta}^\top \Gamma^{-1} \dot{\tilde{\theta}} = -\tilde{\theta}^\top \left(\sum_{k=1}^{M} \mathcal{Y}_k^\top \mathcal{Y}_k \right) \tilde{\theta}.$$

If the FE condition is satisfied from a finite time t_{FE}, namely,

$$\lambda_{\min}\left(\sum_{k=1}^{M} \mathcal{Y}_k^\top \mathcal{Y}_k \right) \geq \lambda > 0, \qquad \forall t \geq t_{\text{FE}},$$

then, for all $t \geq t_{\text{FE}}$,

$$\dot{V}_\theta \leq -\lambda \|\tilde{\theta}\|^2. \tag{A.2}$$

Moreover,

$$\frac{1}{2}\lambda_{\min}(\Gamma^{-1})\|\tilde{\theta}\|^2 \leq V_\theta \leq \frac{1}{2}\lambda_{\max}(\Gamma^{-1})\|\tilde{\theta}\|^2. \tag{A.3}$$

It follows from (A.2) and (A.3) that

$$\dot{V}_\theta \leq -\frac{2\lambda}{\lambda_{\max}(\Gamma^{-1})} V_\theta, \qquad \forall t \geq t_{\text{FE}}.$$

By the comparison lemma,

$$V_\theta(t) \leq V_\theta(t_{\text{FE}}) \exp\left(-\frac{2\lambda}{\lambda_{\max}(\Gamma^{-1})}(t - t_{\text{FE}}) \right), \qquad \forall t \geq t_{\text{FE}}. \tag{A.4}$$

Hence, $\tilde{\theta}(t)$ converges to zero exponentially for all $t \geq t_{\text{FE}}$. If the FE condition holds from the initial time, then $t_{\text{FE}} = 0$. This completes the proof.

B. Proof of Theorem 2

Proof. For the identifier layer, suppose that there exists a finite time $t_{\text{FE}} \geq 0$ such that, for each player i,

$$\lambda_{\min}\left(\sum_{k=1}^{M} \mathcal{Y}_{ik}^\top \mathcal{Y}_{ik} \right) \geq \lambda_i > 0, \qquad \forall t \geq t_{\text{FE}}. \tag{B.1}$$

Then, by Theorem 1, the identifier parameter estimation error $\tilde{\theta}_i(t)$ converges exponentially to zero for all $t \geq t_{\text{FE}}$. Since the identifier is linearly parameterized and all basis functions are bounded on the compact set Ω, the input-gain estimation error

$$\tilde{g}_i(x) = g_i(x) - \hat{g}_i(x)$$

is bounded on Ω and satisfies $\tilde{g}_i(x(t)) \to 0$ as $t \to \infty$. Therefore, the terms induced by $\tilde{g}_i(x)$ are bounded and can be included in the residual terms of the subsequent Lyapunov analysis.

For the critic layer, define the normalized regressor

$$\bar{\phi}_i(t) = \frac{\Delta\phi_i(t)}{\sqrt{m_{si}(t)}}, \qquad m_{si}(t) = 1 + \Delta\phi_i^\top(t)\Delta\phi_i(t).$$

Assume that there exist positive constants T_c and β_i such that

$$\int_t^{t+T_c} \bar{\phi}_i(\tau)\bar{\phi}_i^\top(\tau)\, d\tau \geq \beta_i I, \qquad \forall t \geq t_{\mathrm{FE}}, \quad i = 1, \ldots, N. \tag{B.2}$$

According to Assumption 1 and the NN approximation setting in the main text, the functions $F(x)$, $g_i(x)$, $\hat{g}_i(x)$, $\phi_i(x)$, $\nabla\phi_i(x)$, $\varepsilon_i(x)$, $\nabla\varepsilon_i(x)$, and the exploration signal $e_i(t)$ are bounded on Ω. Hence, there exist positive constants such that

$$\|F(x)\| \leq b_f\|x\|, \quad \|g_i(x)\| \leq b_{g_i}, \quad \|\hat{g}_i(x)\| \leq \bar{g}_i,$$

$$\|\phi_i(x)\| \leq b_{\phi_i}, \quad \|\nabla\phi_i(x)\| \leq b_{\nabla\phi_i}, \quad |\varepsilon_i(x)| \leq b_{\varepsilon_i}, \quad \|\nabla\varepsilon_i(x)\| \leq b_{\nabla\varepsilon_i},$$

$$\|e_i(t)\| \leq \bar{e}_i, \qquad \|W_{1i}\| \leq \bar{W}_{1i}.$$

Therefore, all approximation and identification residuals appearing below are bounded on Ω.

Define the aggregate value function

$$V_\Sigma(x) = \sum_{i=1}^N V_i(x).$$

Since each running cost satisfies

$$r_i(x, u) = x_i^\top Q_i x_i + \sum_{j=1}^N u_j^\top R_{ij} u_j,$$

where $Q_i > 0$ and $R_{ij} \geq 0$, the value function $V_i(x)$ is positive definite with respect to the origin under an admissible policy. Hence, $V_\Sigma(x)$ is positive definite on Ω. Since $V_\Sigma(x)$ is continuous and Ω is compact, there exist positive constants $\underline{c}_V$ and $\bar{c}_V$ such that

$$\underline{c}_V\|x\|^2 \leq V_\Sigma(x) \leq \bar{c}_V\|x\|^2, \qquad \forall x \in \Omega. \tag{B.3}$$

Choose the composite Lyapunov candidate

$$L = V_\Sigma(x) + \frac{1}{2}\sum_{i=1}^N \tilde{W}_{1i}^\top \alpha_{1i}^{-1}\tilde{W}_{1i} + \frac{1}{2}\sum_{i=1}^N \tilde{W}_{2i}^\top \alpha_{2i}^{-1}\tilde{W}_{2i} + \frac{1}{2}\sum_{i=1}^N \tilde{\theta}_i^\top \Gamma_i^{-1}\tilde{\theta}_i. \tag{B.4}$$

Let

$$z = \mathrm{col}\left\{x, \tilde{W}_{11}, \ldots, \tilde{W}_{1N}, \tilde{W}_{21}, \ldots, \tilde{W}_{2N}, \tilde{\theta}_1, \ldots, \tilde{\theta}_N\right\}.$$

Then, by (B.3) and the positive definiteness of α_{1i}, α_{2i}, and Γ_i, there exist positive constants $\underline{c}_L$ and $\bar{c}_L$ such that

$$\underline{c}_L\|z\|^2 \le L \le \bar{c}_L\|z\|^2. \tag{B.5}$$

For all $t \ge t_{\mathrm{FE}}$, taking the time derivative of (B.4) along the closed-loop trajectories yields

$$\dot{L} = \sum_{i=1}^{N} (L_{1i} + L_{2i} + L_{3i} + L_{4i}),$$

where

$$L_{1i} = \dot{V}_i(x), \qquad L_{2i} = \tilde{W}_{1i}^{\top}\alpha_{1i}^{-1}\dot{\tilde{W}}_{1i},$$
$$L_{3i} = \tilde{W}_{2i}^{\top}\alpha_{2i}^{-1}\dot{\tilde{W}}_{2i}, \qquad L_{4i} = \tilde{\theta}_i^{\top}\Gamma_i^{-1}\dot{\tilde{\theta}}_i.$$

First, from (3.9), one has

$$\dot{\tilde{\theta}}_i = -\Gamma_i \sum_{k=1}^{M} \mathcal{Y}_{ik}^{\top}\mathcal{Y}_{ik}\tilde{\theta}_i.$$

Therefore, by (B.1),

$$L_{4i} = -\tilde{\theta}_i^{\top}\left(\sum_{k=1}^{M} \mathcal{Y}_{ik}^{\top}\mathcal{Y}_{ik}\right)\tilde{\theta}_i \le -\lambda_i\|\tilde{\theta}_i\|^2. \tag{B.6}$$

Next, consider the state-related term. The actual closed-loop dynamics can be written as

$$\dot{x} = F(x) + \sum_{j=1}^{N} (\hat{g}_j(x) + \tilde{g}_j(x))(\hat{u}_j(x) + e_j(t)).$$

Using the HJB relation, the actor approximation relation, and the boundedness of the NN approximation errors, the exploration signal, and the identification errors, the derivative of $V_\Sigma(x)$ can be upper bounded by

$$\sum_{i=1}^{N} L_{1i} \le -\eta_x\|x\|^2 + \chi_a \sum_{j=1}^{N} \|\tilde{W}_{2j}\|^2 + \chi_\theta \sum_{j=1}^{N} \|\tilde{\theta}_j\|^2 + d_x, \tag{B.7}$$

where η_x, χ_a, χ_θ, and d_x are positive constants. The constant d_x collects the bounded residuals caused by the approximation error, the exploration signal, and the transient identification error.

For the critic-weight term, from

$$E_{1i} = -\Delta\phi_i^{\top}\tilde{W}_{1i} - \varepsilon_{B_i}(t)$$

and the critic update law (3.17), one obtains

$$\dot{\tilde{W}}_{1i} = -\alpha_{1i}\bar{\phi}_i\bar{\phi}_i^{\top}\tilde{W}_{1i} - \alpha_{1i}\frac{\Delta\phi_i}{m_{si}}\varepsilon_{B_i}(t).$$

Thus,

$$L_{2i} = -\tilde{W}_{1i}^{\top}\bar{\phi}_i\bar{\phi}_i^{\top}\tilde{W}_{1i} - \varepsilon_{B_i}(t)\tilde{W}_{1i}^{\top}\frac{\Delta\phi_i}{m_{si}}.$$

Since $\Delta\phi_i/m_{si}$ and $\varepsilon_{B_i}(t)$ are bounded on Ω, and since the normalized regressor satisfies the PE condition (B.2), the standard PE-based interval estimate for normalized gradient learning implies that there exist positive constants μ_i and d_{2i} such that

$$\int_t^{t+T_c} L_{2i}(\tau)\,d\tau \le -\mu_i \int_t^{t+T_c} \|\tilde{W}_{1i}(\tau)\|^2\,d\tau + d_{2i}T_c, \qquad \forall t \ge t_{\mathrm{FE}}. \tag{B.8}$$

Consequently, $\tilde{W}_{1i}$ is uniformly ultimately bounded. Since W_{1i} is constant, $\hat{W}_{1i} = W_{1i} - \tilde{W}_{1i}$ is also bounded.

Finally, consider the actor-weight term. Let

$$\Xi_i(t) = \frac{T}{4}\sum_{j=1}^{N} \frac{\Delta\phi_j}{m_{sj}^2}\hat{W}_{1j}^{\top} D_{3ij}.$$

Because $\Delta\phi_j/m_{sj}^2$, D_{3ij}, and $\hat{W}_{1j}$ are bounded, there exists a positive constant $\bar{\Xi}_i$ such that

$$|\Xi_i(t)| \le \bar{\Xi}_i.$$

Using $\tilde{W}_{2i} = W_{1i} - \hat{W}_{2i}$, the actor update law (3.22) gives

$$\dot{\tilde{W}}_{2i} = \alpha_{2i}\left[\tanh\!\big(B_i(\tilde{W}_{2i} - \tilde{W}_{1i})\big) + (c_i - \Xi_i(t))\hat{W}_{2i}\right].$$

Therefore,

$$L_{3i} = \tilde{W}_{2i}^{\top}\tanh\!\big(B_i(\tilde{W}_{2i} - \tilde{W}_{1i})\big) + (c_i - \Xi_i(t))\tilde{W}_{2i}^{\top}\hat{W}_{2i}.$$

Since

$$\hat{W}_{2i} = W_{1i} - \tilde{W}_{2i},$$

one has

$$\tilde{W}_{2i}^{\top}\hat{W}_{2i} = \tilde{W}_{2i}^{\top}W_{1i} - \|\tilde{W}_{2i}\|^2 \le \bar{W}_{1i}\|\tilde{W}_{2i}\| - \|\tilde{W}_{2i}\|^2.$$

Moreover, the hyperbolic tangent term is bounded and satisfies, for some positive constants,

$$\left|\tilde{W}_{2i}^{\top}\tanh\!\big(B_i(\tilde{W}_{2i} - \tilde{W}_{1i})\big)\right| \le \chi_{31i}\|\tilde{W}_{2i}\|^2 + \chi_{32i}\|\tilde{W}_{1i}\|^2 + d_{31i}.$$

Using these bounds and applying Young's inequality, one obtains

$$L_{3i} \le -\eta_{3i}\|\tilde{W}_{2i}\|^2 + \chi_{3i}\|\tilde{W}_{1i}\|^2 + d_{3i}, \tag{B.9}$$

where η_{3i}, χ_{3i}, and d_{3i} are positive constants, provided that c_i is selected sufficiently large to dominate the bounded term $\Xi_i(t)$ and the cross terms.

Combining (B.6)–(B.9), and integrating over $[t, t + T_c]$, yields

$$L(t + T_c) - L(t) \le -\sigma\int_t^{t+T_c} \|z(\tau)\|^2\,d\tau + \delta T_c, \qquad \forall t \ge t_{\mathrm{FE}}, \tag{B.10}$$

where σ and δ are positive constants.

Define

$$t_k = t_{\mathrm{FE}} + kT_c, \qquad k = 0, 1, 2, \ldots.$$

Since $\dot{L}$ is bounded on Ω, there exists a constant $\bar{d}_L > 0$ such that

$$|L(\tau) - L(t_k)| \le \bar{d}_L(\tau - t_k), \qquad \forall \tau \in [t_k, t_{k+1}].$$

Using (B.5), one obtains

$$\int_{t_k}^{t_{k+1}} \|z(\tau)\|^2 \, d\tau \ge \frac{1}{\bar{c}_L} \int_{t_k}^{t_{k+1}} L(\tau) \, d\tau \ge \frac{T_c}{\bar{c}_L} L(t_k) - \frac{\bar{d}_L T_c^2}{2\bar{c}_L}.$$

Substituting this estimate into (B.10) with $t = t_k$ gives

$$L(t_{k+1}) \le \left(1 - \frac{\sigma T_c}{\bar{c}_L}\right) L(t_k) + \delta T_c + \frac{\sigma \bar{d}_L T_c^2}{2\bar{c}_L}.$$

Choose the design parameters such that

$$0 < \rho := 1 - \frac{\sigma T_c}{\bar{c}_L} < 1.$$

Then

$$L(t_{k+1}) \le \rho L(t_k) + \bar{\delta}, \qquad \bar{\delta} = \delta T_c + \frac{\sigma \bar{d}_L T_c^2}{2\bar{c}_L}. \tag{B.11}$$

By recursive application of (B.11),

$$L(t_k) \le \rho^k L(t_0) + \frac{\bar{\delta}}{1 - \rho}, \qquad t_0 = t_{\mathrm{FE}}.$$

Therefore, $L(t_k)$ is uniformly ultimately bounded. Since $\dot{L}$ is bounded on each interval $[t_k, t_{k+1}]$, $L(t)$ is also uniformly ultimately bounded in continuous time. Finally, by (B.5), the composite error vector $z(t)$ is uniformly ultimately bounded. In particular, the system state x, the critic weight estimation errors $\tilde{W}_{1i}$, and the actor weight estimation errors $\tilde{W}_{2i}$ are uniformly ultimately bounded. In addition, $\tilde{\theta}_i(t)$ converges exponentially to zero for all $t \ge t_{\mathrm{FE}}$.

Since the interval $[0, t_{\mathrm{FE}}]$ is finite and the closed-loop signals are continuous on the compact operating set Ω, boundedness on $[0, t_{\mathrm{FE}}]$ follows directly. Hence, the conclusion holds for all $t \ge 0$. This completes the proof.

AIMS Press